ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN GEOLOGÍA Y PETRÓLEOS

PREDICCIÓN DE PROBLEMAS EN EL PROCESO DE PERFORACIÓN DE POZOS PETROLEROS APLICANDO APRENDIZAJE DE MÁQUINA SUPERVISADO

VALIDACIÓN DEL MODELO DE APRENDIZAJE DE MÁQUINA A PARTIR UNA BASE DE DATOS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN PETRÓLEOS

JOEL ANDRÉS LLANO ESPÍN

jandresllano7@gmail.com

DIRECTOR: MARIO LAURO ROBLES REYES

mario.robles@epn.edu.ec

DMQ, agosto 2023

CERTIFICACIONES

Yo, JOEL ANDRÉS LLANO ESPÍN declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

_____

JOEL ANDRÉS LLANO ESPÍN

Certifico que el presente trabajo de integración curricular fue desarrollado por JOEL ANDRÉS LLANO ESPÍN, bajo mi supervisión.

_____

MARIO LAURO ROBLES REYES

DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el producto resultante del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

JOEL ANDRÉS LLANO ESPÍN

DIEGO IVÁN CUZCO YAMASCA

JORGE MATEO ROMERO MUÑOZ

DEDICATORIA

Dedico el resultado de esta investigación a mi familia, quienes fueron apoyo fundamental durante todo este proceso en los buenos y malos momentos. Gracias por su compresión, paciencia y ayuda durante todo este tiempo. A mi abuela que desde muy lejos siempre estuvo pendiente de mí. A mis amigos quienes me acompañaron desde que inicio este largo camino y me ayudaron a cumplir mi objetivo.

Andrés Llano

AGRADECIMIENTO

A Dios, quien ha estado siempre presente y me ha orientado en cada paso de mi vida quien me ha dado la fuerza para seguir adelante y no vencerme en el camino.

A mis padres Adela y Juan, por su ayuda durante el proceso de la universidad, su cariño y confianza que siempre han tenido en mí. Gracias por estar siempre presentes en cada etapa de mi vida.

A mis hermanos Lizbeth, Jhojan y María José, por su cariño, paciencia y ayuda en todos los quehaceres del hogar, por siempre levantarme cuando mi energía se agotaba.

A mi abuela Angelita que a pesar de la distancia siempre me acompañaba con sus oraciones y palabras de aliento las cuales me dan fuerza para seguir siempre hacia adelante.

A EP Petroecuador, por concedernos el acceso a la información para que esta investigación sea posible.

Al Ing. Diego Cuzco, quien desarrolló el papel de profesor y director en este trabajo de investigación. Gracias por su paciencia, ayuda y consejos durante el proceso de esta investigación. Además, gracias por todas las enseñanzas impartidas en el salón de clases.

A mi compañero y amigo, Mateo, por tu apoyo, empatía, esfuerzo, ayuda y tiempo dedicado durante esta investigación. Gracias por tu amistad y por comprenderme en mis momentos más complicados.

A mis amigos y profesores, gracias por compartir de su tiempo, por las enseñanzas, por las risas y hacer de esta etapa universitaria mucho más amena.

Andrés Llano

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

RESUMEN

Los algoritmos de aprendizaje de máquina son utilizados para analizar y procesar grandes conjuntos de información con el objetivo tomar decisiones basadas en datos, realizar predicciones precisas y automatizar tareas complejas en una variedad de campos. La industria petrolera aprovecha esta capacidad de aprender de los datos para optimizar procesos, resolver problemas y obtener información valiosa de manera eficiente. En esta investigación se comprobó el grado de precisión de dos modelos desarrollados con XGBoost a partir de la información recopilada de los reportes diarios de perforación, reportes finales de fluidos de perforación y registros litológicos (*masterlog*[1]*)* correspondientes a 104 pozos del Bloque 60 Campo Sacha. La base de prueba contiene 51 casos distribuidos entre condiciones normales, embolamientos, influjos, pérdidas de circulación, pegas mecánicas/empaquetamientos y taponamientos, con el fin de poner a prueba el rendimiento de los modelos XG Boost 1 y XG Boost 2. En el primer caso, el grado de predicción del modelo fue de 88%, con 6 problemas predichos de forma incorrecta. Por otro lado, el segundo modelo acertó el 80% de los casos y obtuvo 10 equivocaciones. El ajuste de los hiperparámetros influye en la precisión de aprendizaje, pero otros factores como la velocidad y capacidad computacional son críticos para seleccionar el modelo. Por tal motivo, se determinó que el modelo XG Boost 1 es el que mejor se ajusta para predicción de datos reales, no solo porque tiene mayor predicción que el segundo modelo, sino que, adicionalmente, una tasa de aprendizaje mayor fue utilizada para su desarrollo, lo que reduce los recursos del computador.

PALABRAS CLAVE: aprendizaje de máquina, XG Boost, modelo, predicción.

ABSTRACT

Machine learning algorithms are employed to analyze and process extensive sets of information with the goal of making data-driven decisions, achieving accurate predictions, and automating complex tasks across a range of fields. The petroleum industry harnesses this data-driven learning capability to optimize processes, troubleshoot issues, and efficiently gain valuable insights. This research verified the accuracy of two models developed using XGBoost based on the information extracted from daily drilling reports, final drilling fluid reports, and lithological records (masterlog1) from 104 wells in Block 60 Sacha Field. The test dataset comprises 51 cases spanning normal conditions, stuck pipe incidents, influxes, lost circulation, mechanical pack-offs, and blockages, aiming to assess the performance of XG Boost 1 and XG Boost 2 models. In the former case, the model achieved a prediction rate of 88%, with 6 misclassified instances. Conversely, the latter model achieved an 80% accuracy rate, with 10 misclassifications. The adjustment of hyperparameters significantly influences learning accuracy; however, other factors such as computational speed and capacity play critical roles in model selection. Hence, it was concluded that XG Boost 1 is better suited for real data prediction. This preference is based not only on its higher prediction accuracy compared to the second model but also on its use of a higher learning rate, which reduces computational resource demands.

KEYWORDS: machine learning, XG Boost, model, prediction.

1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

El desarrollo de la cuarta revolución industrial ha optimizado los procesos manuales y, por lo tanto, la calidad y velocidad de los resultados dentro de las operaciones en distintas industrias, y la Industria Petrolera no es la excepción (CCOO de Industria, 2017).

Mediante la inteligencia artificial en conjunto con el aprendizaje supervisado y no supervisado se han creado diversas herramientas las cuales ayudan a automatizar los procesos en la industria.

Como es el caso de la detección de problemas de pega de tubería, predicción de ROP, problemas de pérdida de circulación, entre otros. El primer acercamiento que integró aprendizaje supervisado fue realizado por (Ubillus & Pacheco, 2021), sin embargo, la muestra utilizada para esta investigación es pequeña y se empleó otro algoritmo (K-Nearest Neighbors).

En este ámbito, se utilizará una muestra de 100 pozos y se generará una base de prueba para realizar la etapa de validación del código el cual implementa el algoritmo XGBoosting. Con el código generado será posible identificar los problemas operacionales que ocurren en la etapa de perforación, con el fin de conocer la incertidumbre con la cual se generó el modelo.

1.1. Objetivo general

Predecir los problemas operacionales en el proceso de perforación de pozos petroleros aplicando aprendizaje de máquina supervisado, para optimizar los tiempos de perforación a través de la reducción de tiempos no productivos (NPT) generados por dichos problemas.

1.2. Objetivos específicos

1. Generación de la base de datos de prueba.
2. Validar la información de los pozos seleccionados.
3. Aplicar el aprendizaje de máquina supervisado XGBoost.
4. Analizar el desempeño del código para conocer si cumple con un grado de precisión de al menos el 75% en la detección de problemas de los pozos de prueba.

1.3. Alcance

Se realizará la validación de los datos que serán recolectados de los reportes finales de perforación de al menos 30 pozos (70 pozos serán utilizados en la etapa de entrenamiento del código) y se proporcionará los parámetros necesarios en el código elaborado en Python, para realizar una evaluación y verificación del funcionamiento adecuado del programa. Con los resultados obtenidos se realizará un análisis comparativo de los problemas operacionales definidos por el código y los problemas que tuvieron lugar en la sección analizada. Mediante esta comparación se determinará la incertidumbre real del modelo.

Es necesario enfatizar que el análisis de los problemas operacionales no se realiza con datos obtenidos a tiempo real. Sin embargo, se analizará si es posible su aplicación en tiempo real.

1.4. Marco teórico

El desarrollo de la cuarta revolución industrial ha optimizado los procesos manuales y, por lo tanto, la calidad y velocidad de los resultados dentro de las operaciones en distintas industrias, y la Industria Petrolera no es la excepción (CCOO de Industria, 2017).

La transformación digital disminuirá los gastos en tiempo de operación de los equipos, plataforma, horas de mano de obra, error humano, vida útil de las herramientas y recursos utilizados (Alsheikh, 2022). El aprendizaje automático, los sensores y la robótica podrían usarse para evaluar y estudiar el pozo, sus formaciones y los parámetros de operación para optimizar la perforación. Las predicciones acertadas permitirán a los ingenieros tomar mejores decisiones mientras trabajan.

En el área de aprendizaje de máquina, las contribuciones más relevantes relacionadas con la detección y solución de problemas en la perforación se muestran en la Tabla 1, las cuales servirán de punto de partida y guía para desarrollar la investigación e implementación de un modelo que permita predecir los posibles problemas durante las operaciones. Adicionalmente, se incluye bibliografía donde se detallan los principales problemas operacionales, sus síntomas y las posibles soluciones.

Tabla 1: Estudios más relevantes que aportan al desarrollo del Trabajo

Autor y Año	Párametros de Entrada	Finalidad / Contribución
(Rabia, 2002)	-	Análisis de los problemas operacionales durante la
(Lake & Mitchell, 2006)		

Autor y Año	Párametros de Entrada	Finalidad / Contribución
(Azar & Robello Samuel, 2007)		perforación, sus síntomas y soluciones
(Hossain & Islam, 2018)		
(Ubillus & Pacheco, 2021)	-	Predicción de problemas operacionales durante perforación con método vecinos más cercanos
(Noshi & Schubert, 2019)	Velocidad de rotación, SPP[1], MD[2], altura del bloque, torque, peso promedio en el gancho, gamma ray, pérdida de presión en el anular, profundidad de la broca, volumen de lodo, WOB[3], datos del desplazamiento positivo del motor.	Predecir ROP mediante la aplicación de los algoritmos RF, ANN[4], GBM[5], *Support Vector Regression, Ridge Regresion*
(Encinas, Tunkiel, & Sui, 2022)	MD, peso promedio en el gancho, WOB, torque, ROP[6], RPM[7], SPP	Predecir ROP mediante la aplicación de los algoritmos RF, RNN (main), *XG-Boost, gradient boosting regresor*
(Li & Samuel, 2019)	WOB, torque, RPM en fondo y superficie	Predecir ROP mediante la aplicación del algoritmo ANN
(Chen, Yu, Shen, & Zhengxin Zhang, 2019)	Cargas en el gancho, SPP, RPM, torque, caudal	Identificar pega de tubería mediante los algoritmos DT, SVM, ANN
(Bayan & Zulkarnain, 2020)	19 parámetros de perforación (ROP, RPM, WOB, torque entre otros)	Identificar pega de tubería mediante el algoritmo ANN
(Elmousalami & Elaskary, 2020)	Caudal, tipo de lodo, tiempo total de perforación, ROP, inclinación máxima.	Identificar pega de tubería mediante los algoritmos ANN, DT, XGBoost.

[1] *Stand Pipe Pressure:* es la pérdida de carga total por fricción en el circuito hidráulico.
[2] *Measured Depth:* es la longitud del hoyo perforado.
[3] *Weight On Bit:* cantidad de peso descendente ejercido sobre la broca.
[4] *Artificial Neural Netwok*: debido a su capacidad de aproximación universal y a su estructura flexible, permiten captar comportamientos no lineales complejos
[5] *Gradient Boosting Machine*: método para convertir a los aprendices débiles en aprendices fuertes.
[6] *Rate of Penetration:* es la velocidad a la que una broca rompe la roca para perforar el pozo.
[7] *Revolutions per Minute:* Indica la velocidad a la cual está girando el motor.

Autor y Año	Párametros de Entrada	Finalidad / Contribución
(Hou, y otros, 2020)	Datos de perforación (MD, WOB, RPM, SPP, ROP), Datos Geológicos (litología, presión de fractura, presión de poro), Datos de fluidos de perforación (MW, PV, YP, contenido de sólidos)	Predecir pérdida de circulación mediante el algoritmo ANN
(Sabah, Talebkeikhah, Agin, Talebkeikhah, & Hasheminasab, 2019)	Profundidad, parámetros de perforación (WOB, caudal, RPM entre otros), propiedades del lodo (viscosidad, gel strength, porcentaje de sólidos, entre otros), litología, trayectoria del pozo	Predecir pérdida de circulación mediante los algorítmos ANN y DT
(Alkinani, Al-Hameedi, & Dunn-Norman, 2020)	Propiedades de lodo (MW, ECD, PV, YP), RPM, WOB, caudal, área de flujo total de la tobera	Predecir pérdida de circulación mediante el algoritmo ANN
(Hou, y otros, 2019)	ROP, RPM, SPP, peso promedio en el gancho, WOB, entrada y salida de caudal	Detectar en tiempo real un influjo de gas, mediante el algoritmo BPNN[8] con PCA
(Yang, y otros, 2019)	Profundidad, peso promedio en el gancho, WOB, RPM, torque, entre otros	Detectar un influjo con el algoritmo ANN con PCA
(Shi, y otros, 2019)	Flujo de entrada y salida, Presión en el anular, temperatura en el anular, peso promedio en el gancho entre otras.	Detectar un influjo con los algoritmos RF[9] y SVM[10]
(Shadravan, Tarrahi, & Aman, 2017)	MW, volumen del aditivo, y temperatura	Predicción de la viscosidad el fluido de perforación con el algoritmo GPR[11] y ANN
(Kuesters, Mason, Gomes, Cockburn, & Lodhi, 2020)	Peso promedio en el gancho, WOB, RPM, ROP, torque, caudal, SPP	Detección de cavernas durante la perforación
(Okoli, Cruz Vega, & Shor, 2019)	Torque, ROP, WOB	Predicción de vibración de la sarta con algoritmo KNN[12],

[8] *Back Propagation in Neural Networks*: técnica que aún se utiliza para entrenar grandes redes de aprendizaje profundo.

