DETERMINACIÓN DE LOS PRINCIPALES FACTORES QUE IMPACTAN LA OCURRENCIA DE BACKREAMING EN POZOS PETROLEROS A TRAVÉS DE LA ANALÍTICA DE DATOS

Integrantes: Jennifer Jackelin Monsalve Leon, Jennifer Astrid Rodriguez Pava y Wilson Rafael Velarde Sanchez
Asesor: Juan Fernando Pérez Bernal Pre-asesora: Natalia Pacheco Carvajal

Resumen

Durante las operaciones de perforación de pozos petroleros, luego de llegar a la profundidad deseada se requiere realizar un viaje a superficie, que en condiciones normales se realiza en elevadores; sin embargo, en algunas ocasiones se hace necesario aplicar backreaming para lograr pasar las restricciones y llegar a superficie con las herramientas, dicho proceso genera tiempos no productivos en la operación, que se traducen en costos adicionales para la compañía. Este proyecto estudia los diferentes factores y parámetros de la perforación, que pueden impactar en la ocurrencia de backreaming, a través de la analítica de datos, aplicando modelos de clasificación binaria para 13 pozos de la compañía Sierracol. El modelo Random Forest resulta ser el mejor prediciendo la ocurrencia de backreaming y disminuyendo la probabilidad de tener Falsos Negativos, con una exactitud de 88.6%. A partir de esto, se desarrolló una herramienta predictiva y analítica con el objetivo de encontrar la combinación de parámetros que ayude a mitigar la frecuencia de esta operación no deseada; preliminarmente, se calcula que el impacto de lograr disminuir la recurrencia de backreaming se vea reflejado en la reducción del 5% de los costos planeados de un pozo, que ascienden aproximadamente a 200 mil USD. A través de los resultados de este estudio se espera generar la base para futuras investigaciones y análisis en las operaciones de perforación en la industria Oil & Gas.

Palabras clave

Backreaming, factores, modelo de clasificación, Random Forest, machine learning, predicción.

Introducción

La analítica de datos se vuelve un aliado clave en la industria petrolera para la toma de decisiones informadas que apuntan a mejorar sus procesos eficientemente. A continuación, se presenta un proyecto pionero para la empresa SierraCol que integra la analítica con procesos de perforación, en búsqueda de prever la recurrencia de backreaming en pozos petroleros. En la extracción de hidrocarburos subterráneos se utiliza un taladro que perfora el suelo y roca subyacente hasta llegar a la capa deseada; en algunas ocasiones al sacar este ensamblaje a superficie se presentan restricciones que generan la necesidad de aplicar operaciones de backreaming, proceso que puede considerarse como "perforar hacia atrás" para salir del pozo. Este proceso no deseado impacta directamente el desempeño de las operaciones de perforación y la calidad del hueco para las operaciones subsiguientes, por lo que es de suma importancia su análisis; siendo así, que en el campo objeto de este estudio para el periodo analizado, el backreaming generó 311.3 hrs de tiempo no productivo (NPT), representando aproximadamente $ 1,387,000.00 USD de sobrecosto.

A través de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) se aplica, en este estudio, modelos analíticos que aprovechan los parámetros de perforación y las características del pozo para anticipar la probabilidad de ocurrencia de backreaming. Con este objetivo no solo se busca optimizar la planificación y ejecución de la perforación, sino también reducir costos y riesgos operativos asociados, así como mejorar la toma de decisiones estratégicas en la industria petrolera.

Con el proceso de selección de variables y tras el análisis de diferentes modelos, se obtiene para el presente estudio, un modelo Random Forest con una exactitud del 88.6% para la predicción del backreaming. Adicionalmente, se integra el método de los SHAP values en el modelo obtenido, con el fin de entender cómo cada variable independiente contribuye al resultado, obteniendo hallazgos muy valiosos como, por ejemplo, la probabilidad de ocurrencia de backreaming se incrementa, dado una mayor profundidad, mayor inclinación, mayor flujo, menor velocidad de perforación o si el pozo es de 3 secciones, en lugar de 2.

Para la culminación de este proyecto se presenta la herramienta denominada "Backreaming Prediction Tool" que apoya la planeación del proceso de perforación, calculando la probabilidad de backreaming para diferentes escenarios, con base en el modelo Random Forest desarrollado; permitiendo tomar decisiones informadas que mitiguen el uso de backreaming. Lo que de manera preliminar se estima puede verse reflejado en la reducción del 5% de los costos planeados de un pozo que ascienden aproximadamente a 200 mil USD.

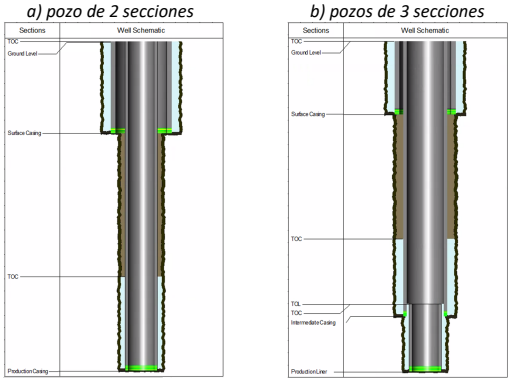
Definiciones claves

La ***perforación***, en la industria Oil & Gas, se refiere al proceso mediante el cual se perfora un agujero con una broca para crear un pozo para la producción de petróleo y/o gas natural. De acuerdo con las condiciones del yacimiento para lograr la producción de los recursos, se define una arquitectura conocida como *estado mecánico*, que es el esquema que indica cómo está estructurado el pozo; en la etapa de perforación, indica la cantidad de

secciones y tamaños de hueco definidos para alcanzar el yacimiento objetivo.

Figura 1. Diagrama Estado Mecánico

a) pozo de 2 secciones *b) pozos de 3 secciones*



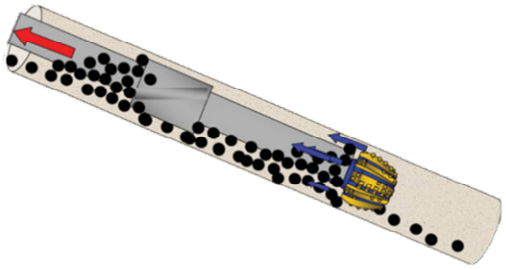
Fuente: Autores

El proceso de perforación de un pozo implica varios pasos importantes y la combinación de varios sistemas. Para realizar la perforación de una sección se inicia con el arme del ensamblaje con el que se va a realizar la perforación, se debe asegurar tener el fluido de perforación en las condiciones adecuadas y los equipos suficientes en superficie para llevar a cabo la perforación y recibir los retornos de ésta. Una vez se llega a la profundidad planeada para la sección, se realizan ciclos de limpieza y se procede a sacar a superficie el ensamblaje de perforación, este viaje a superficie en condiciones normales se realiza en elevadores (libre, sin restricciones); en otros casos, se pueden presentar restricciones que requieran de ayudas mecánicas para poder superarlas, esta ayuda mecánica se conoce como backreaming.

El ***backreaming*** consiste en bombear y girar la sarta de perforación al tiempo que se extrae del pozo; puede considerarse como "perforar hacia atrás" con el fin de salir del pozo cuando se presentan inconvenientes para sacar la tubería del pozo sin rotación ni circulación.

En general, las operaciones de backreaming se han convertido en una solución para las malas condiciones del pozo durante los viajes a superficie, pero también es conocida por causar los mismos problemas que se supone que deben evitar. Si el backreaming no se realiza correctamente, puede complicar las operaciones generando problemas de empaquetamiento (packoffs), inestabilidad de hueco y afectaciones en las herramientas, tanto en la sarta de perforación como en el ensamble de fondo (BHA [1]).

Figura 2. Backreaming



Fuente: A guide to successful backreaming: Real-time case histories.

Las operaciones de backreaming requieren de tiempo adicional de taladro comparado con un viaje a superficie en elevadores. Esto puede estar justificado o no, dependiendo de las condiciones del hueco; es por esto que para algunos Operadores este tiempo se define como **No Productivo.**

Sierracol hace parte de las compañías operadoras para las que el requerimiento de backreaming en sus operaciones se considera tiempo no productivo puesto que impacta directamente el desempeño tanto en tiempos como en costos.

Revisión de Literatura

La ocurrencia de backreaming en las operaciones de perforación es frecuente sin discriminar el tipo de pozo ni la ubicación de éste. Los estudios existentes se han enfocado principalmente en el manejo de riesgos durante dichas operaciones y la definición de procedimientos para evitar complicaciones operativas al ejecutarlo. Dentro de los principales artículos revisados como referencia para reconocer el impacto de este proceso no deseado en las operaciones de perforación, se encontró el modelamiento de vibraciones en el BHA bajo condiciones de backreaming para evitar los daños a las herramientas de fondo asociados a las operaciones de backreaming. Por otro lado, relacionado con los problemas operativos que se pueden presentar al momento de ejecutar el backreaming, se tiene el estudio elaborado por la SPE y Schlumberger (2010) donde se realizó análisis post-mortem de diferentes pozos, donde se presentaron problemas de pega de tubería y empaquetamiento como resultado de mala praxis de dicha operación, dejando como resultado prácticas recomendadas para el backreaming considerando el tamaño de hueco, inclinación y condiciones de pozo, de acuerdo con lo observado durante la perforación.

