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

GENERACIÓN DE UNA BASE DE DATOS A TRAVÉS DEL ANÁLISIS DE CARPETAS DE POZOS Y CODIFICACIÓN DEL MODELO DE APRENDIZAJE DE MÁQUINA

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

JORGE MATEO ROMERO MUÑOZ

cefk.mromero@gmail.com

DIRECTOR: MARIO LAURO ROBLES REYES

mario.robles@epn.edu.ec

DMQ, octubre 2023

CERTIFICACIONES

Yo, JORGE MATEO ROMERO MUÑOZ 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.

_____

JORGE MATEO ROMERO MUÑOZ

Certifico que el presente trabajo de integración curricular fue desarrollado por JORGE MATEO ROMERO MUÑOZ, 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 (los) producto(s) resultante(s) 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.

JORGE MATEO ROMERO MUÑOZ

MARIO LAURO ROBLES REYES

JOEL ANDRÉS LLANO ESPÍN

DEDICATORIA

Dedico este trabajo a todos aquellos que me han brindado su apoyo incondicional a lo largo de este camino. A mi familia, por su amor y paciencia infinitos; a mis amigos, por sus ánimos y risas en los momentos difíciles; y a mis profesores y mentores, por su guía y conocimiento que han iluminado mi camino hacia el logro de este objetivo. Esta tesis es el fruto del esfuerzo conjunto y el resultado de incontables horas de trabajo, y con humildad la dedico a todos aquellos que han contribuido de alguna manera en mi formación académica y personal. Que este trabajo sea un pequeño tributo a su apoyo y un paso hacia un futuro lleno de nuevos desafíos y logros.

Mateo Romero

AGRADECIMIENTO

A Dios, por guiar mis pasos y brindarme las oportunidades necesarias en este camino académico y personal. Su gracia me ha acompañado en cada paso del camino.

A mis queridos padres, Jorge y Mery, por su amor incondicional, apoyo constante y sacrificio.

A EP Petroecuador, por brindarnos el acceso a la información necesaria para el desarrollo de esta investigación.

Al profesor y tutor de tesis MSc. Diego Cuzco, por ser nuestro guía en este proceso. Sus consejos, conocimientos y dedicación han sido invaluables para la culminación exitosa de este trabajo. Agradezco su tiempo y paciencia al orientarnos hacia la excelencia académica.

A Andrés, mi compañero de tesis, por su esfuerzo y dedicación prestada para el desarrollo de este trabajo, y sobre todo por su amistad.

A mis profesores y amigos, con quienes construí un camino lleno de bellos recuerdos en esta etapa universitaria.

Mateo Romero

ÍNDICE DE CONTENIDO

RESUMEN

La implementación de aprendizaje de máquina en la perforación ha permitido ahorrar tiempo y optimizar sus procedimientos, dado que los algoritmos son capaces de extraer información rápidamente correlacionando las variables para visualizar de manera clara conjuntos de datos grandes y complicados. La optimización de parámetros operacionales, predicción de problemas, optimización de trayectorias y control automatizado de tareas son algunos campos en los que el aprendizaje de máquina ha sido aplicado eficientemente. Esta investigación implementó el algoritmo de aprendizaje de máquina supervisado XGBoost para predecir problemas operacionales ocurridos durante la perforación de pozos petroleros en el oriente ecuatoriano. La información recopilada incluye: reportes diarios de perforación, reportes finales de fluidos de perforación y registros litológicos correspondientes a 104 pozos del Bloque 60 Campo Sacha.

El análisis de componentes principales facilitó visualizar la agrupación de los problemas en función de la sección de perforación, representando el conjunto de datos en un gráfico bidimensional. La correcta aplicación de XGBoost permitió generar un modelo inicial con hiperparámetros específicos: n_estimators=38, learning_rate=0.1, max_depth=15 y random_state = 42, el mismo que presentó 100% de predicción. Con el fin de reducir el posible sobreajuste del modelo, los hiperparámetros fueron modificados a n_estimators=20, learning_rate=0.05, max_depth=3 y random_state = 42, obteniendo ahora 93% de predicción en la base de entrenamiento. La finalidad es evaluar cuál de estos modelos ofrece un mejor rendimiento en un conjunto de datos de prueba. Este enfoque permite operaciones seguras y facilita así los procesos de toma de decisiones.