[9] *Random Forest:* realizan predicciones de salida combinando los resultados de una secuencia de árboles de decisión de regresión.

[10] *Support Vector Machine:* puntos que están cercanos al hiperplano e influye en la posición y orientación del mismo.

[11] *Gaussian Process Regression:* Clase notablemente potente de algoritmos no paramétricos de aprendizaje automático para tareas de aprendizaje supervisado.

[12] *K Nearest Neighbors:* Es método que busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de los datos que le rodean.

Autor y Año	Párametros de Entrada	Finalidad / Contribución
		Regresión logística, DT[13], Naive Bayes
(Zha & Pham, 2018)	Torque, tensión, RPM, WOB	Predicción de vibración de la sarta con "deep learning"

Elaborado por: Los autores

A lo largo de los años, varias áreas en la industria petrolera han sido ampliamente estudiadas y enfocadas a la detección de problemas y solución de los mismos. Tales áreas incluyen Levantamiento Artificial, Recuperación Mejorada, Monitoreo de la Producción, Caracterización de yacimiento, entre otras.

En la Figura 1 se observa que la mayor área de aplicación de aprendizaje de máquina es en la perforación. Dependiendo del caso, varios algoritmos han sido utilizados para crear modelos, mostrando diferentes resultados en la precisión de predicción, el cual dependerá de los objetivos de la investigación y el volumen de información utilizado para el desarrollo de los mismos. En la presente investigación, se implementará el aprendizaje supervisado de maquina mediante el modelo XG Boosting, enfocado directamente con los problemas operacionales que ocurren durante el proceso de perforación a partir de la información de 100 pozos donde se incluyen los reportes diarios de perforación y reportes finales de fluidos de perforación.

[13] *Decision Tree:* Método utilizado para clasificación y regresión.

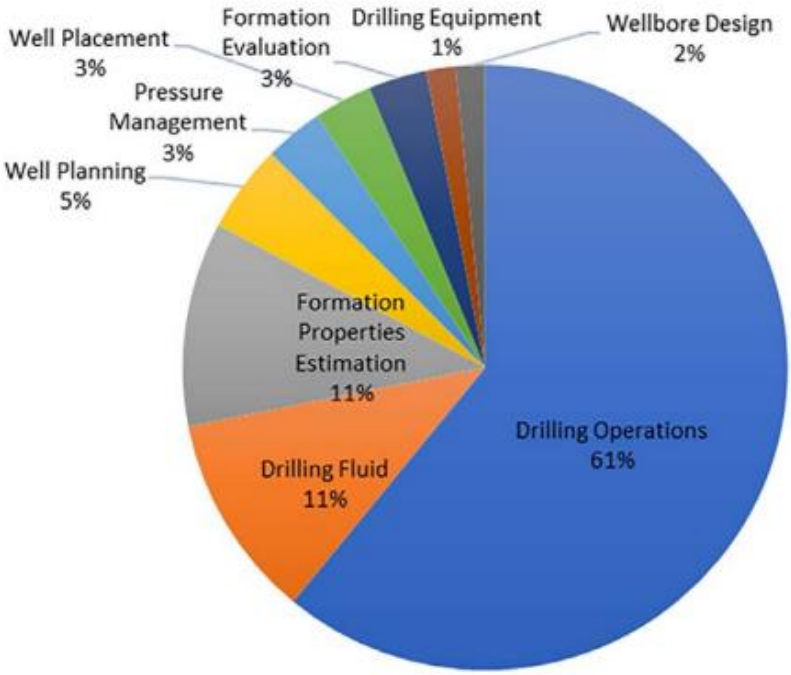

Figura 1: Porcentaje de aplicación de aprendizaje de máquina en la industria petrolera.

Fuente: (Olukoga & Feng, 2021)

El aprendizaje de máquina ha transformado la perforación al brindar la capacidad de análisis avanzado de volúmenes de datos, optimización y predicción de parámetros y problemas en tiempo real. Estas mejoras tienen el potencial de aumentar la eficiencia, la seguridad y la rentabilidad en la industria de la perforación. Sin embargo, con el fin de entender completamente la relación que existe entre el aprendizaje de máquina y la perforación, es importante conocer los conceptos fundamentales en esta área. Los mismos se detallan en el siguiente apartado.

1.4.1 Perforación en Tierra

La perforación es una operación fundamental en la industria petrolera. Su objetivo es comunicar el reservorio con la superficie a través del hoyo perforado, para así producir el hidrocarburo que se encuentra dentro del yacimiento (pozo productor) o inyectar un fluido que mejore la producción del mismo (pozo inyector), entre otras utilidades. (Azar & Robello Samuel, 2007) indica que la perforación se realiza de forma telescópica invertida, es decir, se utiliza una broca de mayor a menor diámetro con las cuales se construye las diferentes secciones, incluyendo: sección de hoyo conductor, sección superficial, sección intermedia y sección de producción, las mismas que se revisten con tuberías de distinto tamaño.

En la Tabla 2 se presentan las brocas y tuberías de revestimiento que son comúnmente utilizadas en Ecuador:

Tabla 2: Brocas y Tuberías de Revestimiento utilizadas en el Ecuador

Sección	**Broca [pg]**	**Tubería de Revestimiento [pg]**
Conductor	26	20
Superficial	16	13 3/8
Intermedio	12 1/4	9 5/8
Producción	8 1/2	7

Elaborado por: Los autores

Las secciones mostradas anteriormente atraviesan diferentes formaciones y profundidades.

- *Casing conductor:* de acuerdo con Cuzco & Ortiz (2013, pág. 30) está asentado a profundidades someras entre 150-200 pies atravesando los boulders de las formaciones Mera/Mesa. En ocasiones suele omitirse la construcción de esta sección siempre y cuando se maneje adecuadamente la hidráulica durante la perforación (incremento gradual de caudal), pues valores altos presión podrían llegar a fracturar las formaciones someras.
- *Casing superficial:* Jaramillo (2018, pág. 27) indica que es comúnmente asentado hasta la formación Orteguenza atravesando las formaciones Chambira, Curaray, Arajuno, Chalcana.
- *Casing intermedio:* Cuzco & Ortiz (2013, pág. 140) mencionan que se suele asentar hasta el tope de Basal Tena atravesando las formaciones de Orteguaza, Tiyuyacu, Tena superior y Tena inferior.
- *Casing de producción:* Rabia (2002, pág. 107) sostiene que esta tubería de revestimiento se asienta hasta la formación productora. En Ecuador las arenas productoras varían dependiendo del campo. Sin embargo, de forma general se incluyen Basal Tena, M1, M2, Napo U, Napo T, Hollín superior y Hollín inferior. (Baby, Rivadeneira, & Barragán, 2014).

La Figura 2 muestra las formaciones mencionadas, su edad geológica, su litología principal y el evento tectónico ocurrido para su desarrollo.

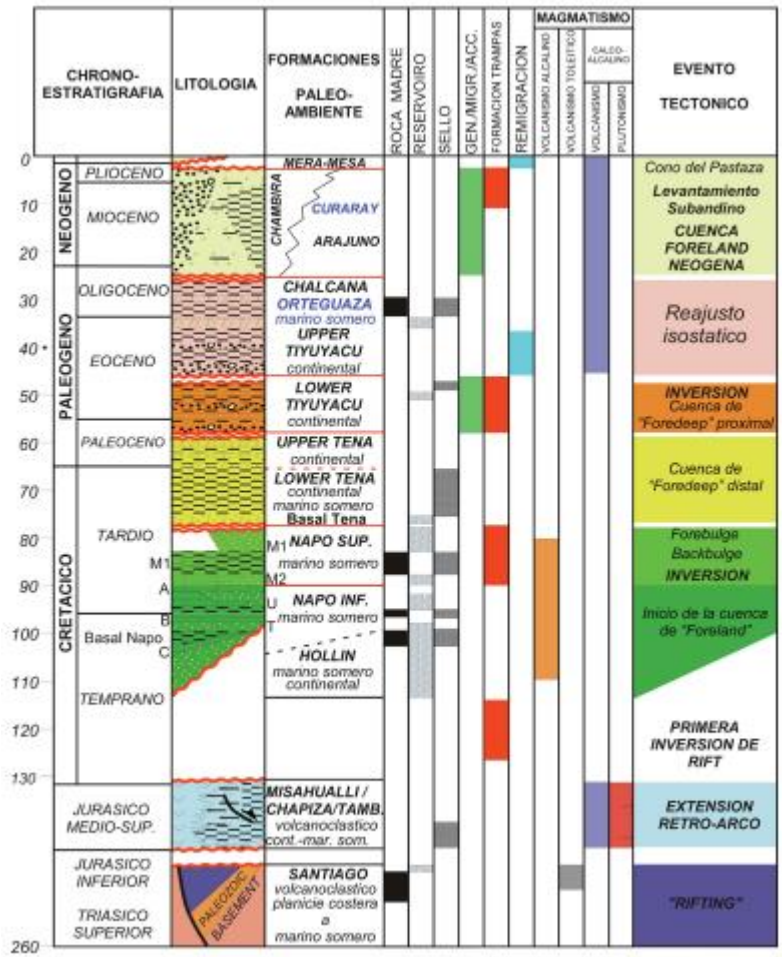

Figura 2: Columna estratigráfica Cuenca Oriente.

Fuente: (Baby, Rivadeneira, & Barragán, 2014).

(Cuzco & Ortiz, 2013, pág. 10) indican que la información litológica permitirá desarrollar programas adecuados en las diferentes disciplinas incluidas en el proceso de la perforación (ej. cementación, direccional, fluidos) en función de las condiciones geológicas esperadas para tomar medidas que ayuden a optimizar su proceso. No obstante, no es el único parámetro del que depende llegar de forma adecuada al objetivo. Según (Baberli, 1998, pág. 92) para ejecutar y asegurar una perforación exitosa, es necesario que los sistemas del taladro de perforación funcionen adecuadamente. Entre estos están:

- Sistema de Izaje
- Sistema Rotatorio
- Sistema de Circulación
- Sistema de Generación de Energía

- Sistema de Control o Prevención de Reventones

El correcto dimensionamiento de los componentes (ej. potencia del equipo, capacidad de carga) de cada uno de los sistemas evitará posibles problemas durante la perforación.

1.4.1.1 Fluidos de perforación

Es también conocido como lodo de perforación. Es un fluido compuesto por una fase continua (agua, aceite o gas) y una fase discontinua (sólidos). Este fluido es bombeado a través de la sarta de perforación hasta la broca, donde entra en contacto con el fondo y las paredes del pozo y retorna a superficie por el espacio anular. (Lake & Mitchell, 2006, pág. 89) señala que el fluido de perforación es el único componente del proceso de construcción de pozos que permanece en contacto con las formaciones atravesadas por el pozo durante toda la operación de perforación. Por lo tanto, resulta de suma importancia que la selección de las propiedades y tipo de lodo sea adecuada para garantizar que la operación de cada una de las distintas secciones sea segura y eficiente. Para ello, el lodo debe cumplir con las siguientes funciones:

- Refrigerar y lubricar la broca y la sarta de perforación
- Limpiar el fondo del pozo
- Transportar los recortes de perforación a superficie
- Controlar presiones de fondo
- Crear costra de lodo
- Transmitir potencia hidráulica a la broca
- Proveer de estabilidad al pozo
- Garantizar la evaluación de las formaciones

(Rabia, 2002, pág. 275) indica que cualquier problema causado por el incumplimiento en las funciones del fluido de perforación puede resultar extremadamente costoso en materiales y tiempos no productivos (ej. daño de equipos, retrasos en la perforación). Adicionalmente, la falta de cumplimiento de los estándares de seguridad de un correcto lodo de perforación puede comprometer el éxito de las operaciones, con posibles consecuencias negativas, como la pérdida del pozo o la incapacidad para alcanzar el objetivo geológico previsto. De manera adicional, estas fallas podrían derivar en

situaciones más graves, como reventones, que representarían un riesgo significativo para la integridad y seguridad del personal a cargo del taladro.