Como se evidencia, el backreaming es una operación que genera disfuncionalidades en las herramientas e ineficiencias en el desempeño de perforación que han sido analizadas y para las cuales se han generado guías de ejecución; sin embargo, hasta el momento, dentro de la

[1] BHA: por sus siglas en inglés Bottom Hole Assembly (Ensamblaje de Fondo)

revisión realizada, no se tienen resultados de estudios de analítica relacionados con la prevención y/o mitigación de este, es decir, enfocados en la definición de estrategias preventivas para evitar la recurrencia del backreaming.

Metodología

Este caso de estudio se abordó con base en los planteamientos de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual representa un método probado para orientar los trabajos de minería de datos (IBM Corporation, 2021), en la que se plantean seis fases, como se muestra en la Figura 3.

Figura 3. Diagrama de Metología aplicada

Fuente: Instituto de Ingeniería del conocimiento

La aplicación de esta metodología sobre el caso de estudio se presenta a continuación, utilizando el software libre **Python** integrando distintas fuentes de información de la compañía, que se encontraban en las bases de datos de OpenWells y Compass (Landamrk), e información adquirida con los sensores de los taladros de perforación, tales como Totco WellData (NOV) y Rig cloud (Nabors).

CASO DE ESTUDIO

¿Es posible definir la probabilidad de ocurrencia de backreaming y obtener una herramienta que ayude a mitigar la necesidad de realizar esta operación durante los viajes a superficie?

1. Comprensión del negocio

Este proyecto se realizó para la empresa SierraCol Energy la cual es una organización independiente de exploración y producción de hidrocarburos en Colombia. Se trabajó con el departamento de Perforación para el área de la cuenca de los Llanos con los pozos de desarrollo, de diseño convencional de dos y tres secciones: superficie y producción; y superficie, intermedio y producción, respectivamente.

En la mayoría de los pozos perforados, en las últimas campañas de perforación, ha sido necesario realizar operaciones de backreaming para poder sacar a superficie el ensamblaje con el cual se realizó la perforación. El impacto de estas operaciones no solamente se evidencia en el desempeño de la perforación relacionado a tiempos y costos, sino también en los recursos que son asignados anualmente para la ejecución de nuevos proyectos.

En el caso de Sierracol, en las campañas de perforación de 2019 a 2023, en el campo de estudio, se ha tenido un total de **311.3 hrs** de tiempo no productivo (NPT) a causa del backreaming, lo que representa aproximadamente **$ 1,387,000.00 USD** que equivale al 24% y 36% del costo total de un pozo nuevo de 3 y 2 secciones, respectivamente; en cuanto al tiempo, **12.5 días** que corresponde a la mitad del tiempo de perforación de un pozo tipo del campo en estudio. Por otro lado, la duración del backreaming por pozo, es más del 20% del tiempo que toma la perforación, y no solamente afecta los tiempos planeados para las operaciones, sino que genera un impacto en la calidad del hueco para las corridas de revestimiento y cementaciones, que son las operaciones siguientes para el aseguramiento de la sección.

Las estadísticas compartidas sobre los tiempos y costos que ha asumido Sierracol con las operaciones de backreaming, evidencian la importancia de evaluar e identificar las posibles causas de la condición de hueco, que influyen en la necesidad de realizar backreaming en los pozos que se perforan. Se puede decir que el lograr mitigar y, en el mejor de los casos, eliminar el backreaming, permitiría al departamento de perforación incluir mayor número de pozos en sus campañas de perforación y reducir aproximadamente un 5% (200 mil USD) el AFE (Authorization for expenditure) de sus pozos.

2. Comprensión de los datos

De acuerdo a lo revisado con el equipo de ingeniería de Sierracol, las bases de datos proporcionadas y consideradas como las más relevantes para la identificación de backreaming son: los parámetros de perforación en cada una de las formaciones que se atraviesan en el pozo; la trayectoria direccional, el tipo de fluido de perforación, propiedades y aditivos que pueden generar reacciones en las formaciones que las hace cerrarse pasada una cantidad de tiempo; entre otros.

Se decide trabajar con la información de pozos perforados desde 2019 hasta junio de 2023, en total **13 pozos**. A continuación, se presenta el número de registros para cada una de las bases de datos oficiales, de la compañía y terceras compañías de servicios aliadas, que fueron compartidas para el desarrollo del caso de estudio.

Tabla 1. Registros Base de Datos

		Registros
Bases de datos	Parámetros de perforación	1'206'857
	Trayectorias	2'483

Litología	294
NPT Back	63
BHA	1'051
Fluid Properties	941
Productos Fluidos	16'618
Hole Sections	72

Fuente: Autores

En esta fase se identificaron diferentes tipos de problemas como: datos perdidos, errores de datos, incoherencias de codificación, características que no eran útiles dada la variable objetivo, entre otros.

Descripción de variables

Parte fundamental para la comprensión de la data es el ejercicio de dar una definición concreta a cada una de las variables de las diferentes bases de datos. En la Tabla 2 se muestra la definición de las principales variables a evaluar.

Tabla 2. Descripción principales variables

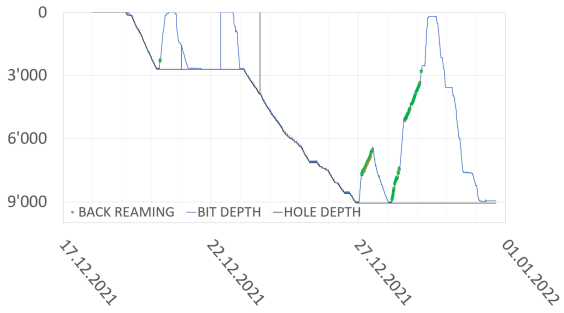
Variable	Traducción	Definición
Depth	Profundidad (ft)	Indica la distancia desde la superficie hasta la profundidad alcanzada en el pozo.
Bit Position	Posición de la broca (ft)	Ubicación específica en la que se encuentra la broca dentro del pozo.
Block Height	Altura del Bloque (ft)	Indica la altura del bloque, utilizando como referencia cero la mesa (rig-floor), con valores positivos hacia arriba de la mesa.
Bit Weight	Peso de la Broca (Klbs)	Carga aplicada sobre la broca para permitir la perforación.
Flow In Rate	Tasa de Flujo de Entrada (GPM)	Velocidad de flujo del fluido dentro del pozo, crucial para lubricar la broca, mantener la integridad del agujero y controlar la circulación del fluido.
Pump Pressure	Presión de la Bomba (psi)	Presión generada por las bombas, determinada por la velocidad de la bomba y ajustada según simulaciones para garantizar la presión adecuada para la operación.
Hook Load	Carga del Gancho (klbs)	Fuerza ejercida sobre el gancho, utilizada para monitorear la operación y detectar posibles problemas como bloqueos.
Top Drive RPM	Revoluciones por Minuto del Top Drive (RPM)	Velocidad de rotación del top drive, responsable de hacer girar la tubería durante la perforación, medida en revoluciones por minuto (RPM).
Top Drive Torque	Torque del Top Drive (lbft)	Fuerza de torsión generada en la broca al entrar en contacto con la roca
String Speed	Velocidad de la Sarta (ft/min)	Velocidad a la que se desplaza la sarta de perforación en el pozo.
ROP - Average	Tasa de Penetración – Promedio (ft/hr)	Velocidad promedio de perforación, medida en pies por hora.
Diff Press	Presión Diferencial (psi)	Diferencia de presión, indicativa del desempeño del pozo.

Fuente: Autores

Variable objetivo

Se identificó la presencia y ausencia de backreaming durante el viaje a superficie de cada uno de los pozos en estudio, y se definió la variable categórica **"BR_1"** que permitiera reconocer las profundidades a las cuales se tuvo necesidad de realizar operaciones de backreaming.

Figura 4. Identificación de Backreaming



Fuente: Autores

En la Figura 4 se incluyen las curvas de Hole Depth, profundidad del hueco; Bit Depth, posición de la broca en el momento; y sobre estas los puntos de backreaming, identificados en verde, que permiten reconocer la posición de la broca en el momento que se requirió esta operación.

Es importante mencionar, que la necesidad de aplicar el backreaming surge de las condiciones del hueco generadas al momento de la perforación, por lo que los parámetros asociados a dichas condiciones son los evaluados en el modelo, y no los obtenidos durante el viaje a superficie.

Tabla 3. Definición de variable objetivo backreaming

Variable objetivo	*Presencia*	*Ausencia*
Backreaming	1	0

Fuente: Autores

3. Preparación de los datos

Considerando la cantidad de registros de la base de datos que contenía los *parámetros de perforación*, donde se tenía información de avance cada 0.1ft, se define un intervalo de profundidad de **10 ft**, para unificar en paralelo las diferentes bases de datos con el uso de la variable **"Bit position"**. El tratamiento de datos faltantes se realizó siguiendo las recomendaciones de expertos en el negocio, asegurando la integridad y fiabilidad de los datos.

Se realiza el Análisis Exploratorio de Datos (EDA) y se valida con el negocio que los resultados sean coherentes según su conocimiento y experiencia en el tema. El EDA se centra en una exhaustiva comprensión de las características del conjunto de datos empleando técnicas de visualización como histogramas y diagramas de caja, esenciales para identificar posibles patrones; así como análisis estadísticos y de transformación de datos como la métrica Weight of Evidence (WOE), Information Value (IV) y coeficientes de correlación como el punto biserial y Cramér's.

- ***Weight of Evidence (WOE):*** Es una técnica que mide la fortaleza de la relación entre una variable

predictora y la variable objetivo, usualmente en problemas de clasificación binaria. Según Zeng (2013), el indicador WOE contribuye a la medición de la fuerza de cada atributo de la variable independiente.