Por otro lado, (Lake & Mitchell, 2006, pág. 96) indican que es imperativo la correcta utilización de aditivos para monitorear las propiedades del fluido de perforación en función de la sección que se está perforando. Estas propiedades se enumeran a continuación:

- Densidad
- Viscosidad
- Punto de cedencia
- pH
- Dureza
- Fuerza del gel
- Contenido de sólidos
- Reología

Luego de abordar la importancia de los lodos de perforación en el proceso de perforación, es imperativo analizar de forma detallada y específica los diversos problemas que pueden surgir durante la perforación de pozos. Estos desafíos, aunque complejos, son parte integral de la exploración y producción de hidrocarburos y requieren una comprensión minuciosa para garantizar el éxito operativo y la seguridad del personal.

1.4.1.2 Problemas Operacionales durante la Perforación en Tierra

En la industria petrolera, comúnmente presentan problemas y riesgos en sus diferentes etapas (ej. Exploración, producción, transporte y refinación), y la perforación no es la excepción. Durante las operaciones de perforación, algún tipo de problema ocurrirá con certeza, sin importar que se utilicen los mejores equipos, personal capacitado, los mejores materiales y el mejor programa de perforación (Baberli, 1998, pág. 133).

Hossain e Islam (2018, pág. 12) sugieren que durante la planificación, la clave para alcanzar los objetivos geológicos con éxito es diseñar los programas de perforación en base a la anticipación de los problemas potenciales en lugar de la precaución y contención. Por otro lado, durante las operaciones, mientras más efectiva sea la detección de los síntomas, más sencillo será identificar el problema y sugerir una solución. Por lo tanto, la

optimización durante la planificación y las operaciones de perforación proviene de una correcta predicción de los problemas.

Algunos de los problemas operacionales más comunes durante la perforación, los indicadores y sus posibles soluciones se incluyen en la Tabla 3:

Tabla 3: Problemas y Posibles soluciones en la perforación de pozos.

Problemas			**Indicadores**	**Soluciones**
Pega de tubería	Mecánica	Falta de limpieza del hoyo (empaquetamiento)	- Aumento de torque y arrastre - Incremento de la densidad equivalente de circulación - Decremento de la tasa de penetración - Poca o nula circulación - Incremento en la presión de la bomba	- Incremento del Punto de cedencia y viscosidad del lodo - Aumentar caudal - Circular y reciprocar la sarta - Aplicar tren de píldoras de limpieza (píldora dispersa y viscosa)
		Inestabilidad de lutita	- Aumento de la viscosidad plástica - Aumento de presión de circulación - Aumento de arrastre - Torque errático - Poca o nula circulación de lodo	- Reciprocar en casos extremos - Adicionar inhibidores de arcilla - Incrementar salinidad del fluido de perforación (detiene hidratación de arcillas[14]) -Incrementar lubricidad del lodo
	Diferencial		- Incremento de torque y arrastre - Inhabilidad para reciprocar y rotar la sarta - Circulación no es interrumpida	- Reducir peso de lodo - Lavar con aceite sobre la tubería atascada
Pérdida de Circulación			- Reducción de flujo de lodo en superficie - ROP aumenta - Torque alto y errático -Baja el nivel de los tanques de retorno	- Bombear lodo con aditivo de control de pérdida de filtrado - Sellar la zona con cemento o tapones - Disminuir el caudal - Disminuir el peso del lodo

[14] La adición de sal en el lodo de perforación provoca que el agua presente en las arcillas se desplace hacia el fluido de perforación con mayor concentración de sal. Como resultado, las partículas de arcilla retienen menos agua y dejan de hidratarse de manera significativa.

Problemas	Indicadores	Soluciones
Influjos	- Decremento de presión en la bomba - Incremento de nivel en los tanques - Aumento del peso sobre el gancho - Incremento abrupto del ROP - Incremento del caudal de flujo	- Circular influjo fuera del pozo con método de control de pozo - Desplazar el lodo liviano con un lodo más pesado
Broca embolada	- Incremento de presión - ROP nula o reducida - Torque errático	- Agitar la broca - Bombear píldora dispersa - Utilizar detergente, inhibidores de arcilla y pequeño porcentaje de glicol para dispersar arcilla pegada

Fuentes: (Azar & Robello Samuel, 2007), (Hossain & Islam, 2018), (Lake & Mitchell, 2006) & (Rabia, 2002).

Entender correctamente los síntomas asociados a diferentes problemas operacionales permite a los ingenieros llevar a cabo un diagnóstico preciso, posibilitando la implementación rápida de medidas correctivas y preventivas para enfrentar dicho problema. De forma adicional, la aplicación de términos estadísticos y algebraicos a las variables que gobiernan la ocurrencia de un problema puede proporcionar información valiosa sobre patrones y tendencias asociadas, permitiendo predecir su ocurrencia futura. Los términos estadísticos más relevantes para esta investigación serán abordados en la siguiente sección.

1.4.2 Descripción de términos Estadísticos y Algebraicos en el Aprendizaje de Máquina

(Johnson, 2012, pág. 1) indica que la estadística engloba la recolección, el procesamiento, el análisis y la interpretación de datos; con los cuales es posible realizar cálculos, seleccionar modelos y hacer predicciones. Entre ellos, la recolección de información es probablemente el trabajo más importante dado que mientras más grande sea el volumen de información y la misma se encuentre validada, el modelo generado presentará mejores resultados.

Por lo tanto, el mismo autor sugiere que para la correcta aplicación de la estadística, se deben tomar en cuenta los siguientes cuatro pasos:

1. Establecer y definir metas para la investigación
2. Determinar qué información será necesaria y cómo se va a recolectar
3. Aplicar métodos estadísticos apropiados para una extracción eficiente de los datos
4. Interpretar la información y extraer conclusiones

En este proyecto se tomará en cuenta tanto la estadística descriptiva como la estadística inferencial. Por un lado, mientras la estadística descriptiva permite conocer la relación entre las variables que están directamente involucradas en un problema operacional a partir de gráficas, la estadística inferencial permite conocer el grado de relación entre dichas variables para para posteriormente realizar predicciones.

(Walpole , Myers, Myers, & Ye, 2007, pág. 1) explican que a estadística descriptiva permite conocer el sentido de la ubicación de los datos, su variabilidad y la naturaleza general de su distribución. Usualmente, va acompañada de gráficas (ej. histogramas, diagramas de caja, gráficos de dispersión). Por otro lado, la estadística inferencial permite obtener conclusiones acerca de las características de los datos hipotéticos que se tomen de la población con base a cálculos probabilísticos (ej. intervalos de confianza, chi cuadrado, proporciones). Este tipo de estadística se aplica en los modelos predictivos, técnicas de aprendizaje de máquina e inteligencia artificial.

Finalmente, para el entendimiento de la metodología que desarrollará el presente trabajo, es necesario conocer términos estadísticos adicionales, los cuales se detallan a continuación:

1.4.2.1 Curtosis

La curtosis es una medida que sirve para analizar el grado de concentración que presentan los valores de una distribución alrededor de su media (Spiegel & Stephens, 2009, pág. 125).

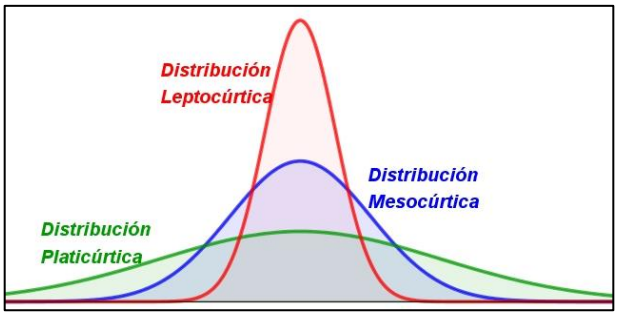

Figura 3: Curva de Curtosis

Fuente: Colón, A. & Christersson, M. (2022)

El coeficiente de curtosis permite explicar lo mencionado anteriormente de forma matemática, el cual se presenta en la Ecuación 1:

$$b_2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x_n})^4 / Sn^4$$

Ecuación 1: Coeficiente de Curtosis

Como la suma de las cuartas potencias es siempre positiva, b2 ≥ 0. Por otro lado, la curtosis de la distribución normal estándar es 3. Esto significa que, independientemente de si los datos son discretos o continuos, podríamos restar 3 para obtener $\gamma_2 = \beta_2 - 3$ como medida estandarizada de la curtosis, tal como sugiere Fisher. La Ecuación 2 se describe a continuación (Shanmugam & Chattamvelli, 2015, pág. 119):

$$\gamma_2 = \frac{1}{n}\frac{\sum_{j=1}^{n}(x_j - \overline{x_n})^4}{Sn^4} - 3$$

Ecuación 2: Medida Estandarizada de la Curtosis

Donde:

- γ_2: Medida estandarizada de Curtosis
- n: es el número total de datos
- x_j: es la marca de clase del grupo i-ésimo
- $\overline{x_n}$: es la media aritmética de la distribución
- Sn^4: es la desviación estándar (o desviación típica) de la distribución.

Con la medida estandarizada de la curtosis se puede comparar con el gráfico obtenido, es decir:

- $\gamma_2 > 0$, indica distribuciones leptocúrticas[15]
- $\gamma_2 < 0$, indica distribuciones platicúrticas[16]

1.4.2.2 Asimetría

La asimetría o sesgo es la desviación de la distribución de los valores de una variable alrededor de su media o promedio en un conjunto de datos. Si la curva se desplaza hacia la izquierda o hacia la derecha, se dice que está sesgada (Figura 4).

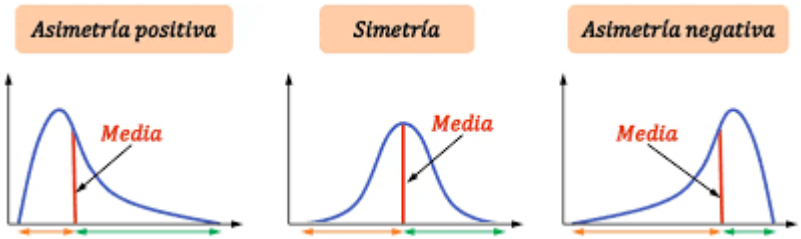

Figura 4: Curva de Asimetría

Fuente: (Juárez, 2022)

De acuerdo con (Spiegel & Stephens, 2009, pág. 125) se pueden observar distintas características en las asimetrías. En el caso de datos sesgados positivamente, la media de los datos será mayor que la mediana. En el caso de datos sesgados negativamente, la media será menor que la mediana.

Por otro lado, la moda de los datos sesgados de forma positiva será menor que la mediana y la media ($M_d<M_e<\bar{X}$), mientras que, en el caso de los datos sesgados de forma negativa, la moda será mayor que la mediana y la media ($M_d>M_e>\bar{X}$).

$$Sesgo = \frac{\bar{X} - moda}{s}$$

Ecuación 3: Sesgo a partir de la moda

$$Sesgo = \frac{3 * (\bar{X} - mediana)}{s}$$

Ecuación 4: Sesgo a partir de la mediana

[15] Leptocúrtica: Es aquella que tiene una mayor concentración de datos cerca de la media y colas más delgadas en comparación con la distribución normal.
[16] Platicúrtica: Se caracteriza porque se muestran colas más anchas y una menor concentración de frecuencias alrededor de la media.

Donde:

$\bar{X}$: media de los datos obtenidos

s: desviación estándar de los datos obtenidos

1.4.2.3 Análisis de Componentes Principales

También conocido como PCA (Principal Component Analysis), es un enfoque ampliamente utilizado para reducir la dimensionalidad de conjuntos de datos extensos. Su objetivo es transformar un conjunto de variables en uno más compacto, preservando la mayor parte de la información presente en el conjunto original (Sircar, Yadav, Rayavarapu, Bist, & Oza, 2021).

Es importante tener en cuenta que, al disminuir el número de variables en un conjunto de datos, se produce un ligero sacrificio en términos de precisión. Sin embargo, la reducción de dimensionalidad busca obtener la simplicidad como contrapartida a esta pérdida. Al disponer de conjuntos de datos más pequeños, se facilita la exploración y visualización de los mismos, agilizando el análisis de los puntos de datos y permitiendo un procesamiento más eficiente para algoritmos de aprendizaje de máquina al no tener que lidiar con variables superfluas (Richardson, 2009, pág. 2).

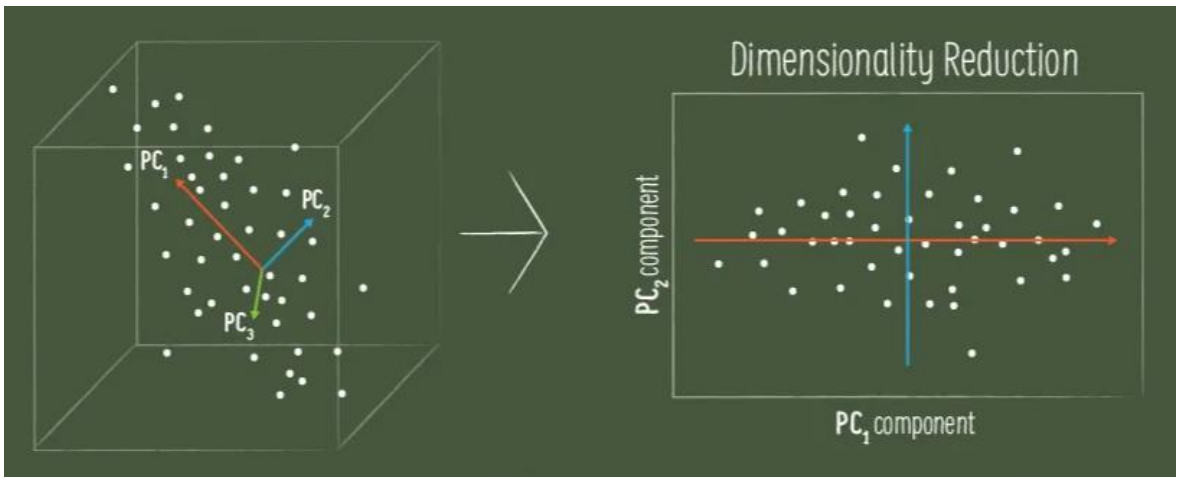

Figura 5: Análisis de Componentes Principales

Fuente: (Sharma, 2019)

Los componentes principales son variables nuevas que se generan a partir de combinaciones lineales o mezclas de las variables originales. Estas combinaciones se realizan de manera que los componentes principales resultantes no están relacionados entre sí, y se concentra la mayor cantidad de información de las variables originales en los primeros componentes. En resumen, la idea es que, aunque los datos tengan 10 dimensiones, el análisis de componentes principales busca maximizar la información en el

primer componente, luego en el segundo, y así sucesivamente, hasta lograr una representación más compacta.

1.4.3 Descripción de términos para el desarrollo del modelo de Aprendizaje de Máquina

En esta sección, se detallan los términos asociados al aprendizaje de máquina de manera gradual y accesible, permitiendo una comprensión sólida de los fundamentos utilizados en el desarrollo del modelo propuesto para la presente investigación.

1.4.3.1 Clasificación

La clasificación es una técnica de aprendizaje automático supervisado en la cual el modelo se esfuerza por predecir la etiqueta correcta de un conjunto de datos de entrada determinado. En el proceso de clasificación, el modelo se entrena utilizando un conjunto de datos de entrenamiento etiquetados, y posteriormente se evalúa en un conjunto separado de datos de prueba para evaluar su rendimiento. Esta evaluación permite que el modelo realice predicciones precisas cuando se enfrenta a datos nuevos que no se han visto (Keita, 2022).

En contraste, la clasificación multiclase implica la presencia de al menos dos etiquetas de clase excluyentes entre sí, donde el propósito es predecir a cuál clase pertenece un ejemplo de entrada dado. La mayoría de los algoritmos de clasificación binaria también son aplicables a la clasificación multiclase. Estos algoritmos incluyen, pero no se limitan a, los siguientes:

- Random Forest
- Naive Bayes
- K-Nearest Neighbors
- Gradient Boosting
- Extreme Gradient Boosting

1.4.3.2 Árboles de Decisión

(Hernandez, 2022, pág. 28) menciona que un árbol de decisión es una estructura jerárquica en forma de árbol invertido que se asemeja a un diagrama de flujo. Los modelos basados en árboles se componen de una o más sentencias que se basan en características específicas para dividir los datos. Dentro de estas divisiones, se utiliza un modelo para

predecir los resultados. Los componentes principales de este modelo son los nodos, las ramas y las hojas. Cada nodo interno representa la evaluación de una característica, cada rama representa el resultado de dicha evaluación, y cada hoja contiene la etiqueta correspondiente a una clase.

Dentro de los nodos, se distingue el nodo superior o raíz, que representa el conjunto de datos completo. Este nodo representa la primera división que se realiza en el conjunto de datos. Durante el proceso de entrenamiento en los árboles de decisión, las muestras presentes en cada nodo interno o nodo de decisión se subdividen en subconjuntos basados en un atributo específico. El objetivo final es crear un modelo predictivo capaz de tomar observaciones sobre una muestra y hacer conclusiones precisas sobre el valor objetivo de esa muestra. Este proceso se repite de forma recursiva en cada subconjunto derivado, en un enfoque conocido como "partición recursiva" (Figura 6).

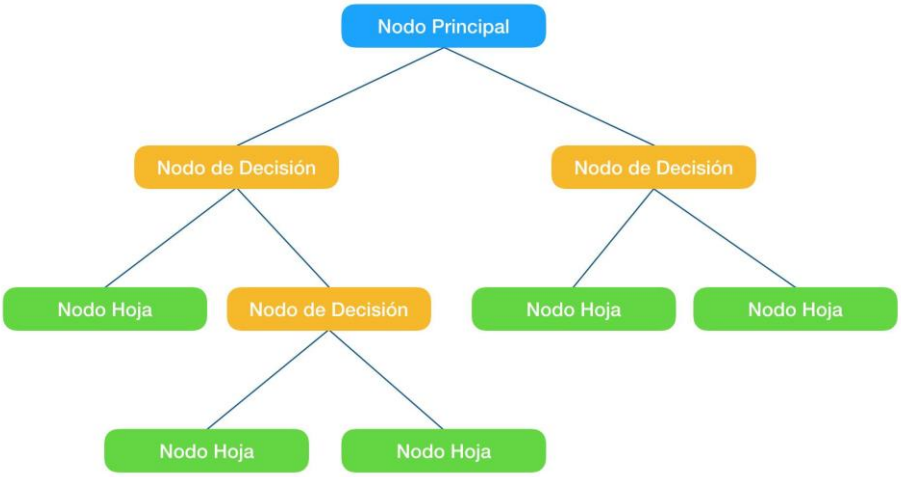

Figura 6: Estructura de un Árbol de Decisiones

Fuente: (Rodriguez V. , 2018).

1.4.3.3 Extreme Gradient Boosting

Extreme Gradiente Boosting, también llamado XGBoost, es una mejora del algoritmo Gradient Boosting. La principal diferencia radica en que XGBoost utiliza un modelo más regularizado, lo que ayuda a evitar el sobreajuste. XGBoost trabaja al combinar varios modelos de aprendizaje débiles para formar un modelo fuerte. Un modelo débil es uno que tiene un rendimiento ligeramente mejor que adivinar al azar. Sin embargo, cuando se combinan estos modelos débiles, pueden formar un modelo fuerte que es mucho más preciso (Mahendra, 2023).

XGBoost funciona entrenando varios árboles de decisión. Cada árbol se entrena con una porción de los datos y las predicciones de cada árbol se combinan para formar la predicción final. (Chen & Guestrin, 2016, pág. 1) indican que, en lugar de construir un solo árbol de decisión, XGBoost construye una serie de árboles en secuencia, donde cada árbol intenta corregir los errores del árbol anterior. En cada iteración, se calculan los gradientes y se ajusta el modelo para minimizar la función de pérdida. Esto permite que el modelo se enfoque en los casos más difíciles y mejore gradualmente su rendimiento. Por esta razón, XGBoost permite manejar datos atípicos y de características complejas.

1.4.3.4 Matriz de Confusión

Es una herramienta fundamental en la evaluación de modelos de clasificación en aprendizaje automático. (Ting, 2011, pág. 209) indica que esta matriz representa la eficacia del modelo al comparar las predicciones realizadas con las clases reales del conjunto de datos. Es una tabla que muestra el número de aciertos y errores para cada clase de salida, lo que permite identificar la capacidad de discriminación del modelo y su habilidad para clasificar correctamente las distintas clases.

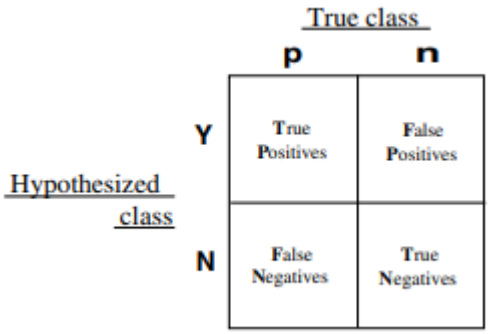

Figura 7: Matriz de Confusión

Fuente: (Fawcett, 2005)

1.4.3.5 Mapa de Calor

Según (Wilkinson & Friendly, 2012) es uno de los métodos de gráfica multivariante para mostrar las relaciones entre dos o más conjuntos de datos. Esta gráfica utiliza colores para mostrar la densidad de una variable en una determinada área geográfica o visual. Mediante el uso de un gradiente de color, representa diferentes niveles de concentración de la variable en cuestión. Estos mapas son muy útiles para visualizar grandes conjuntos de datos, identificar patrones o tendencias en los datos (Figura 8).

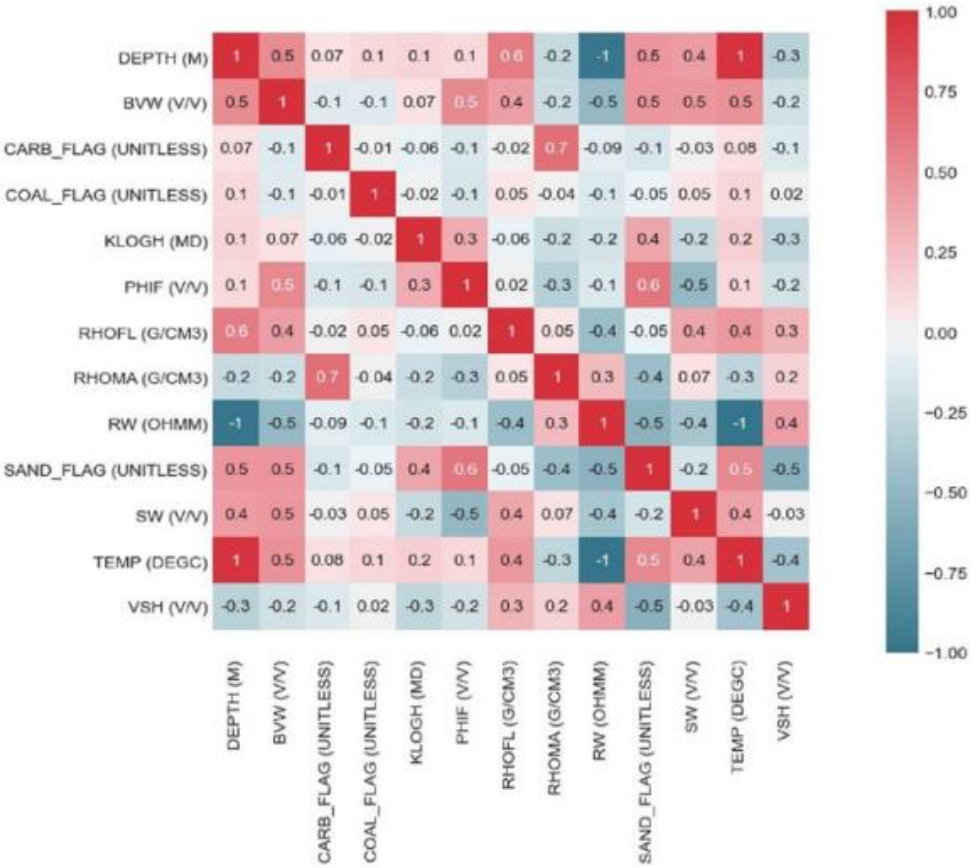

Figura 8: Mapa de Calor

Fuente: (Pandey, Rastogi, Kainkaryam, Bhattacharya, & Saputelli, 2020)

2 METODOLOGÍA

La preparación de una base de datos para modelos de aprendizaje de máquina es una etapa crítica en el desarrollo del modelo. Es fundamental garantizar la precisión de los conjuntos de datos, ya que las entradas de mala calidad o atípicos pueden dar lugar a un rendimiento deficiente del modelo. Además, el tamaño de la base de datos desempeña un papel sustancial en el proceso de aprendizaje. Un conjunto de datos suficientemente grande permite un entrenamiento más eficaz del modelo, mejorando su rendimiento global.

Para el desarrollo de la base de datos de prueba utilizada en la presente investigación se recopiló la información de los reportes diarios de perforación, reportes finales de fluidos de perforación y registros litológicos (*masterlog*[17]*)* correspondientes a 104 pozos del Bloque 60 Campo Sacha, la misma que fue validada mediante la comparación de datos entre los reportes diarios de perforación y reportes finales de fluidos.

En la Tabla 4 se enlistan las variables que fueron utilizadas como parámetros de entrada para el desarrollo de la base de prueba:

Tabla 4: Parámetros de Entrada

Parámetro	Simbología	Unidades
Profundidad Vertical Verdadera	tvd	ft
Profundidad Medida	md	ft
Inclinación	incl	grados
Azimuth	azim	grados
Dogleg	dogleg	-
Rata de Penetración	ROP_min, ROP_max	ft/hr
Peso sobre la broca	wob_min, wob_max	lbf
Revoluciones por Minuto	rpm_min, rpm_max	rpm
Caudal	q_min, q_max	bbl
Presiones	p_min, p_max	psi
Formaciones	Orteguaza, Tiyuyacu, Tena, Napo, Calizas de Napo, Areniscas de Napo, Hollín	-
Torque	toq_min, toq_max	lb-in

[17] *Masterlog:* Hace referencia a un registro detallado y completo de las características litológicas y geológicas encontradas durante la perforación de un pozo.

Parámetro	Simbología	Unidades
Densidad	den	lpg
Viscosidad	vis	sec/qt
Viscosidad Plástica	PV	Centipoise
Punto de Cedencia	YP	lbf/100ft2
Prueba de azul de metileno	MBT	lb/bbl
Ph	PH	ph
Geles	gel_10seg, gel_10min, gel_30min	lbf/100ft2
Filtrado	fil	ml/30min
Litología	Arenisca, Conglomerado, Limolita, Arcillolita, Caliza, Lutita, Caolinita y Anhidrita	-
Rotando / Deslizando	R, D	-

Elaborado por: Los autores

Por otro lado, la Tabla 5 muestra la categorización asignada a cada problema operacional, la cual será necesaria para entender los resultados mostrados en las predicciones realizadas por el modelo.