La transformación del WOE de un predictor categórico se puede definir como sigue (Raymaekers, J., Verbeke, W., & Verdonck, T. (2022)): Suponiendo una categoría j con Nj elementos. Denotando Pj como el número de casos verdaderos en la categoría y Fj el número de casos falsos. Siendo P el número total de casos verdaderos y F el número total de casos falsos en los datos. El valor de WOE de la categoría j se da por:

$$\log\left(\frac{P_j/P}{F_j/F}\right)$$

De acuerdo con Dassatti (2019) se deben tener en cuenta las siguientes consideraciones:

a. Un indicador con valores negativos implica una alta proporción de malos (casos falsos) sobre buenos (casos verdaderos).
b. Para que el WOE esté definido, ninguna de las clases puede estar conformada únicamente con buenos malos.
c. Tanto la tasa de malos como el valor del WOE deben ser lo suficientemente diferentes entre cada grupo; la agrupación o segmentación se debe realizar de forma tal que se maximice la diferencia entre buenos y malos. Cuanto mayor sea la diferencia entre el WOE de los grupos, mayor será la capacidad predictiva asociada a ese grupo o atributo.

Finalmente, cabe mencionar que mientras el WOE describe la relación entre una variable independiente y la dependiente, existe el information value (IV) que mide la fuerza de esa relación.

- ***Information Value (IV):*** Es un indicador del poder predictivo de cada una de las variables independientes, que se construye a partir de la suma de los WOE a lo largo de las categorías, ponderados por las diferencias proporcionales halladas en dichos rangos. Para un número k de categorías, se tiene la siguiente expresión:

$$IV = \sum_{i=1}^{k} (\%Buenos_i - \%Malos_i) * WOE_i$$

Las variables con un IV menor a 0.10 se consideran generalmente como débiles, valores mayores a 0.30 se asocian con variables con fuerte poder predictivo y valores superiores a 0.5 suelen asociarse con variables que sobre-predicen a la variable independiente, por lo que se aconseja no considerar dichas variables o considerarlas de forma controlada (Siddiqi 2012), (Zeng 2013).

- ***Cramér's V*:** Es una medida del tamaño del efecto para la prueba de chi-cuadrado de independencia. Mide qué tan fuertemente están asociados dos campos categóricos (IBM 2023).

La fórmula para el Cramer's V esta dada por:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

donde N es el número total de observaciones y k es el menor entre el número de filas y columnas.

Tabla 4. Interpretación en Cramér's V

Tamaño del efecto (ES)	Interpretación
ES ≤ 0.2	El resultado es débil. Aunque el resultado es estadísticamente significativo, los campos sólo están débilmente asociados.
0.2 < ES ≤ 0.6	El resultado es moderado. Los campos están moderadamente asociados.
ES > 0.6	El resultado es fuerte. Los campos están fuertemente asociados.

Fuente: IBM

- ***Punto biserial:*** es una medida de la relación entre una variable dicotómica (binaria) y una variable continua.

Los valores que puede asumir este coeficiente se encuentran entre menos uno y uno, con el cero indicando que no hay correlación alguna entre las variables que se están comparando.

Una fórmula para calcular este coeficiente es la siguiente:

$$rpbis = \frac{M_1 - M_0}{S_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

Donde,
M_1= Media del puntaje global del instrumento del grupo que contestó de manera positiva a la variable binaria.
M_0 = Media del puntaje global del instrumento del grupo que contestó de manera negativa a la variable binaria.
S_n = Desviación estándar del instrumento.
n = Tamaño de la población que contestó el instrumento.
n_1 = Tamaño del grupo que contestó de manera positiva a la variable binaria.
n_0 = Tamaño del grupo que contestó de manera negativa a la variable binaria.

Selección Variables significativas

Una vez realizado el EDA y la aplicación de los indicadores explicados anteriormente, se procede con el proceso de selección de variables diferenciándolas entre categóricas y numéricas.

- ***Categóricas:***

Se analiza las variables en función a los indicadores Information Value (IV) y el coeficiente de Cramer, seleccionando aquellas que presenten una capacidad predictiva a partir de moderada y una correlación mayor al **20%** (a partir de una asociación moderada).

A continuación, se muestra los resultados obtenidos:

Tabla 5. Exploración y Análisis de variables categóricas

Variable	IV	Capacidad Predictiva	Correlación
Formation	77%	Muy Fuerte	41%
Block	18%	Moderada	21%
Well	67%	Muy Fuerte	38%
Site	56%	Muy Fuerte	36%
Drilling Phase	113%	Muy Fuerte	47%
Secciones	28%	Moderada	26%
Assembly Name	52%	Muy Fuerte	37%

Fuente: Autores

Como se observa, todas las variables cuentan con indicadores relevantes. Sin embargo, dado que se busca desarrollar un modelo de predicción aplicable a nuevos pozos en general, en lugar de uno específico para cada pozo, se descartan aquellas variables cuyos valores están directamente asociados a un pozo en particular, tales como: Block, Site y Well.

Adicionalmente, de acuerdo con lo explicado por los expertos de negocio, se descarta la variable Assembly Name, ya que se indica que a pesar de tener las mismas categorías en esta variable se pueden tener diferencias en la configuración de los componentes, por lo que no es una variable de identificación única para determinar la influencia del BHA y, por tanto, no sería útil para la generación del modelo y posterior predicción.

Finalmente, las variables categóricas elegidas de forma preliminar son: ***Drilling Phase, Secciones y Formation***.

- ***Numéricas:***

A continuación, se pasa a evaluar 84 variables numéricas analizando su distribución, cuartiles, diagramas de cajas por clase y correlación de punto biserial, seleccionando aquellas variables que mantengan una correlación mayor al **20%** y que sean estadísticamente significativas.

Se obtienen 23 variables que cumplen con estas características; sin embargo, 4 de ellas (Dispersante, Inhibidor_Arcilla, Lump_Sum, Bactericida) presentan más del 70% de valores en cero y son descartadas. Por tanto, preliminarmente se eligen las siguientes variables numéricas:

- Total_Depth
- Depth_Bit
- Flow_In
- Pump_Pressure
- Hookload
- Top_Drive_RPM
- Top_Drive_Torque
- ROP_Depth_Hour
- Inclinación
- Av_Density
- Av_PV
- Av_YP
- Av_Gels_10m
- Av_Gels_30m
- Viscosificante
- Detergente
- Anti_acrecion
- Control_Lutitas
- Control_perdidas

Posteriormente, se analiza la correlación entre las variables y los Factores Infladores de Varianza (VIF) a través de un modelo logístico simple, para seleccionar aquellas variables correlacionadas con un mayor nivel de importancia en el modelo y descartar aquellas que ayuden a disminuir el VIF, con el fin de mitigar el riesgo de multicolinealidad entre las variables predictoras. A partir de esto, se reduce finalmente el número de variables a 11 (2 categóricas y 9 numéricas), las cuales se muestran en la siguiente tabla.

Tabla 6. Variables Seleccionadas

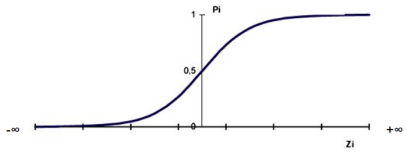
Feature	Tipo de Variable	Correlac. con Backraming	Information Value	VIF
Secciones	Categórica	26%	27.7%	-
Formation	Categórica	41%	77.1%	-
Total Depth	Numérica	52%	-	8.39
Inclinación	Numérica	42%	-	7.07
Flow In	Numérica	29%	-	10.48
Av Gels 30m	Numérica	30%	-	7.64
ROP Depth Hour	Numérica	-20%	-	3.35
Viscosificante	Numérica	-20%	-	1.30
Detergente	Numérica	-22%	-	2.06
Control Lutitas	Numérica	20%	-	1.52
Control Pérdidas	Numérica	-22%	-	1.20

4. Modelado

Considerando que la variable principal de análisis de este estudio es la presencia/ausencia de backreaming dadas unas condiciones específicas, se decide trabajar con cinco de los ***modelos de clasificación*** que se tienen en la literatura. Se ajustaron modelos de: **Regresión Logística, Support Vector Machines (SVM,) Árboles de decisión, Random Forest y XGBoost.**

- ***Regresión logística:*** Es un método lineal de clasificación que se basa en la función logística o sigmoidal, que es una curva en forma de "S" que permite modelar la probabilidad de que la variable dependiente tenga un valor determinado en función de las variables independientes. La función logística convierte cualquier valor de entrada en un valor entre 0 y 1, lo que se interpreta como la probabilidad de que el evento ocurra.

Figura 5. Función sigmoide



Fuente: Clases MIIA - Universidad de los Andes

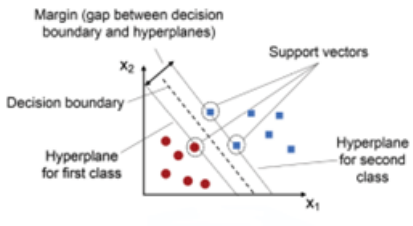
$$Pi = \frac{1}{1 + e^{-(b_1.x_1 + b_p.x_{p1} + bo)}}$$

Donde:

$\boldsymbol{P_i}$ es la probabilidad de que ocurra el evento de interés.
$\boldsymbol{x_1, x_{p1}}$ son las variables independientes asociadas al evento de interés.
$\boldsymbol{b_1, b_p}$ son los coeficientes asociados a cada variable independiente

- ***SVM (Super vector machine classifier):*** Es un algoritmo de aprendizaje supervisado en el que el objetivo es encontrar un hiperplano que separe al máximo las dos clases. Produce límites no lineales construyendo un límite lineal en una versión amplia y transformada del espacio de características en evaluación. El hiperplano se selecciona de forma que maximice la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase, que se denominan vectores de soporte.