Tabla 5: Variable categórica "Problema" a variable numérica

Problema	Valor Numérico
Condición Normal	0
Embolamiento	1
Influjo	2
Pega Mecánica/Empaquetamiento	3
Pérdida Circulación	4
Taponamiento	5

2.1 Análisis y Procesamiento de Datos

La base de datos resultante, empleada como sustento en la construcción del modelo propuesto en el presente estudio, incorpora las siguientes consideraciones metodológicas:

- El proceso de depuración de datos, enfocado en la eliminación de inconsistencias y datos erróneos.

- La codificación precisa de variables categóricas con el propósito de adecuarlas al análisis posterior.
- La estandarización meticulosa de las variables, a fin de mitigar posibles distorsiones y asegurar un tratamiento homogéneo.

Los procedimientos descritos anteriormente han sido ampliamente detallados en el Trabajo de Titulación e Integración Curricular (TIC) realizado por (Romero, 2023), en el cual se empleó el lenguaje de programación Python como herramienta primordial para su ejecución.

Una vez obtenida la base de prueba final, se procede a su incorporación y ejecución en el entorno del código generado, permitiendo así la realización de análisis y predicciones pertinentes en concordancia con los objetivos de la investigación.

2.1.1 Modelo XG Boost 1 (n_estimators=38, learning_rate=0.1, max_depth=15 y random_state = 42)

A partir de los hiperparámetros considerados en el Modelo XG Boost 1 y la base de datos de prueba, la cual contiene 51 problemas distribuidos aleatoriamente (Condición Normal, Embolamiento, Influjo, Pega Mecánica/Empaquetamiento, Pérdida de Circulación y Taponamiento), se realiza la predicción en el código generado en *Python.* Para ilustrar de forma gráfica los resultados obtenidos por el modelo se realiza una matriz de confusión.

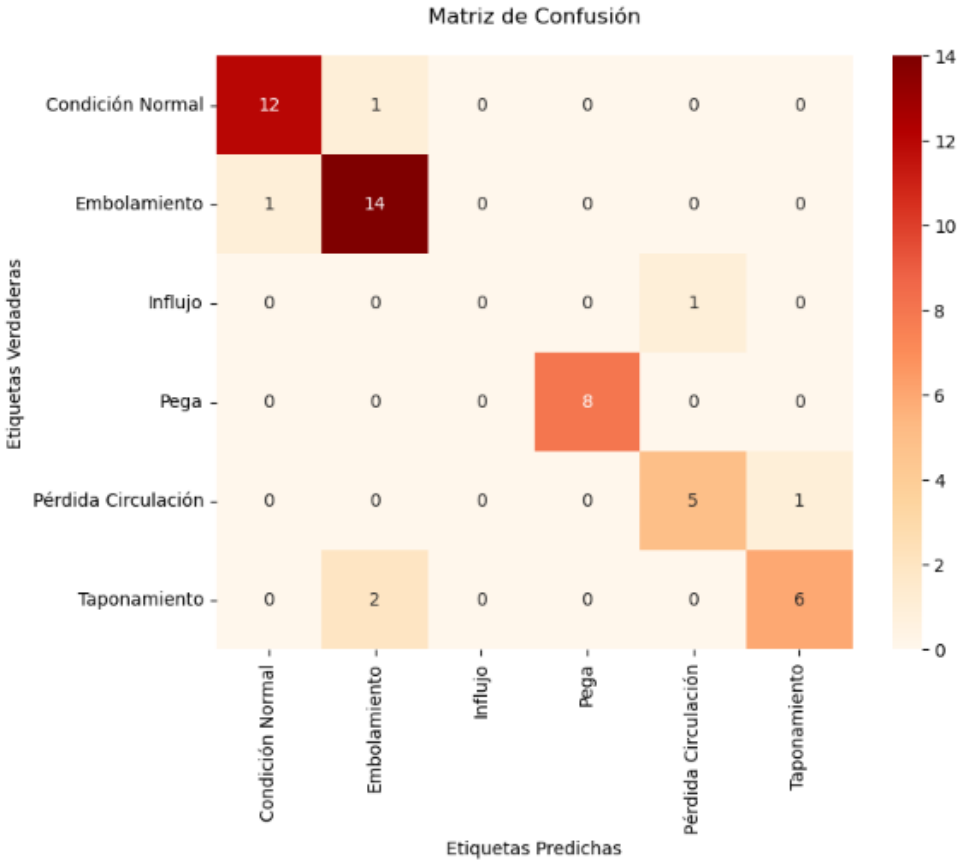

Figura 9: Matriz de Confusión de las Predicciones obtenidas (Modelo XG Boost 1)

Elaborado por: Andrés Llano

En la Figura 9 se muestran las predicciones efectuadas por el modelo, evidenciando una correspondencia precisa en 45 de los problemas anticipados, mientras que en 6 ocasiones se presenta una discordancia entre la predicción y la situación real.

Para conocer por qué el modelo realiza esta predicción, se emprende una exploración comparativa destinada a determinar los niveles de acierto de cada problema individualmente. Este análisis se realiza a través del empleo del lenguaje de programación Python, donde se desarrolla una tabla de comparación que contrasta los problemas identificados por el modelo con sus equivalentes reales. Dicha representación tabular se encuentra presentada en la Tabla 6: Comparación Real vs Predicho (Modelo XG Boost 1)Tabla 6, brindando una visualización resumida de los problemas reales y los predichos por el modelo.

Tabla 6: Comparación Real vs Predicho (Modelo XG Boost 1)

	Real	Predecir
166	2	4
126	5	1
70	5	1
25	0	1
52	4	5
114	1	0

Elaborado por: Andrés Llano

Una vez que los problemas han sido identificados, se procede a la construcción de una matriz que refleja los porcentajes de decisión asociados a cada uno de los problemas durante las predicciones realizadas por el modelo. Esta matriz ilustra de manera cuantitativa la certidumbre relativa con la que se aborda cada problema antes del resultado predictivo. Este análisis de ponderación se encuentra representado en la Tabla 7:

Tabla 7: Porcentaje de decisión (Modelo XG Boost 1)

	0	1	2	3	4	5
166	0,14	0,18	0,09	0,04	0,28	0,26
126	0,14	0,80	0,01	0,03	0,01	0,01
70	0,08	0,28	0,07	0,08	0,26	0,24
25	0,07	0,84	0,01	0,05	0,01	0,01
52	0,01	0,05	0,01	0,01	0,02	0,90
114	0,86	0,04	0,01	0,05	0,02	0,02

Elaborado por: Andrés Llano

En la Tabla 7 se presenta de manera gráfica el análisis de los porcentajes de predicción asignados a cada problema por parte del modelo. Conforme a esta representación, es posible deducir que el modelo, al enfrentarse a los datos de prueba, determinará su respuesta de acuerdo con el porcentaje más elevado obtenido en la predicción correspondiente. Por ejemplo, en el primer caso se observa que los porcentajes más altos son 28% (correspondiente a pérdida de circulación) y 26% (correspondiente a taponamiento) la predicción del algoritmo será pérdida de circulación, debido a que es el valor más alto entre todas las categorías. Por lo tanto, este proceso de toma de decisión

se fundamenta en la atribución de pesos relativos a cada posible solución, con la elección final guiada por la predicción que obtenga el mayor valor porcentual en el conjunto de posibilidades evaluadas.

2.1.2 Modelo XG Boost 2 (n_estimators=20, learning_rate=0.05, max_depth=3 y random_state = 42)

Para entender los resultados del modelo XG Boost 2 se procederá a la replicación del análisis efectuado en el primer modelo. Por consiguiente, los procedimientos metodológicos previamente empleados serán aplicados nuevamente al segundo modelo, asegurando una coherencia metodológica y facilitando la comparación entre los resultados obtenidos por ambos enfoques. En este contexto, se presentan los resultados en la Figura 10, Tabla 8, y Tabla 9.

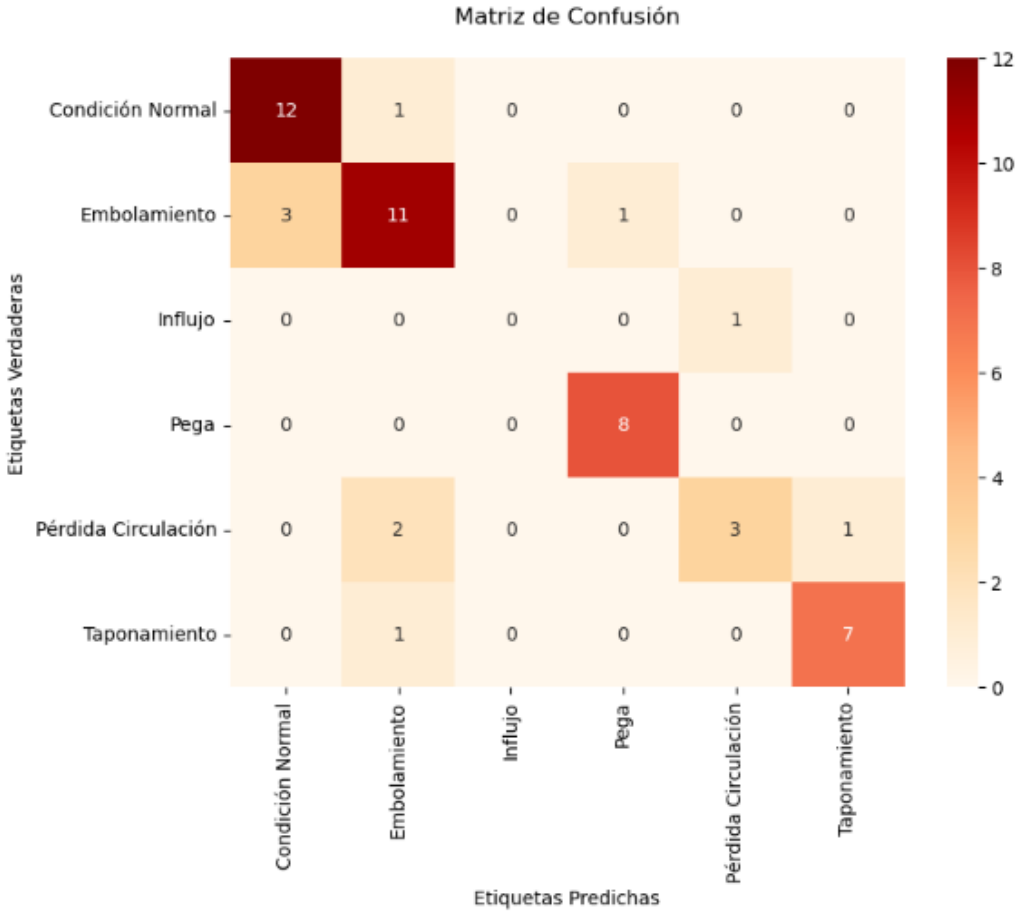

Figura 10: Matriz de Confusión de las Predicciones obtenidas (Modelo XG Boost 2)

Elaborado por: Andrés Llano

En la Figura 10 se muestran las predicciones efectuadas por el modelo, evidenciando una correspondencia precisa en 41 de los problemas anticipados, mientras que en 10

ocasiones se presenta una discordancia entre la predicción y la situación real. Los problemas en los cuales el modelo tuvo una predicción errónea se muestran de forma resumida en la Tabla 8, la cual ilustra el problema real y la predicción realizada, estos problemas se encuentra categorizados en función a la Tabla 5. En un enfoque complementario, la Tabla 9 se presenta un análisis en torno a los porcentajes de predicción asignados a cada problema por parte del modelo, proporcionando una perspectiva clara de las decisiones asumidas.

Tabla 8: Comparación Real vs Predicho (Modelo XG Boost 2)

	Real	Predecir
66	4	1
141	1	3
166	2	4
126	5	1
25	0	1
52	4	5
109	1	0
116	1	0
65	4	1
114	1	0

Elaborado por: Andrés Llano

Tabla 9: Porcentaje de decisión (Modelo XG Boost 2)

	0	1	2	3	4	5
66	0,15	0,27	0,13	0,18	0,13	0,14
141	0,20	0,21	0,10	0,28	0,10	0,10
166	0,16	0,16	0,14	0,09	0,25	0,20
126	0,23	0,41	0,06	0,17	0,06	0,07
25	0,13	0,46	0,09	0,14	0,09	0,09
52	0,09	0,15	0,09	0,09	0,09	0,49
109	0,34	0,21	0,11	0,14	0,11	0,11
116	0,32	0,21	0,10	0,17	0,10	0,10
65	0,14	0,23	0,13	0,22	0,14	0,14
114	0,45	0,11	0,09	0,17	0,09	0,09

Elaborado por: Andrés Llano

2.1.3 Evaluación y selección del Modelo XG Boost

Una vez completado el análisis y la exposición de los resultados de predicción derivados de los dos modelos propuestos (Modelo XG Boost 1 y Modelo XG Boost 2), se inicia la etapa de comparación y evaluación de los mismos. Con la finalidad de discernir y determinar cuál de los modelos se ajusta de manera óptima a los requerimientos de la presente investigación e identificar la alternativa más pertinente y efectiva para su implementación en los resultados.