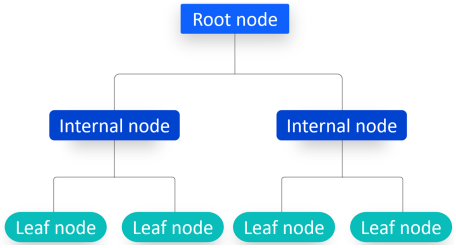
Figura 6. SVM



Fuente: Analytics Vidhya

- ***Árboles de Decisión (Decision Tree):*** Es un algoritmo para clasificar utilizando particiones sucesivas. Es de carácter descriptivo, por tanto, no es interpretable a partir de parámetros. En los árboles de decisión se encuentran los siguientes componentes: nodos, ramas y hojas. Los nodos son las variables de entrada, las ramas representan los posibles valores de las variables de entrada y las hojas son los posibles valores de la variable de salida.

Figura 7. Decision tree configuration



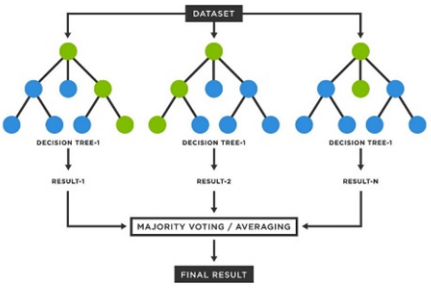
Fuente: IBM

El aprendizaje de árboles de decisión emplea una estrategia de búsqueda codiciosa para identificar los puntos de división óptimos dentro de un árbol. Este proceso de división se repite de forma descendente y recursiva hasta que todos o la mayoría de los registros se han clasificado con etiquetas de clase específicas.

Como alternativa para mejorar el desempeño de los modelos de aprendizaje individuales se tienen los ***métodos de ensamble***, los cuales se componen de un conjunto de clasificadores. Los métodos más conocidos son el *bagging* [2] y el *boosting* [3] y el objetivo de su aplicación es minimizar la varianza y maximizar el sesgo.

- ***Random Forest (Bosques Aleatorios):*** Es un algoritmo que combina los resultados de varios árboles de decisión para llegar a un único resultado. Utiliza el método bagging, generando un subconjunto aleatorio de características que garantiza la baja correlación entre los árboles de decisión.

Figura 8. Random forest configuration



Fuente: GeeksforGeeks

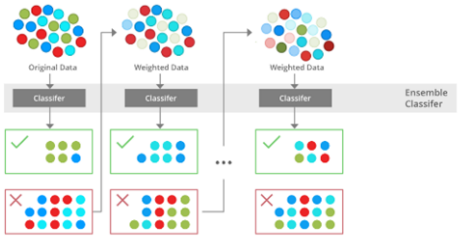
- ***XGBoost (Extreme Gradient Boosting):*** Es un algoritmo que combina los resultados de varios

[2] Bagging: Aprendizaje paralelo. Se aplican cuando se tiene alta varianza y sesgo bajo.

[3] Boosting: Aprendizaje secuencial. Se aplican cuando se tiene baja varianza y sesgo alto.

modelos de clasificación para llegar a un único resultado. Utiliza el método boosting y la potencia de los modelos de árboles de decisión para mejorar la precisión predictiva. Tiene un enfoque de optimización de gradiente para mejorar de manera iterativa la precisión del modelo, minimizando la función de pérdida.

Figura 9. XGBoost configuration



Fuente: GeeksforGeeks

Desarrollo de Modelos

Para el desarrollo de los modelos se siguen los siguientes pasos:

- Se dividen los datos en: base de construcción (con información de 11 pozos y 11,183 registros) y base de validación (con información de 2 pozos y 2,058 registros), que el modelo no conocerá en la fase de construcción.
- Se divide la base de construcción en entrenamiento (70%) y prueba (30%) de forma estratificada.
- Se evalúa la proporción de los datos de la clase a predecir, verificando que no se requiere balanceo ya que se cuenta con 5,999 datos de la clase No Backreaming y 5,184 de la clase Backreaming, siendo la proporción de 1.16 de una clase sobre la otra.
- Se realiza el preprocesamiento de los datos empleando Standard Scaler para las variables numéricas y One Hot Encoder para las variables categóricas.
- Se ejecutan los 5 modelos con los parámetros por defecto, y se itera (forward y backward) con las 11 variables seleccionadas, eligiendo la combinación con la que se obtengan los mejores resultados, siendo ésta la combinación de 6 variables: **Total Depth, Flow In, Inclination, Av_Gels_30m, ROP Depth Hour y Sections**.
- Posteriormente, con las variables elegidas, se procede a buscar los mejores valores para los hiperparámetros de los modelos, con el fin de generar los mejores resultados intentando reducir el sobre ajuste.
- Finalmente se calculan los indicadores de los modelos a través de validación cruzada y los datos de prueba.

5. Evaluación y selección del modelo

Para realizar la evaluación de los modelos aplicados, se incluyeron todos los indicadores de desempeño que hacen parte de la *matriz de confusión*: precisión, exactitud, especificidad, sensibilidad, y puntuación f1; que permiten valorar la capacidad de cada modelo para identificar el backreaming.

Tabla 7. Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: QUORA

- Verdaderos positivos: número de eventos positivos correctamente predichos.
- Falsos positivos: número de eventos predichos como positivos siendo realmente negativos.
- Falsos negativos: número de eventos predichos como negativos siendo realmente positivos.
- Verdaderos negativos: número de eventos negativos correctamente predichos.

Tabla 8. Indicadores de desempeño clasificación

Indicador	Definición	Ecuación
Sensibilidad	Proporción de los verdaderos éxitos que han sido correctamente clasificados.	$\frac{VP}{FN+VP}$
Especificidad	Proporción de verdaderos fracasos que han sido correctamente clasificados.	$\frac{VN}{VN+FP}$
Precisión	Proporción de los éxitos predichos que son verdaderos éxitos.	$\frac{VP}{FP+VP}$
Exactitud	Proporción global que se ha clasificado correctamente.	$\frac{VP+VN}{n}$
F1 score	Media armónica de precisión y recuperación. En general, es una medida de la precisión y robustez de su modelo.	$\frac{2VP}{2VP+FP+FN}$

Fuente: Autores

Lo ideal en los modelos de clasificación, es que todos los indicadores de la *Tabla 8* sean iguales a 1, lo que indicaría una predicción perfecta de la variable objetivo.

Es pertinente precisar, que para el objetivo del caso de estudio el error más crítico, es del tipo II un **falso negativo**, es decir que, el modelo prediga la no ocurrencia de backreaming cuando en realidad sí se presenta, por lo que se pondrá mayor énfasis en los indicadores de Exactitud (Accuracy) y Sensibilidad (Recall).

En la Tabla 9, se comparan los resultados obtenidos con los 5 modelos entrenados para realizar la predicción de ocurrencia de backreaming. En el Anexo 3 se puede apreciar el detalle de los resultados obtenidos para los datos de test como los de validación.

Tabla 9. Resultados modelos

Modelo	Total	
	Accuracy	Recall Back
Random Forest	88.6%	88.3%
XG Boost	87.8%	86.1%
Decision Tree	85.6%	86.8%
SVM	84.9%	83.2%
Regresión Logística	79.7%	78.0%

Fuente: Autores

De los resultados, se observa que todos los modelos presentan métricas aceptables siendo los mejores modelos Random Forest y XGBoost con exactitud de 88.6% y 87.8% respectivamente, mientras que el modelo de más bajo desempeño es el Logístico con una exactitud de 79.7%.

Finalmente, el modelo seleccionado es el de **Random Forest**, el cual es ligeramente superior al de XGBoost y muestra un buen desempeño en la base de prueba, en la base de validación y en el total de los datos tanto en el indicador de exactitud, como en el de sensibilidad; además de mantenerse estable y sin sobre ajuste en todas las bases evaluadas.

Principales factores que impactan en la ocurrencia de Backreaming

Los modelos de Machine learning a menudo se perciben como cajas negras, ya que reciben un conjunto de características como entrada y producen predicciones como resultado. A pesar de tener indicadores de rendimiento excelentes, se mantiene la necesidad de comprender cómo cada característica influye en las predicciones y cuáles son las variables más relevantes para estos resultados. Por lo anterior, en este estudio se analizan 2 métodos para entender los principales factores que impactan en el backreaming: importancia de variables y los valores de Shapley, como métodos de explicación del modelo.

- ***Importancia de variables:*** El método de importancia de variables en Random Forest obtiene una medida que refleja cuánto contribuye cada variable a la reducción de la impureza de Gini, cuanto mayor sea dicha disminución, más importante se considera la variable en el proceso de toma de decisiones del modelo.

  Es importante destacar que esta medida de importancia no necesariamente indica la dirección de la relación entre la variable y la respuesta, solo indican la influencia en la predicción del modelo.

En la Tabla 10, se muestran los valores obtenidos por este método:

Tabla 10. Importancia de variables

Variable	*Importancia*
Total Depth	0.3640
Inclination	0.2592
Flow in	0.1657
Av Gels 30m	0.1014
ROP (ft/hr)	0.0433
3 Sections	0.0409
2 Sections	0.0255

Se observa que las principales variables que contribuyen a reducir la impureza Gini son: Total Depth, Inclination y Flow in.