Para entender de forma adecuada los resultados obtenidos de los dos modelos es necesario realizar la introducción de tres métricas fundamentales en la evaluación de modelos de clasificación en el aprendizaje de máquina. Cada una de estas métricas ofrece una perspectiva específica sobre el rendimiento del modelo en diferentes aspectos de la clasificación:

- *Recall:* Representa la proporción de ejemplos positivos reales que fueron correctamente identificados por el modelo.
- *Precision:* Representa la proporción de ejemplos clasificados identificados correctamente (positivos por el modelo que realmente son positivos).
- *f1-score:* Represente una métrica de evaluación del rendimiento de un modelo de clasificación que combina la precisión y el *recall* en un solo valor.

Una vez detalladas las tres métricas se presentan los resultados obtenidos de la predicción de la base de prueba en la Tabla 10 y Tabla 11 correspondientes al Modelo XG Boost 1 y Modelo XG Boost 2 respectivamente.

Tabla 10: Reporte de Clasificación (Modelo XG Boost 1)

	precision	***recall***	***f1-score***	***support***
0	0,92	0,92	0,92	13
1	0,82	0,93	0,87	15
2	0,00	0,00	0,00	1
3	1,00	1,00	1,00	8
4	0,83	0,83	0,83	6
5	0,86	0,75	0,80	8
				51
accuracy			0,88	

Elaborado por: Andrés Llano

Tabla 11: Reporte de Clasificación (Modelo XG Boost 2)

	precision	*recall*	*f1-score*	*support*
0	0,80	0,92	0,86	13
1	0,73	0,73	0,73	15
2	0,00	0,00	0,00	1
3	0,83	1,00	0,94	8
4	0,75	0,50	0,60	6
5	0,88	0,88	0,88	8
				51
accuracy			0,80	

Elaborado por: Andrés Llano

A partir de los resultados se manifiesta que el Modelo XG Boost 1 muestra un nivel de precisión del 88%, en contraste con el Modelo XG Boost 2 que alcanza una precisión del 80%. Esto se refleja en la efectividad de las predicciones a los problemas, ya que el primer modelo únicamente incurre en 6 fallos en la identificación, mientras que el segundo modelo registra 10 desaciertos. Asimismo, en la evaluación de parámetros como la *precisión, el recall y el f1-score*, se destaca la superioridad del primer modelo en la predicción de cada uno de los problemas, manifestando un mejor rendimiento en la clasificación al abordar los datos contenidos en la base de prueba.

A pesar de que el autor (Romero, 2023) recomienda el segundo modelo, debido a que en el entrenamiento obtuvo una precisión del 93% y este resultado quiere decir que el modelo no se encuentra sobre ajustado por lo que se pueden obtener mejores resultados en la predicción, después de comparar los dos modelos para esta investigación el Modelo XG Boost 1 con una precisión de entrenamiento del 100% es al que mejor se adapta el conjunto de datos de prueba, razón por la cual será utilizado para para definir los resultados de la investigación.

Después de definir el modelo a utilizar se presentan las fronteras de decisión de decisión en la Figura 11, donde se definen la separación de las diferentes clases o categorías del modelo de aprendizaje de máquina en la cual cada color representa un caso en particular:

- **Azul:** Condición Normal
- **Naranja:** Embolamiento
- **Verde:** Influjo

- **Rojo:** Pega Mecánica/Empaquetamiento
- **Morado:** Pérdida de Circulación
- **Café:** Taponamiento

En este contexto, si un dato de la base de prueba se posiciona en una zona en particular la predicción se dará respecto a la frontera de decisión (Figura 11) y tomando en cuenta el porcentaje de decisión (Tabla 7).

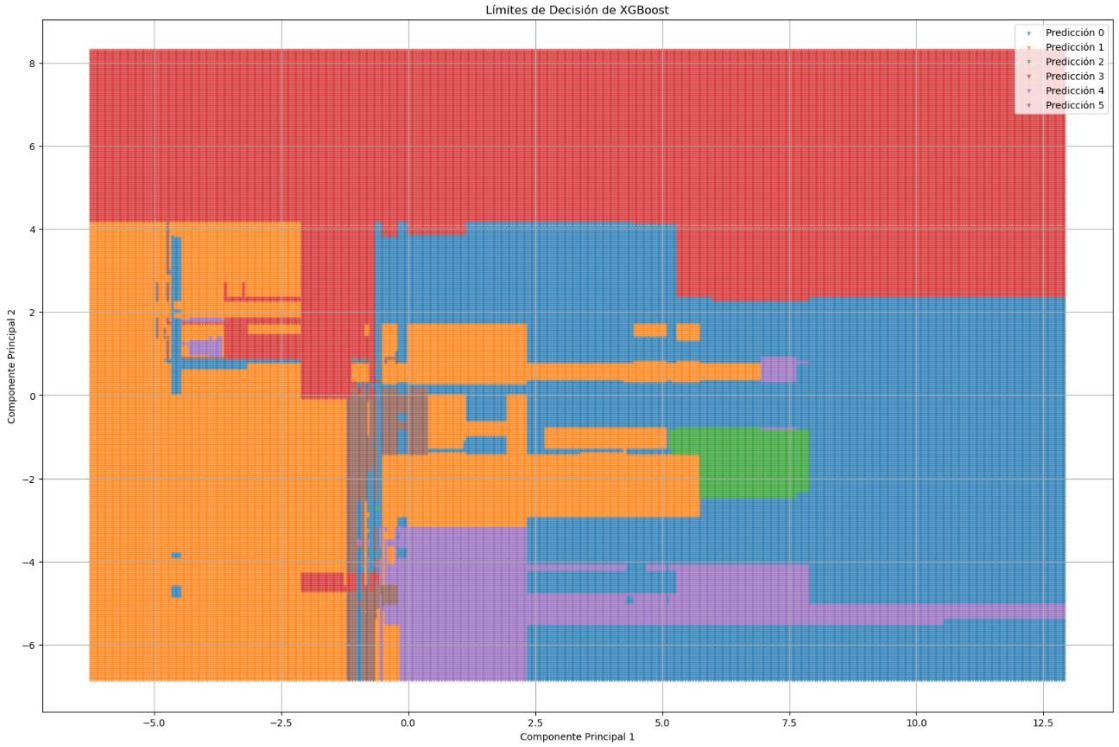

Figura 11: Fronteras de Decisión

Elaborado por: Andrés Llano

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

Con el fin de realizar un análisis profundo de las predicciones equivocadas obtenidas mediante la implementación del Modelo XG Boost 1, se examinarán caso a caso las predicciones incorrectas para comprender las razones detrás de la decisión del algoritmo.

Para abordar este análisis, se considerarán los porcentajes de predicción asignados a cada problema por el modelo (Tabla 7), lo cual permitirá identificar los casos en los cuales las predicciones erróneas se basaron en una menor confianza de parte del modelo. Además,

se llevará a cabo un análisis utilizando el Análisis de Componentes Principales (PCA) los cuales están distribuidos en PCA1 y PCA2, los que permitirán ilustrar la distribución del problema real con el problema predicho.

La combinación de estos enfoques permitirá una apreciación general de las predicciones equivocadas, brindando una perspectiva tanto cuantitativa como cualitativa. A través de este análisis, se pretende conocer las posibles limitaciones del modelo XG Boost y su comportamiento en situaciones específicas, contribuyendo así a una comprensión más completa de su capacidad predictiva.

3.1.1 Problema 1 (Real: Influjo, Predicción: Pérdida de Circulación)

En la Figura 12, se puede observar la dispersión de los datos de Influjo y Pérdida de Circulación, el punto de color naranja es el dato que se va a predecir. Este caso en particular es fácil de analizar, debido a que, el error de predicción se debe a la falta de datos para los problemas relacionados con el influjo, puesto que solo obtuvieron 3 problemas en particular, de los cuales 2 se utilizaron para la base de entrenamiento del modelo y 1 para la base prueba. Esto se ve reflejado en la Tabla 7: Porcentaje de decisión (Modelo XG Boost 1) en la primera fila, obteniendo un porcentaje para predecir el influjo de tan solo 9%, mientras que la pérdida de circulación alcanza el valor de 28%.

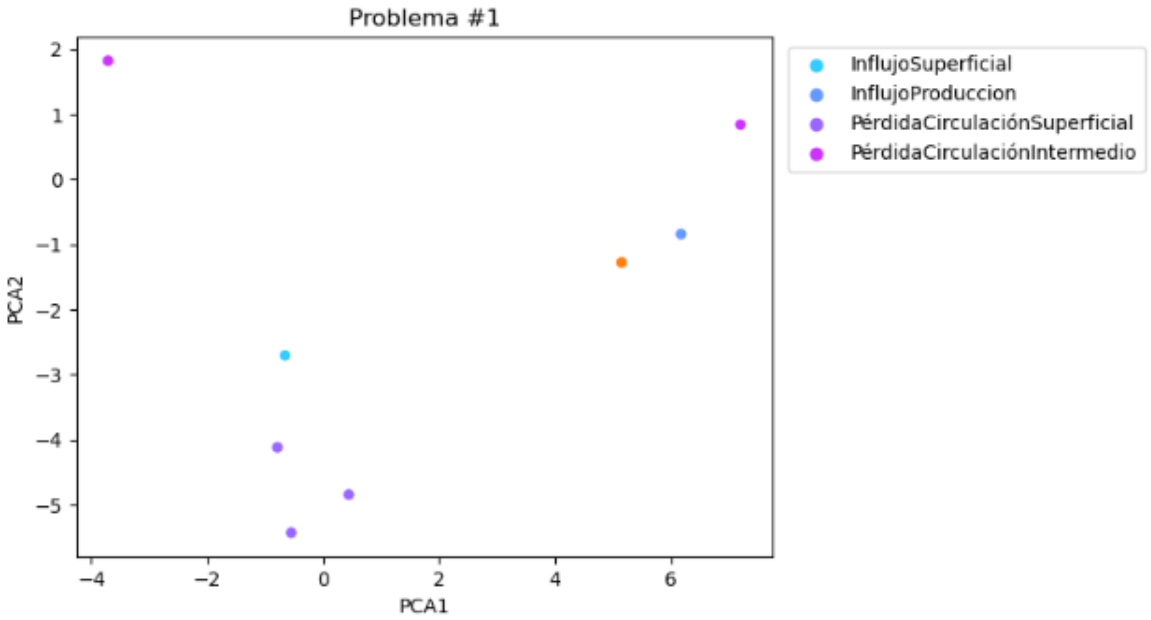

Figura 12: Análisis de Componentes Principales (Problema 1)

Elaborado por: Andrés Llano

3.1.2 Problema 2 y 3 (Real: Taponamiento, Predicción: Embolamiento)

En la Figura 13 se presentan la distribución de los datos de taponamiento y embolamiento, se puede observar solo existe un dato cercano de taponamiento perteneciente a la sección intermedia alrededor del dato que se va a predecir (punto naranja), además, este punto cae en el área de los datos de embolamiento de la sección intermedia, por tal motivo la probabilidad que este dato se ha predicho como embolamiento es alta. En la Tabla 7 en la segunda fila se puede notar que el porcentaje de predicción para el taponamiento es de 1%, mientras que se obtiene un 80% para categorizar como embolamiento.

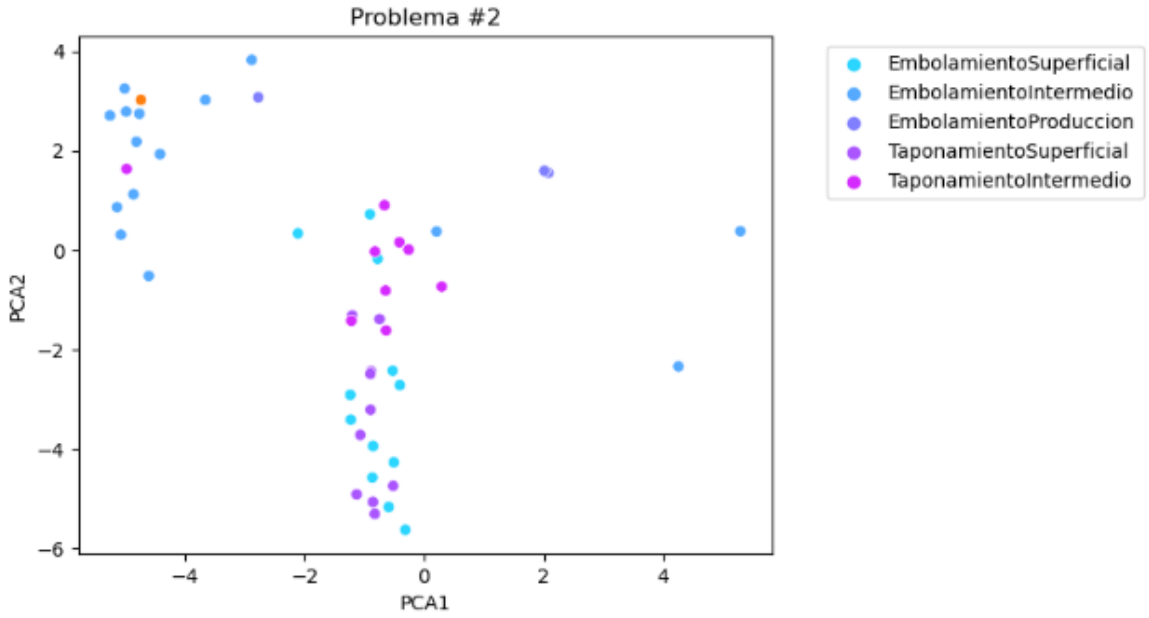

Figura 13: Análisis de Componentes Principales (Problema 2)

Elaborado por: Andrés Llano

Para el caso de la Figura 14 se puede evidenciar que el dato de prueba (punto naranja) se encuentra en un área donde existen problemas de embolamiento y taponamiento superficial, sin embargo, es claro notar que el dato de prueba se acerca más a los problemas de embolamiento, por este motivo el algoritmo lo predice como tal. Además, en la Tabla 7 en la tercera fila se puede observar los porcentajes de predicción que para embolamiento es de 28% y el taponamiento alcanza un porcentaje de 24%, es decir la diferencia es tan solo del 2%.