- ***SHAP (Shapley Values):*** Es un método para mostrar el impacto relativo de cada variable sobre la salida final del modelo, comparando el efecto relativo de las entradas frente al promedio.

  De acuerdo con Fadel (2022), la explicación técnica del concepto de SHAP es el cálculo de los valores de Shapley a partir de la teoría de juegos coalicionales. El valor de Shapley busca medir la contribución de cada jugador al juego, considerándose una competencia entre coaliciones en lugar de entre jugadores individuales. El valor de Shapley se define como la contribución marginal del valor de la variable a la predicción entre todas las coaliciones o subconjuntos concebibles de características. Es un enfoque para redistribuir las ganancias totales entre los participantes. La cantidad que recibe cada característica después de un juego se define de la siguiente manera:

$$\phi_i x = \sum_{S \subset F\{i\}} \frac{|S|!|F|-|S|-1!}{|F|!} f_{S\cup\{i\}} x_{S\cup\{i\}} - f_S x_S$$

Donde,

x: la entrada de la observación
$\phi i(x)$: valor de Shapley para la característica i para la entrada x para el juego/modelo f.
F: el conjunto de todas las características
f_S : el modelo entrenado en el subconjunto de características S.
$f_{S\cup i}$: el modelo entrenado en el subconjunto de características S y $\{i\}$.
x_S: la entrada restringida de x dada el subconjunto de características S.
$x_{S\cup i}$: la entrada restringida de x dada el subconjunto de características S y $\{i\}$.

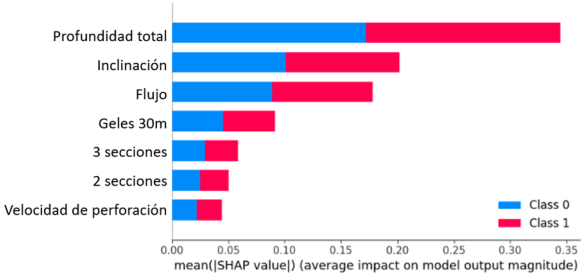
Tabla 11. Cuatro propiedades del valor de Shapley

Definición	Descripción
Eficiencia	La suma de los valores de Shapley de todas las características es igual al valor de la predicción entrenada con todas las características, de modo que la predicción total se distribuye entre las características.
Simetría	Las contribuciones de dos valores de características deben ser iguales si contribuyen de manera igual a todas las posibles coaliciones.
Dummy	Una característica que no cambia el valor predicho independientemente de a qué coalición de valores de características se agregue, debería tener un valor de Shapley de 0.
Linealidad	Si se combinan dos modelos descritos por las funciones de predicción f y g, la predicción distribuida debería corresponder a las contribuciones derivadas de f y las contribuciones derivadas de g.

Fuente: Statistics Canada

A continuación, se muestran los resultados del SHAP value para el modelo Random Forest:

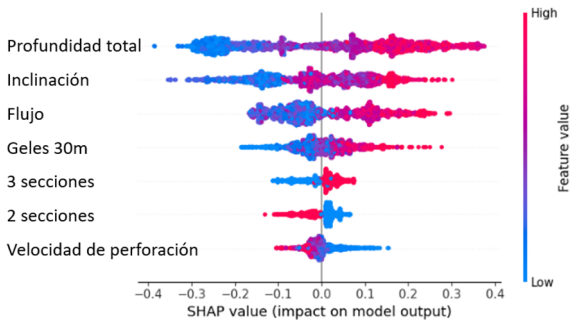
Figura 10. SHAP Value: Impacto Medio en la Magnitud de la Salida del Modelo



Fuente: Autores

Se observa que, principalmente las variables Total Depth, Inclination y Flow in, en ese orden, desempeñan un papel importante a la hora de determinar la predicción.

Figura 11. SHAP Value: Impacto en la salida del modelo (ocurrencia de backreaming)



Fuente: Autores

De la Figura 11 se interpreta, que la probabilidad de backreaming incrementa bajo las siguientes condiciones:

- A mayor profundidad.
- A mayor inclinación.
- A mayor flujo (GPM).
- Cuando la propiedad de Geles a 30m es mayor.
- Si el pozo es de 3 secciones, en lugar de 2.
- A menor velocidad de perforación.

En resumen, bajo ambos métodos, se observa que los resultados son similares, y contrastados con los expertos del negocio, resultan coherentes, interpretables y de utilidad.

6. Despliegue

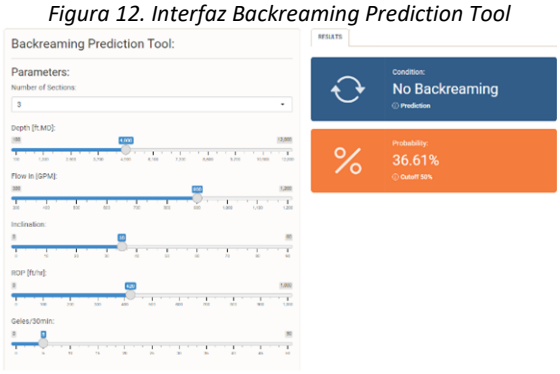
Una vez evaluado y seleccionado el modelo que logra la mejor predicción de ocurrencia de backreaming se procede a la generación de la herramienta que apoya la toma de decisiones, sobre cambios en los parámetros de diseño de los pozos, desde la planeación, para mitigar la recurrencia del backreaming e impactar positivamente en los tiempos, costos y calidad de pozos a entregar en producción.

Herramienta de Predicción

Backreaming Prediction Tool se desarrolla a través de la librería *shiny*[4] de R, donde el modelo Random Forest, generado para la predicción de backreaming, es llamado a través de una interfaz de programación (api) que se conecta en el servidor de Sierracol.

La herramienta se diseña considerando el rango de valores que normalmente tienen las variables que resultaron importantes para la predicción del backremaing, de manera que el usuario, en este caso los ingenieros de perforación encargados de definir los parámetros de perforación para un pozo, puedan realizar cambios en estos parámetros, desde la planeación, y observar el efecto en la probabilidad de presencia de backreaming a una profundidad determinada, la cual les va a permitir tomar acciones preventivas para mitigar el efecto de esta operación.

Figura 12. Interfaz Backreaming Prediction Tool



[4] Shiny: es un paquete de R de código abierto que permite convertir análisis en aplicaciones web interactivas.

Como se puede observar en la Figura 12, para cada una de las 6 variables del modelo, se crea un control deslizante que permite seleccionar el valor exacto que se quiere evaluar en cada variable con el fin de obtener como resultado la probabilidad de backreaming, la cual tiene un cutoff del **50%**, es decir, una vez la probabilidad de ocurrencia llega a este valor, el modelo lo identifica como presencia de backreaming, dando el mensaje ***"Backreaming"***; en caso contrario, se identifica como ***"No backreaming"***.

Por otro lado, se incluye un análisis de sensibilidad de parámetros en dos dimensiones, dejando como variable fija la profundidad, ya que es el parámetro principal para la perforación de pozos, porque nos muestra el avance en esta operación. El objetivo es analizar una a una, la variación de cada una de las variables de interés respecto a la profundidad, manteniendo valores por defecto para las variables fijas que no entran en el análisis, con el fin de identificar las zonas de menor probabilidad de backreaming, visualizadas en tonos verdes; y las zonas de mayor probabilidad de backreaming identificadas en tonos rojos.

Figura 13. Sensibilidad de Profundidad vs Inclinación @ 4900ft MD

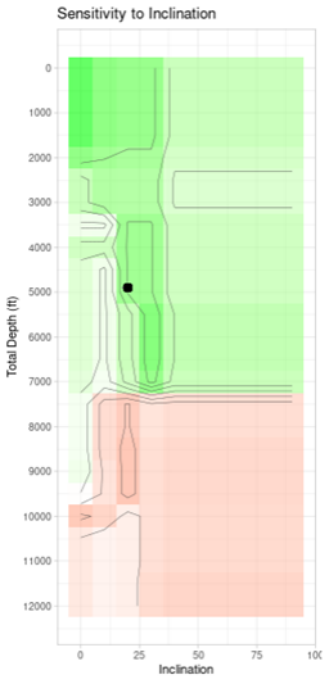
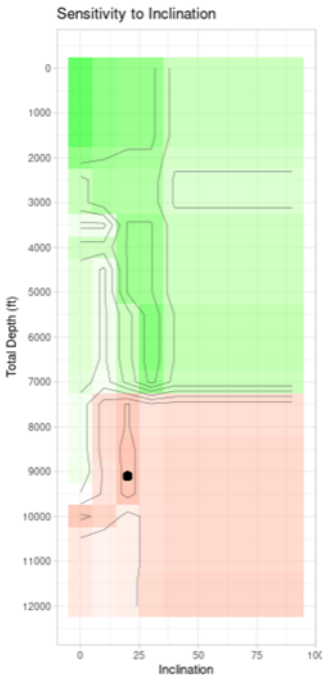


Figura 14. Sensibilidad de Profundidad vs Inclinación @ 9100ft MD



En las Figuras Sensibilidad de Profundidad vs Inclinación @ 4900ft MD y Sensibilidad de Profundidad vs Inclinación @ 9100ft MD, se visualiza el efecto de la profundidad en el incremento de la probabilidad de backreaming. El punto negro, identifica los parámetros definidos para la variable de interés y de acuerdo con su ubicación, en zonas verdes o rojas, se puede determinar el efecto en la probabilidad de ocurrencia de dicha operación. En estos dos casos, se evidencia que, para las condiciones evaluadas, luego de los 10,000ft MD es muy poco probable lograr mitigar la probabilidad del backreaming, asumiendo los valores por defecto para las demás variables.

Es importante mencionar, que para la toma de decisiones sobre los parámetros a utilizar en las operaciones de perforación se debe utilizar la interfaz donde se consideran todas las variables que están incluidas en el modelo. El módulo de sensibilidad se genera solamente como referencia del comportamiento de cada variable e identificación de posibles acciones que puedan ayudar a disminuir la probabilidad de backreaming al considerarlo en todo el modelo.

<u>Discusión de resultados</u>

Se evidencia un gran apoyo al proceso de planificación, para la definición de parámetros de perforación y diseño de los pozos, a pesar de que la implementación oficial del Backreaming Prediction Tool se encuentra aún en proceso.

Preliminarmente, al realizar variaciones de las variables en una profundidad de interés se encuentra que las inclinaciones mayores (> 40°) en zonas profundas tiene

mayor probabilidad de backreaming que en inclinaciones menores (< 20°). Esto impacta directamente al diseño de las trayectorias direccionales, generando como parámetro trabajar en la disminución de la inclinación para atravesar las formaciones más profundas.

Por otro lado, el modelo obtenido y la herramienta desarrollada permiten confirmar consideraciones que el equipo de perforación ya presentía sobre el efecto de la cantidad de secciones en el pozo. En este caso, se confirma y soporta con los datos que, al perforar pozos en 3 secciones se tiene mayor probabilidad de backreaming.

Adicionalmente, al identificar que a mayor profundidad se incrementa la probabilidad de backreaming se puede determinar manejar parámetros que aplaquen un mejor desempeño en las zonas someras, para luego utilizar parámetros controlados en zonas profundas y lograr tener un porcentaje de backreaming mucho menor al obtenido históricamente sin afectar el desempeño de la perforación, equilibrando los riesgos.

Con el estudio realizado se ratifica la importancia de realizar análisis de los datos adquiridos durante las operaciones de perforación para generar estrategias que permitan atacar directamente los tiempos no productivos (NPT) que impactan severamente los costos de las campañas de perforación. Como fue mencionado anteriormente, el lograr mitigar la recurrencia de backreaming representa para Sierracol la posibilidad de aumentar la cantidad de pozos asignados a las campañas de perforación anuales, como consecuencia de la reducción de aproximadamente un 5% (200 KUSD) del AFE, que hace parte de los fondos asignados a contingencias y/o eventos operativos que se pueden presentar durante la perforación de un pozo, donde el backreaming está considerado.

Conclusiones

Este estudio permitió determinar los principales factores que impactan la ocurrencia de backreaming, a partir de las bases de datos compartidas por la operadora Sierracol, los cuales son: profundidad total, rata de flujo hacia el pozo, la velocidad de perforación, la inclinación, el número de secciones y la propiedad del fluido relacionada a los geles.

A partir del análisis exploratorio y descriptivo se identificó que algunas de las variables que la compañía consideraba cruciales para el desarrollo del modelo por su posible influencia sobre la recurrencia de backreaming, tales como: tortuosidad, nombre de la fase de perforación y tamaño del hueco; estaban correlacionadas con otras variables o definitivamente su comportamiento no indicaban una relación directa con el efecto del backreaming.

Por otro lado, algunas variables con capacidad predictiva fuerte no pudieron ser utilizadas en el diseño del modelo de clasificación, debido a que los datos dentro de ellas no eran concluyentes y/o comparables, es decir, incluían información única relacionada a cada pozo lo que no brindaba información adicional al modelo. Por tanto, se sugiere a la compañía desarrollar una manera de recolección de datos que permita identificar mayor información en este tipo de variables como, por ejemplo, configuraciones de BHA, las cuales, de acuerdo con lo revisado con los expertos se consideran que impactan directamente en la calidad del hueco y, por ende, podrían relacionarse con la recurrencia de backreaming.

Referencias

- ¿Qué es el confusion matrix? (2023). Quora. Recuperado de https://es.quora.com/Qu%C3%A9-es-el-confusion-matrix
- An introduction to explainable AI with Shapley values. (2018). SHAP. Recuperado de https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html
- Arboles de clasificación (2019). Estadística y Machine Learning con R. Recuperado de https://bookdown.org/content/2274/metodos-de-clasificacion.html#arboles-de-clasificacion
- Babu, S. C., Gajanan, S. N., & Sanyal, P. (2014). Chapter 4 - Effects of Technology Adoption and Gender of Household Head: The Issue, Its Importance in Food Security—Application of Cramer's V and Phi Coefficient. En S. C. Babu, S. N. Gajanan y P. Sanyal (Eds.), Food Security, Poverty and Nutrition Policy Analysis (Second Edition) (pp. 93-115). Academic Press. ISBN 9780124058644. https://doi.org/10.1016/B978-0-12-405864-4.00004-1.
- CERTIFIED ANALYTICS PROFESSIONAL (CAP®), Examination Study Guide. CAP INFORMS
- Comprensión de la Matriz de Confusión y Cómo Implementarla en Python (2020). DataSource.ai. Recuperado de https://www.datasource.ai/es/data-science-articles/comprension-de-la-matriz-de-confusion-y-como-implementarla-en-python
- Cramér's V. (2023). IBM. Recuperado de: https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-cramrs-v
- Dassatti, Cecilia. (2009). Modelos de Score Crediticio: revisión metodológica y análisis a partir de datos de encuesta. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3443515
- DATAtab Team. (2023). DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. Recuperado de https://datatab.es/tutorial/point-biserial-correlation
- DRILLING. STUDENT ENERGY. Recuperado de https://studentenergy.org/production/drilling/
- Fadel, Soufiane (2022). Explainable Machine Learning, Game Theory, and Shapley Values: A technical review. Statistics Canada. Recuperado de

https://www.statcan.gc.ca/en/data-science/network/explainable-learning
- Guide on Support Vector Machine (SVM) Algorithm (2023). Analytics Vidhya. Recuperado de Shttps://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
- IBM Corporation, Guía de CRISP-DM de IBM SPSS. Recuperado de https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf
- IQR (Rango intercuartílico) (2023). ORACLE. Recuperado de https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_IQR.htm#PFUSU-GUID-CF37CAEA-730B-4346-801E-64612719FF6B
- La metodología CRISP-DM en ciencia de datos (2021). Instituto de Ingeniería del conocimiento. Recuperado de https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/
- LinkedIn. Understanding Torque & Drag: Concepts and Analysis. Recuperado de https://www.linkedin.com/pulse/understanding-torque-drag-concepts-analysis-sixto-romero#:~:text=Torque%2C%20is%20the%20rotational%20force,element%20of%20the%20drill%20string
- Mendoza Juan (2019). RPubs by RStudio. Correlación biserial puntual - Psicometría con R. Recuperado de: https://rpubs.com/jboscomendoza/correlacion_biserial_puntual_r
- Molnar, Christoph. (2023). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Shapley Values. (2023). Recuperado de https://christophm.github.io/interpretable-ml-book/shapley.html
- Raymaekers, J., Verbeke, W., & Verdonck, T. (2022). Weight-of-evidence through shrinkage and spline binning for interpretable nonlinear classification. Applied Soft Computing, 115, 108160. Recuperado de: https://www.sciencedirect.com/science/article/pii/S1568494621010218
- Regresión logística (2021). Gamco. Recuperado de https://gamco.es/glosario/regresion-logistica/
- Samuel, R., & Mirani, A. (2015). Vibration modeling and analysis under backreaming condition. Paper presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, USA. https://doi.org/10.2118/174797-MS
- Schlumberger. Energy Glossary en Español. Recuperado de https://glossary.slb.com/es/terms/d/drilling_fluid
- Support Vector Machine (SVM) Python Example (2023). Analytics Yogi. Reimagining Data-driven Society with Data Science & AI. Recuperado de https://vitalflux.com/classification-model-svm-classifier-python-example/
- Understand Weight of Evidence and Information Value! (2021). Analytics Vidhya. Recuperado de https://www.analyticsvidhya.com/blog/2021/06/understand-weight-of-evidence-and-information-value/
- What is a Decision Tree? (2023). IBM. Recuperado de (https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.
- What is random forest? (2023). IBM. Recuperado de https://www.ibm.com/topics/random-forest
- XGBoost (2023). GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/xgboost/
- Yarim, G., Ritchie, G. M., & May, R. B. B. (2010). A guide to successful backreaming: Real-time case histories. SPE Drill & Completion, 25, 27–38. https://doi.org/10.2118/116555-PA
- Easy web applications in R. Shiny. Recuperado de: https://www.rstudio.com/products/shiny/

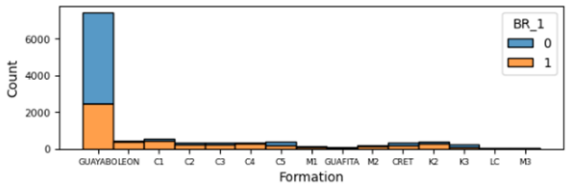
Anexos

Anexo 1: EDA

- ***Formation:***

Ilustración 1. Distribución de la variable Formation

	Feature	Cantidad	Partic.%	Clase_F	Clase_T	%_Clase_F
0	GUAYABO	7432	66.458	2451	4981	32.979
1	K3	224	2.00304	97	127	43.3036
2	C5	380	3.39801	187	193	49.2105
3	M3	46	0.411339	27	19	58.6957
4	CRET	354	3.16552	208	146	58.7571
5	GUAFITA	89	0.795851	53	36	59.5506
6	M1	143	1.27873	88	55	61.5385
7	M2	171	1.52911	124	47	72.5146
8	C2	330	2.95091	263	67	79.697
9	C3	317	2.83466	257	60	81.0726
10	C4	326	2.91514	267	59	81.9018
11	K2	372	3.32648	308	64	82.7957
12	C1	525	4.69463	441	84	84
13	LEON	452	4.04185	391	61	86.5044
14	LC	22	0.196727	22	0	100



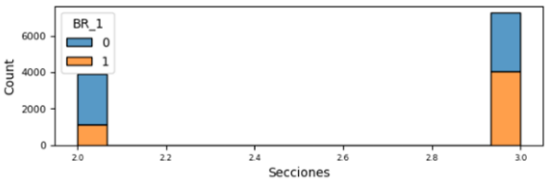
Information Value: 0.7708
La variable tiene una capacidad predictiva muy fuerte.