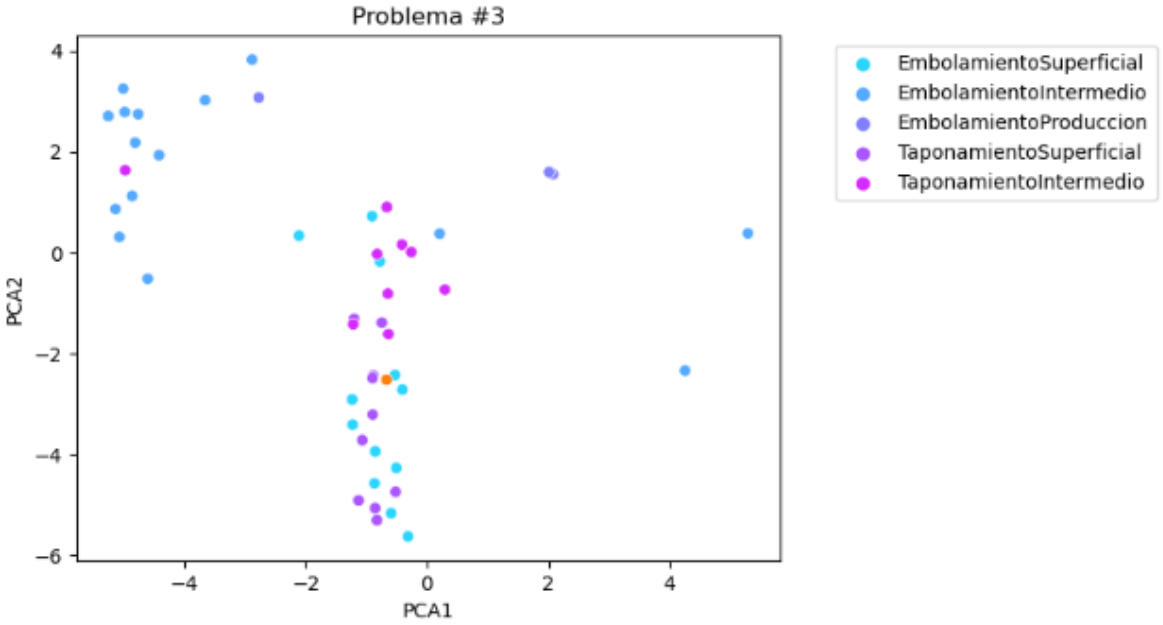

Figura 14: Análisis de Componentes Principales (Problema 3)

Elaborado por: Andrés Llano

3.1.3 Problema 4 (Real: Condición Normal, Predicción: Embolamiento)

La Figura 15 muestra el conjunto de datos de problemas de condición normal y embolamiento, se puede observar que existe tan solo 2 datos del problema de condición normal, pero hay una alta concentración de datos de embolamiento de la sección intermedia alrededor del dato a predecir (punto naranja), por esta razón el algoritmo le asigna este resultado al dato de predicción. También, en la Tabla 7 en la cuarta fila se puede observar que el porcentaje de predicción para el embolamiento es del 84%, mientras que para la condición normal es del 7%.

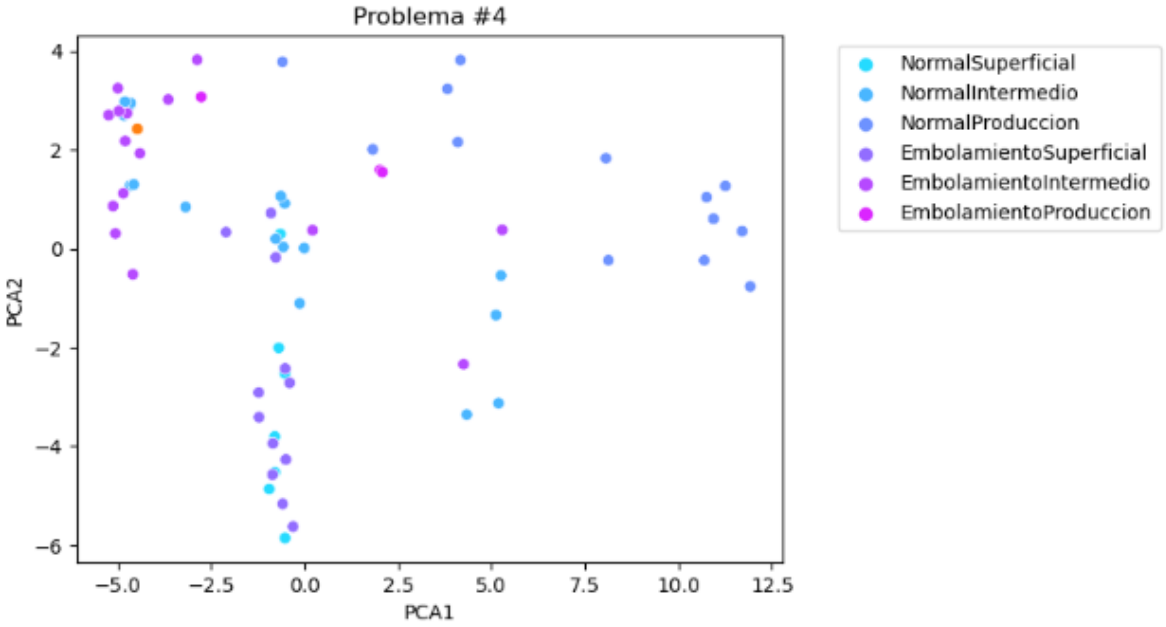

Figura 15: Análisis de Componentes Principales (Problema 4)

Elaborado por: Andrés Llano

3.1.4 Problema 5 (Real: Pérdida de Circulación, Predicción: Taponamiento)

En la Figura 16 se observa que el dato a predecir (punto naranja) se encuentra alrededor de los datos de problemas de taponamiento superficial, además se evidencia una mayor densidad problemas de taponamiento que de pérdida de circulación en el presente gráfico, por esta razón el algoritmo clasifica este dato de predicción como taponamiento. En la Tabla 7 en la quinta fila se observa que el porcentaje de predicción para el tamponamiento es del 90% y tan solo con un 2% para la pérdida de circulación.

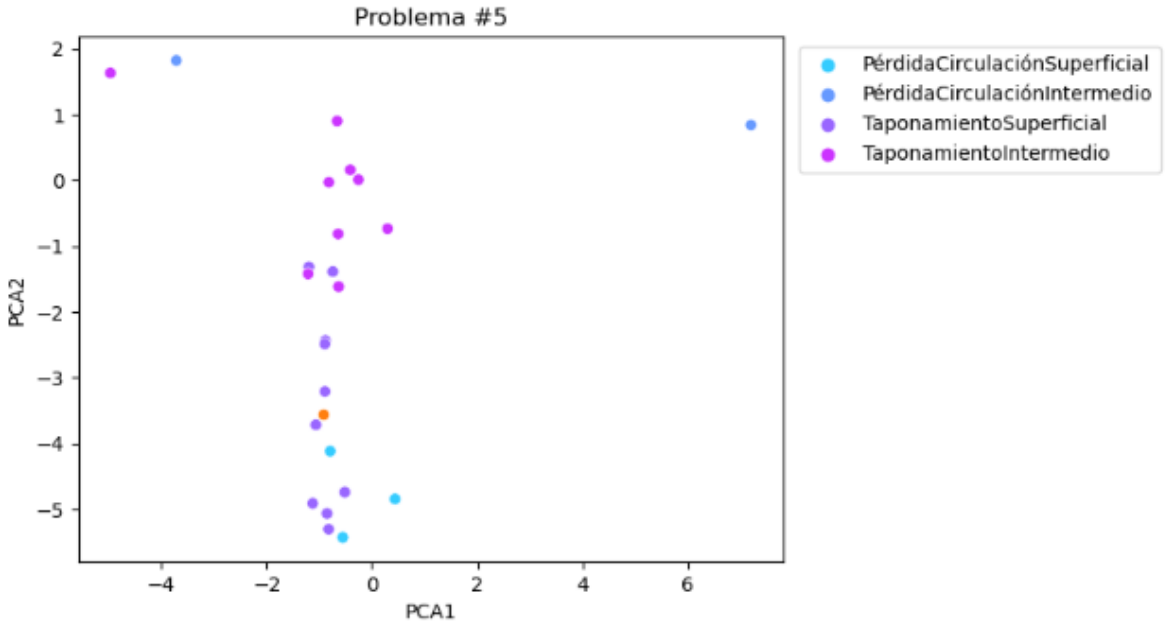

Figura 16: Análisis de Componentes Principales (Problema 5)

Elaborado por: Andrés Llano

3.1.5 Problema 6 (Real: Embolamiento, Predicción: Condición Normal)

Finalmente, en la Figura 17 se presenta el conjunto de datos de problemas de condición de normal y embolamiento, se observa que el dato a predecir (punto naranja) está alrededor de los datos de condición normal de la sección superficial (2 puntos aislados). Además, en la Tabla 7 se muestra en la sexta fila el porcentaje de predicción para condición normal alcanza el 86%, mientras que el embolamiento el 4%. Por esta razón, el algoritmo clasifica a este dato de prueba como embolamiento.

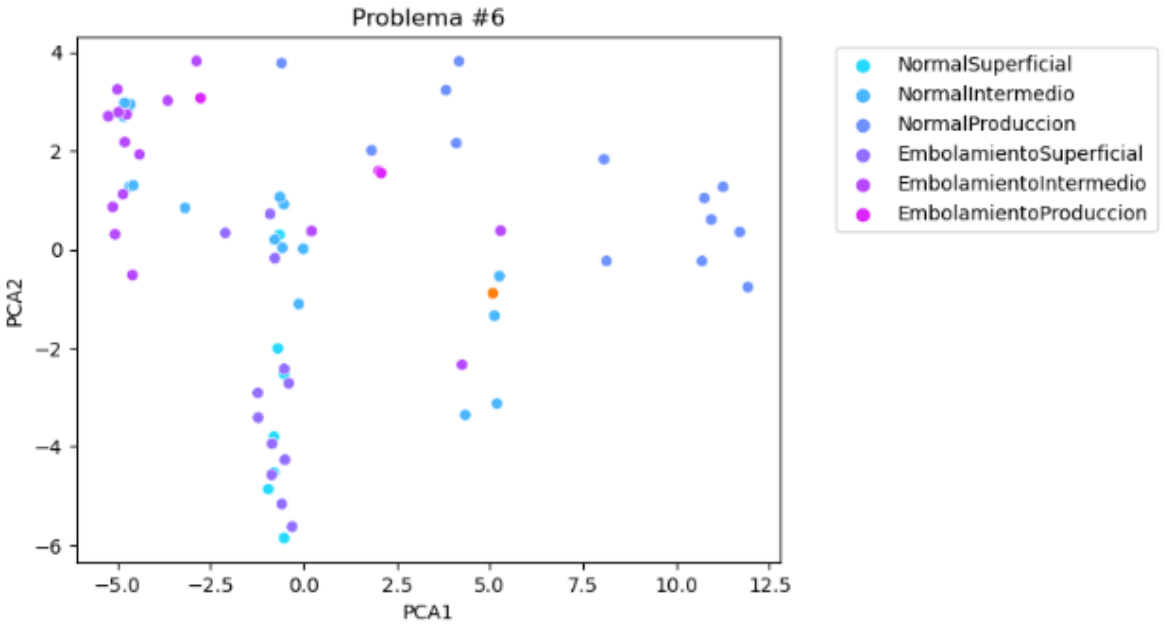

Figura 17: Análisis de Componentes Principales (Problema 6)

Elaborado por: Andrés Llano

3.2 Conclusiones

- Para la generación de la base de prueba se seleccionaron pozos provenientes de diversas áreas del campo, buscando garantizar una representación variada de condiciones geológicas y geográficas. Esta elección fue orientada para asegurar una distribución adecuada de los datos en el ámbito del yacimiento. Cabe recalcar que la información fue sometida a un proceso de validación, con el objetivo de asegurar la integridad y calidad de los datos consignados en la base de datos desarrollada.
- En el análisis de los resultados obtenidos a través de la implementación de los modelos XG Boost 1 y 2, surge una inclinación sólidamente fundamentada hacia la preferencia del Modelo XG Boost 1. Esta elección está anclada en la búsqueda de una solución que garantice tanto un alto grado de precisión en las predicciones como una relevante adaptabilidad a la realidad operativa del contexto estudiado. Siendo el objetivo central de este estudio la identificación temprana de problemas en las operaciones de perforación, la elección del Modelo XG Boost 1 con una precisión del 88% se alinea de manera óptima con los requerimientos de este propósito.