Coeficiente de contingencia (Cramér's V): 0.4134
P-valor: 0.0
Existe una relación moderada.

Fuente: Autores

- ***Secciones***

Ilustración 2. Distribución de la variable Secciones



Information Value: 0.2769
La variable tiene una capacidad predictiva moderada.

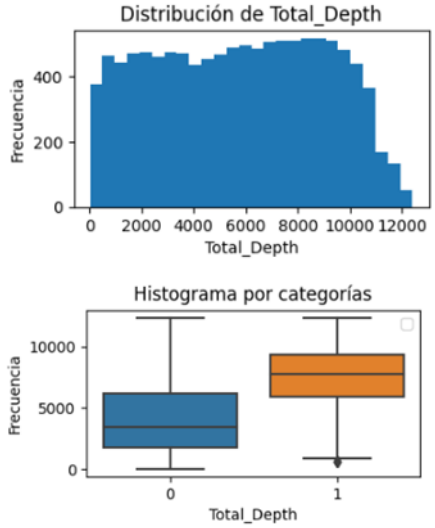
Coeficiente de contingencia (Cramér's V): 0.2561
P-valor: 0.0
Existe una relación moderada.

	Feature	Cantidad	Partic.%	Clase_F	Clase_T	%_Clase_F
0	2	3896	34.8386	1125	2771	28.8758
1	3	7287	65.1614	4059	3228	55.7019

Fuente: Autores

- ***Total_Depth***

Ilustración 3. Distribución de la variable Total_Depth



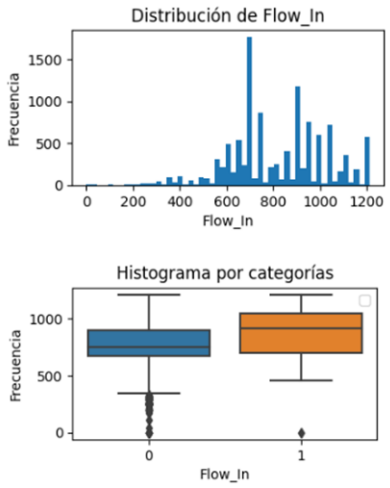

```
Coeficiente de correlación punto biserial: 0.52
Valor de p: 0.0

**********************************************
Distribución de Total_Depth
**********************************************
count    11183.000000
mean      5754.101980
std       3211.223221
min         19.610000
25%       2989.770000
50%       5840.139160
75%       8479.555000
max      12399.580000
```

Fuente: Autores

- ***Flow_In***

Ilustración 4. Distribución de la variable Flow_In



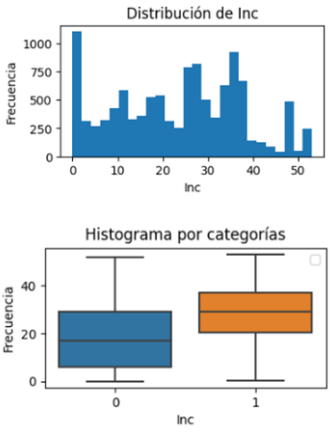

```
Coeficiente de correlación punto biserial: 0.29
Valor de p: 0.0

**********************************************
Distribución de Flow_In
**********************************************
count    11183.000000
mean       822.850387
std        203.500874
min          0.000000
25%        699.421500
50%        804.000000
75%        957.000000
max       1213.739100
```

Fuente: Autores

- ***Inclinación***

Ilustración 5. Distribución de la variable Inclinación



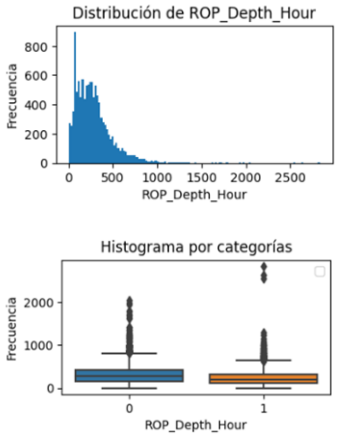

```
Coeficiente de correlación punto biserial: 0.42
Valor de p: 0.0

************************************************
Distribución de Inc
************************************************
count    11183.000000
mean        23.335465
std         14.032977
min          0.000000
25%         11.410000
50%         25.600000
75%         34.610000
max         53.080000
```

Fuente: Autores

- ***ROP_Depth_Hour***

Ilustración 6. Distribución de la variable ROP_Depth_Hour



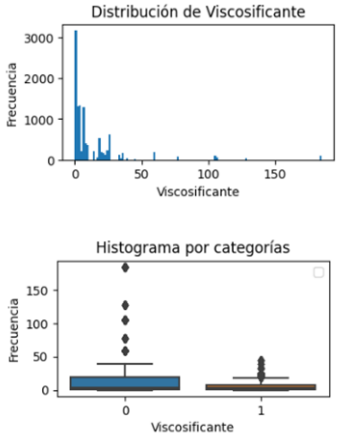

```
Coeficiente de correlación punto biserial: -0.2
Valor de p: 0.0

*********************************************
Distribución de ROP_Depth_Hour
*********************************************
count    11183.000000
mean       276.371171
std        203.000175
min          0.000000
25%        126.200000
50%        242.695450
75%        364.105710
max       2834.200000
```

Fuente: Autores

- ***Viscosificante***

Ilustración 7. Distribución de la variable Viscosificante



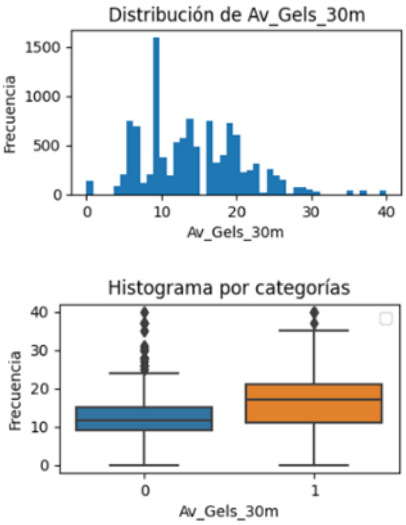

```
Coeficiente de correlación punto biserial: -0.2
Valor de p: 0.0

*********************************
Distribución de Viscosificante
*********************************
count    11183.000000
mean        12.490477
std         24.759376
min          0.000000
25%          1.000000
50%          4.000000
75%         17.500000
max        185.000000
```

Fuente: Autores

- ***Av_Gels_30m***

Ilustración 8. Distribución de la variable Av_Gels_30m



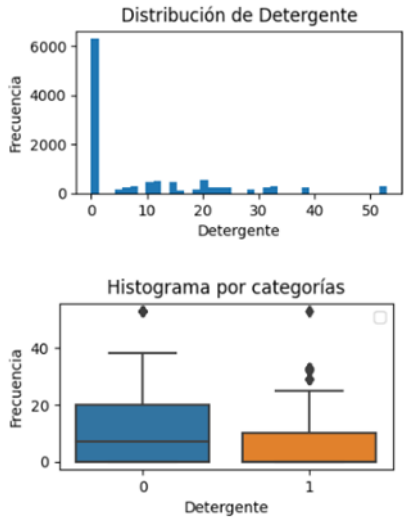

```
Coeficiente de correlación punto biserial: 0.3
Valor de p: 0.0

************************************************
Distribución de Av_Gels_30m
************************************************
count    11183.000000
mean        14.139853
std          6.550778
min          0.000000
25%          9.000000
50%         14.000000
75%         19.000000
max         40.000000
```

Fuente: Autores

- ***Detergente***

Ilustración 9. Distribución de la variable Detergente



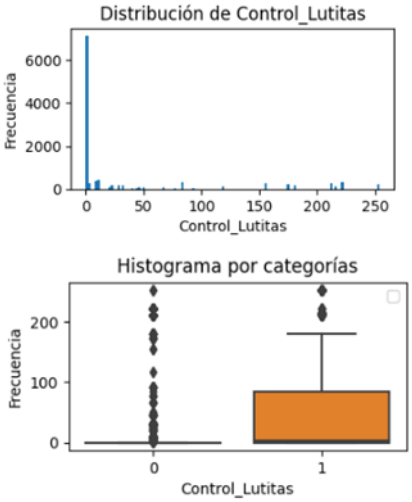

```
Coeficiente de correlación punto biserial: -0.22
Valor de p: 0.0

********************************
Distribución de Detergente
********************************
count    11183.000000
mean         9.128588
std         13.133354
min          0.000000
25%          0.000000
50%          0.000000
75%         16.000000
max         53.000000
```