- La ausencia de sobreajuste en la fase de entrenamiento de un modelo no implica necesariamente la obtención de resultados superiores con la base de prueba, ni viceversa en un modelo que exhiba sobreajuste durante la etapa de entrenamiento. Este contraste se evidencia en el caso del Modelo XG Boost 2, que, a pesar de carecer de sobreajuste en la fase de entrenamiento, manifiesta una menor precisión en su rendimiento con la base de prueba. Por el contrario, el Modelo XG Boost 1, al mantener una adaptabilidad más efectiva en el proceso predictivo tiene como resultado una predicción precisa de los casos presentados en la base de prueba.

- A pesar de la naturaleza iterativa y acumulativa del algoritmo XG Boosting, en el cual se agregan sucesivos árboles capaces de prever los residuos de errores de sus predecesores, se evidencia una limitación intrínseca en el contexto específico del problema de taponamiento. Dicha limitación se deriva de la cantidad escasa de datos disponibles para este problema en la base total, donde únicamente se registran 3 instancias (2 en la fase de entrenamiento y 1 en la de prueba). Este insuficiente volumen de datos en particular dificulta de manera significativa la capacidad predictiva del modelo en relación con el problema operacional de taponamiento.

3.3 Recomendaciones

- Se recomienda la validación y depuración de la información recolectada en el proceso de desarrollo de la base de datos. La existencia de valores atípicos, ausencia de datos o perturbaciones en la información puede traducirse en resultados imprecisos y decisiones desorientadas. Siempre es necesario la adopción de prácticas rigurosas de depuración y validación de datos como una estrategia esencial para garantizar la robustez y utilidad del modelo resultante.

- Recopilar toda la información de los reportes diarios de perforación, reportes finales de fluidos de perforación y registros litológicos (*masterlogs*) de cada uno de los pozos para realizar una validación de los datos con el fin de no tener espacios en blanco y no sea necesario llenar con el promedio de la sección o del caso en particular.

- Para la base de datos entrenamiento obtener una información equitativa en cada uno de los casos (problemas), debido a que la uniformidad en la cantidad de datos contribuye a una robustez estadística, permitiendo una evaluación más precisa y eficaz del modelo.

REFERENCIAS BIBLIOGRÁFICAS

Abdalla, R., Samara, H., Perozo, N., Paz Carvajal, C., & Jaeger, P. (2022). *Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs).* Clausthal: ACS Omega.

Alkinani, H. H., Al-Hameedi, A. T., & Dunn-Norman, S. (2020). Data–driven decision–making for lost circulation treatments: A machine learning approach. *Energy and AI*.

Alsheikh, M. (13 de Julio de 2022). *Operational Digitalization in Advancement of Oil and Gas: A Young Professional's Perspective*. Obtenido de Journal of Petroleum Technology: https://jpt.spe.org/operational-digitalization-in-advancement-of-oil-and-gas-a-young-professional-perspective

Azar, J., & Robello Samuel, R. (2007). *Drilling Engineering.* Tulsa: PennWell Corporation.

Baberli, E. (1998). *El Pozo Ilustrado.* Caracas: Fondo Editorial del Centro Internacional de Educación y Desarrollo.

Baby, P., Rivadeneira, M., & Barragán, R. (2014). *La Cuenca Oriente: Geología y Petróleo.* Travaux de l'Institut Francais d'Études Andines.

Bayan, M., & Zulkarnain. (2020). Stuck pipe prediction in geothermal well drilling at Darajat using statistical and machine learning application. *Asia Pacific Conference on Research in Industrial and Systems Engineering*.

Bikmukhametov, T., & J¨aschke, J. (2022). *Oil Production Monitoring using Gradient Boosting Machine Learning Algorithm.* Trondheim: Norwegian University of Science and Technology.

CCOO de Industria. (Septiembre de 2017). Impacto industrial y laboral. *La Digitalización y la Industria 4.0*. Madrid: CCOO de Industria.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* San Francisco: Association for Computing Machinery. doi:https://doi.org/10.1145/2939672.2939785

Chen, W., Yu, Y., Shen, Y., & Zhengxin Zhang, V. V. (2019). Automatic Drilling Dynamics Interpretation Using Deep Learning. *Society of Petroleum Engineers (SPE)*.

Cuzco, D., & Ortiz, O. (2013). *ESTUDIO DE LA TECNOLOGÍA DE PERFORACIÓN, DISEÑO Y PLANIFICACIÓN DE UN POZO MULTILATERAL NIVEL 5 DE DOS RAMALES EN UN CAMPO PETROLERO DEL ORIENTE ECUATORIANO.* Quito: Escuela Politécnica Nacional.

Elmousalami, H. H., & Elaskary, M. (2020). Drilling stuck pipe classification and mitigation in the Gulf of Suez oil fields using artificial intelligence. *Journal of Petroleum Exploration and Production Technology*.

Encinas, M., Tunkiel, A., & Sui, D. (2022). Downhole data correction for data-driven rate of penetration prediction modeling. *Journal of Petroleum Science and Engineering*.

Fawcett, T. (19 de Diciembre de 2005). *An Introduction to ROC Analysis.* Obtenido de Elsevier: https://people.inf.elte.hu/kiss/11dwhdm/roc.pdf

Hernandez, L. (Febrero de 2022). *Univerisidad Politécnica de Madrid .* Obtenido de Análisis predictivo de funcionamiento de Sistema Híbrido Off Grid mediante Machine Learning: https://oa.upm.es/72650/

Hossain, M. E., & Islam, M. R. (2018). *Drilling Engineering Problems and Solutions.* Beverly: Scrivener Publishing.

Hou, X., Yang, J., Yin, Q., Chen, L., Cao, B., Xu, J., . . . Zhao, X. (2019). Automatic Gas Influxes Detection in Offshore Drilling Based on Machine Learning Technology. *SPE Gas & Oil Technology Showcase and Conference*.

Hou, X., Yang, J., Yin, Q., Liu, H., Chen, H., Zheng, J., . . . Liu, X. (2020). Lost Circulation Prediction in South China Sea Using Machine Learning and Big Data Technology. *Offshore Technology Conference*.

Jaramillo, E. A. (2018). *ANÁLISIS DEL DISEÑO DE ENSAMBLAJE DE FONDO (BHA) Y LA TUBERÍA DE PERFORACIÓN UTILIZADO EN LA CONSTRUCCIÓN DEL POZO XDIRECCIONAL TIPO "J" EN EL CAMPO AUCA EN EL ORIENTE ECUATORIANO.* Quito: Universidad Tecnológica Equinoccial.

Johnson, R. (2012). *Probabilidad y estadística para ingenieros.* Naucalpan de Juárez: PEARSON EDUCACIÓN.

Juárez, M. G. (2022). *Tipos de Asimetría*. Obtenido de Probabilidad y Estadística.Net: https://www.probabilidadyestadistica.net/tipos-de-asimetria/

Keita, Z. (Septiembre de 2022). *¿Qué es la clasificación en el aprendizaje automático?* Obtenido de Clasificación en el aprendizaje automático: Introducción: https://www.datacamp.com/blog/classification-machine-learning

Kuesters, A., Mason, C., Gomes, P., Cockburn, C., & Lodhi, H. (2020). Drillstring Failure Prevention–A Data Driven Approach to Early Washout Detection. *International Drilling Conference and Exhibition*.

Lake, L., & Mitchell, R. (2006). *Petroleum Engineering Handbook.* Richardson: Society of Petroleum Engineers.

Li, Y., & Samuel, R. (2019). Prediction of Penetration Rate Ahead of the Bit Through Real-Time Updated Machine Learning Models. *Society of Petroleum Engineers (SPE)*.

Mahendra, S. (26 de Junio de 2023). *Artificial Intelligence.* Obtenido de Introduction to XGBoost and its Uses in Machine Learning: https://www.aiplusinfo.com/blog/introduction-to-xgboost-and-its-uses-in-machine-learning/

Markovic, S., Bryan, J., Rezaee, R., Turakhanov, A., Cheremisin, A., Kantzas, A., & Koroteev, D. (17 de agosto de 2022). *Application of XGBoost model for in-situ water saturation determination in Canadian oil-sands by LF-NMR and density data.* Scientific Reports.

Martín Guareño, J. J. (2016). *Support vector regression : propiedades y aplicaciones.* Sevilla: Universidad de Sevilla. Obtenido de http://hdl.handle.net/11441/43808

Mendenhall, W., Beaver, R., & Beaver, B. (2010). *Introducción a la Probabilidad y Estadística.* Ciudad de México: Cengage Learning.

Noshi, C., & Schubert, J. (2019). Application of Data Science and Machine Learning Algorithms for ROP Prediction. *Offshore Technology Conference*. doi:https://doi.org/10.4043/29288-MS

Okoli, P., Cruz Vega, J., & Shor, R. (2019). Estimating Downhole Vibration via Machine Learning Techniques Using Only Surface Drilling Parameters. *SPE Western Regional Meeting*.

Olukoga, T., & Feng, Y. (2021). *Practical Machine-Learning Applications in Well-Drilling Operations.* Louisiana : SPE Drilling and Completion.

Pandey, Y. N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., & Saputelli, L. (2020). *Machine Learning in the Oil and Gas Industry.* New York: Apress.

Rabia, H. (2002). *Well Engineering & Construction.* London: Entrac Consulting Limited.

Richardson, M. (Mayo de 2009). *Principal Component Analysis.* Obtenido de https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf

Rodriguez, M., & Mora, R. (2001). Análisis de regresión múltiple. En *Estadística informática : casos y ejemplos con el SPSS.* Alicante: Publicaciones de la Universidad de Alicante. Obtenido de http://hdl.handle.net/10045/8143

Rodriguez, V. (17 de Octubre de 2018). *Decision trees / Árboles de decisión para clasificar en python.* Obtenido de https://vincentblog.xyz/posts/decision-trees-arboles-de-decision-para-clasificar-en-python

Romero, J. M. (2023). *Predicción de problemas en el procesos de perforación de pozos petroleros aplicando aprendizaje de máquina supervisado. .* Quito: Escuela Politécnica Nacional.

Sabah, M., Talebkeikhah, M., Agin, F., Talebkeikhah, F., & Hasheminasab, E. (2019). Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field. Journal of Petroleum Science and Engineering. *Journal of Petroleum Science and Engineering*.

Sethi, A. (27 de Marzo de 2020). *Support Vector Regression Tutorial for Machine Learning.* Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

Shadravan, A., Tarrahi, M., & Aman, M. (2017). Intelligent Tool To Design Drilling, Spacer, Cement Slurry, and Fracturing Fluids by Use of Machine-Learning Algorithms. *SPE Drilling & Completation*.

Shanmugam, R., & Chattamvelli, R. (2015). *Statistics for Scientists and Engineers.* Hoboken: Wiley.

Shaowei, P., Zechen, Z., Zhi, G., & Haining, L. (2022). *An optimized XGBoost method for predicting reservoir porosity using petrophysical logs.* Journal of Petroleum

Science and Engineering. Beijing: Elsevier. doi:https://doi.org/10.1016/j.petrol.2021.109520.

Sharma, M. (15 de August de 2019). *Analytics Vidhya.* Obtenido de Guide to Principal Component Analysis: https://medium.com/analytics-vidhya/guide-to-principal-component-analysis-ab04a8a9c305

Shi, X., Zhou, Y., Zhao, Q., Jiang, H., Zhao, L., Liu, Y., & Yang, G. (2019). A New Method to Detect Influx and Loss During Drilling Based on Machine Learning. *International Petroleum Technology Conference*.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of Machine Learning and Artificial intelligence in oil and gas industry. *Petroleum Research 6*, 383.

Spiegel, M., & Stephens, L. (2009). *Estadística.* Ciudad de México: Mc Graw-Hill.

Ting, K. (2011). Confusion Matrix. En C. Samut, *Encyclopedia of Machine Learning* (pág. 209). Boston: Sringer. doi: https://doi.org/10.1007/978-0-387-30164-8_157

Trenchlesspedia. (Diciembre de 2022). *What Does Conductor Casing Mean?* Obtenido de Trenchlesspedia: https://www.trenchlesspedia.com/definition/2252/conductor-casing#:~:text=A%20conductor%20casing%20may%20not,likely%20to%20be%20a%20problem.

Ubillus, J., & Pacheco, W. (2021). *Desarrollo de una herramienta computacional de evaluación de problemas operacionales en la perforación de pozos en el Campo Sacha.* Quito: Escuela Politécnica Nacional.

Walpole , R., Myers, R., Myers, S., & Ye, K. (2007). *Probability & Statistics for Engineers & Scientists 8th Edition.* New Jersey: Pearson College Div.

Wilkinson, L., & Friendly, M. (2012). The History of the Cluster Heat Map. *The American Statistician*.

Xie, Z., F. Q., Zhang, J., Shao, X., Zhang, X., & Wang, Z. (2021). *Prediction of Conformance Control Performance for Cyclic-Steam-Stimulated Horizontal Well Using the XGBoost.* Energies. doi:http://dx.doi.org/10.3390/en14238161

Yang, J., Sun, T., Zhao, Y., Borujeni, A. T., Shi, H., & Yang, H. (2019). Advanced Real-Time Gas Kick Detection Using Machine Learning Technology. *The 29th International Ocean and Polar Engineering Conference*.

Zha, Y., & Pham, S. (2018). Monitoring Downhole Drilling Vibrations Using Surface Data Through Deep Learning. *2018 SEG International Exposition and Annual Meeting*.