Fuente: Autores

- ***Control_Lutitas***

Ilustración 10. Distribución de la variable Control_Lutitas



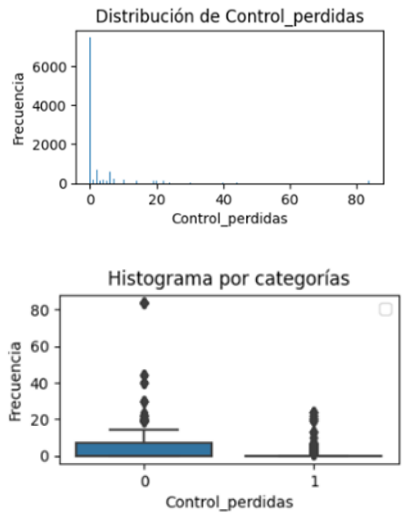

```
Coeficiente de correlación punto biserial: 0.2
Valor de p: 0.0

*********************************************
Distribución de Control_Lutitas
*********************************************
count    11183.000000
mean        35.968702
std         71.243176
min          0.000000
25%          0.000000
50%          0.000000
75%         22.000000
max        254.000000
```

Fuente: Autores

- ***Control_perdidas***

Ilustración 11. Distribución variable Control_perdidas



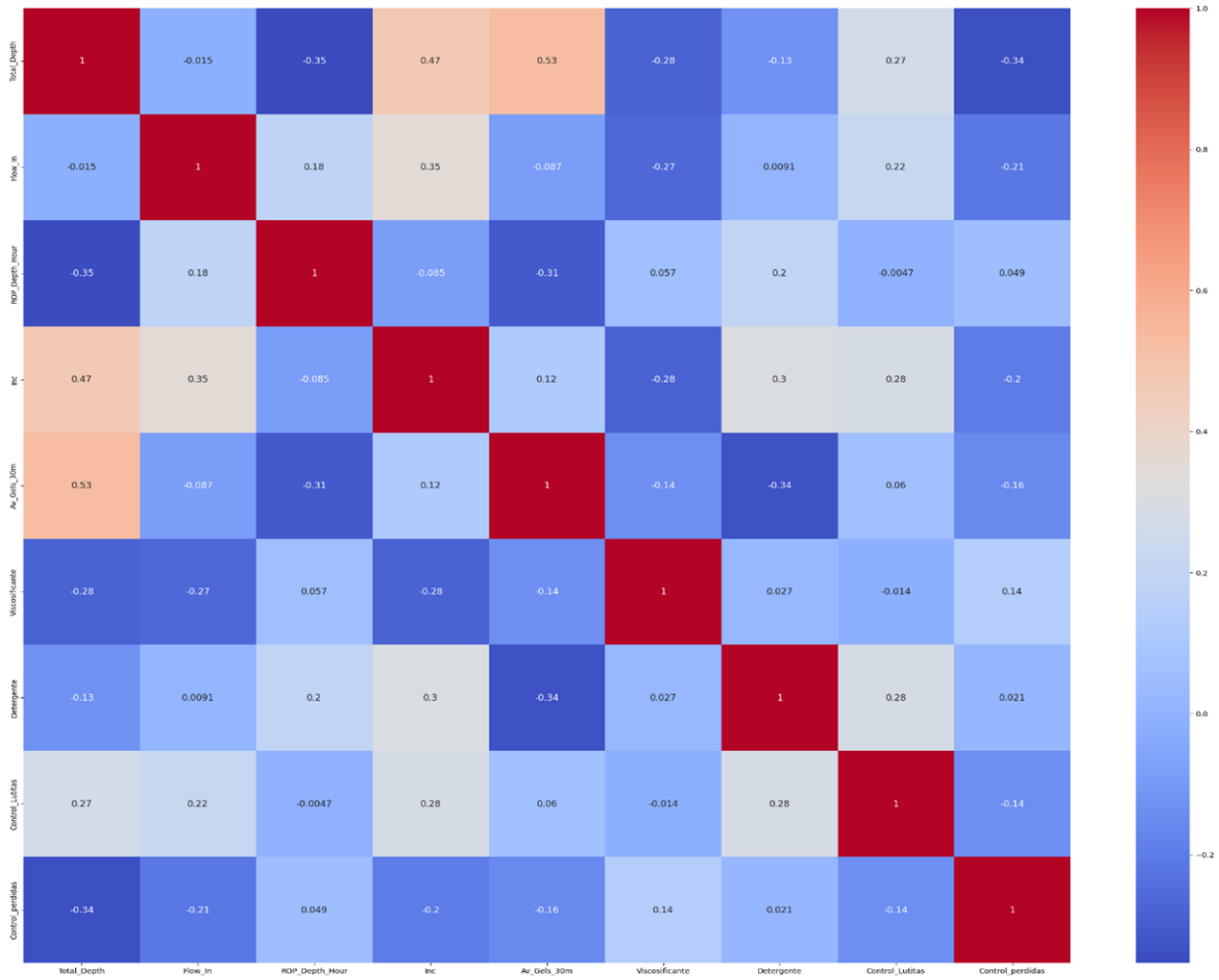

```
Coeficiente de correlación punto biserial: -0.22
Valor de p: 0.0
*********************************************
Distribución de Control_perdidas
*********************************************
count    11183.000000
mean         3.879281
std         11.178888
min          0.000000
25%          0.000000
50%          0.000000
75%          3.000000
max         84.000000
```

Fuente: Autores

Anexo 2

Anexo 2: Análisis de correlación de variables seleccionadas

Figura 15. Análisis de correlación de variables

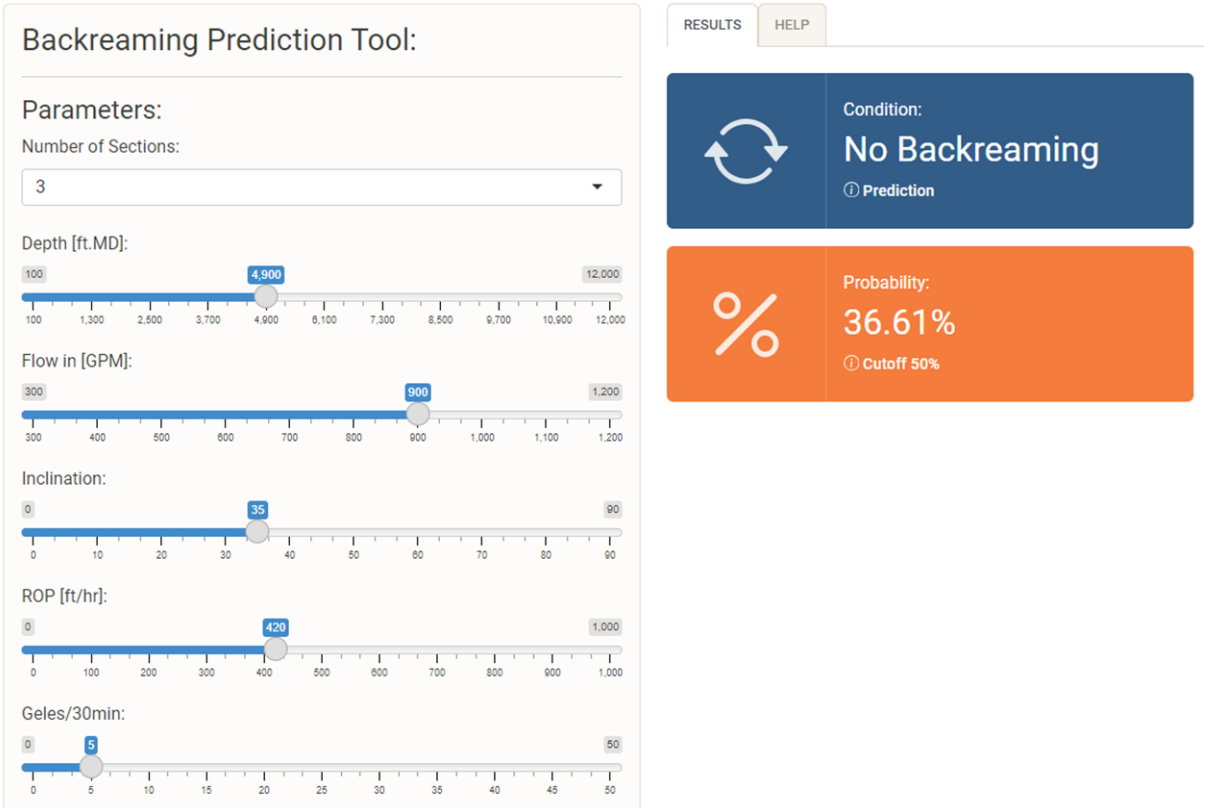


Fuente: Autores

Anexo 3: Resultados de modelos

Modelo	Test		Validación		Total	
	Accuracy	Recall Back	Accuracy	Recall Back	Accuracy	Recall Back
Random Forest	88.6%	87.5%	86.3%	85.7%	88.6%	88.3%
XG Boost	87.3%	85.1%	86.4%	84.5%	87.8%	86.1%
Decision Tree	86.2%	86.8%	85.0%	87.1%	85.6%	86.8%
SVM	85.0%	82.4%	84.2%	83.0%	84.9%	83.2%
Regresión Logistica	79.3%	77.3%	83.1%	78.9%	79.7%	78.0%

Anexo 4 : Backreaming Prediction Tool



Anexo 5: Sensibilidades Backreaming Prediction Tool