PALABRAS CLAVE: aprendizaje de máquina, XGBoost, problemas operacionales, análisis de componentes principales, hiperparámetros.

ABSTRACT

The implementation of machine learning in drilling has enabled time savings and procedure optimization, as algorithms can swiftly extract information and vividly visualize vast and intricate datasets. Optimization of operational parameters, problem prediction, trajectory optimization, and automated task control are some of the fields in which machine learning has been efficiently applied. This research employed the supervised machine learning algorithm XGBoost to predict operational problems during oil well drilling in the Ecuadorian East. The gathered information comprises daily drilling reports, final drilling fluid reports, and corresponding lithological records for 104 wells in Block 60 Sacha Field.

Principal Component Analysis (PCA) facilitated the visualization of problem grouping based on the drilling section, representing the dataset in a two-dimensional graph. The proper implementation of XGBoost allowed the generation of an initial model with specific hyperparameters: n_estimators=38, learning_rate=0.1, max_depth=15, and random_state=42, which demonstrated 100% prediction accuracy. In order to mitigate potential overfitting of the model, the hyperparameters were adjusted to n_estimators=20, learning_rate=0.05, max_depth=3, and random_state=42, resulting in a prediction accuracy of 93% on the training set. The objective is to assess which of these models performs better on a test dataset. This approach enables secure operations and thereby facilitates decision-making processes.

KEYWORDS: machine learning, XGBoost, operational drilling problems, principal component analysis, hyperparameters

1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

El proceso de perforación implica el desarrollo de varias actividades incluyendo: el armado de paradas de tubería de perforación, armado del ensamblaje de fondo, construcción y revestimiento del pozo, registros de evaluación de formaciones y cementación. Todas estas actividades incluyen un riesgo de presencia de problemas operacionales causados por fallos de los equipos o el comportamiento aleatorio de las formaciones y los parámetros utilizados en la ejecución del programa.

Los problemas (ej. Pega mecánica y diferencial, pérdida de circulación, influjo, entre otros) causados por las formaciones, parámetros operacionales incorrectos y pobres diseños de lodos de perforación, han sido ampliamente estudiados, sin embargo, aún no se ha desarrollado una herramienta tecnológica que permita predecir su aparición durante las operaciones con un grado alto de precisión. Por otro lado, el desarrollo de este tipo de tecnología permitirá optimizar los tiempos de perforación a través de la reducción de tiempos no productivos (NPT) generados por dichos problemas y por lo tanto, disminuir los costos finales de esta actividad.

En este contexto, se plantea generar una base de datos a través del uso de un volumen de información de mínimo 100 pozos, del cual se recopilará la información correspondiente a geometría del hoyo, parámetros operacionales, propiedades de lodo e información litológica que permitan catalogar si bajo esas condiciones, se presentó o no un problema operacional en la perforación. Posteriormente, mediante la aplicación del algoritmo Extreme Gradient Boosting (XGBoost) como algoritmo de aprendizaje de máquina supervisado, se pretende desarrollar un modelo validado que prediga los problemas operacionales con una baja incertidumbre.

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. Seleccionar los 50 pozos para la generación de la base de datos.
2. Validar la información de los pozos seleccionados.
3. Generar un base de datos de los pozos seleccionados.

4. Desarrollar un código generado en *Python* para el procesamiento del conjunto de datos.

5. Aplicar el algoritmo de aprendizaje de máquina supervisado XGBoost.

1.3. Alcance

Elaborar un modelo de aprendizaje de máquina supervisado a través de la herramienta de programación Python y sus librerías, además del análisis de al menos 100 carpetas de pozos perforados en un campo petrolero ecuatoriano, para analizar y predecir problemas operacionales que ocurren durante la perforación de un pozo.

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 perforación, sus síntomas y soluciones
(Lake & Mitchell, 2006)		
(Azar & Robello Samuel, 2007)		
(Hossain & Islam, 2018)		

Autor y Año	Párametros de Entrada	Finalidad / Contribución
(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 con aprendizaje de máquina. 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.

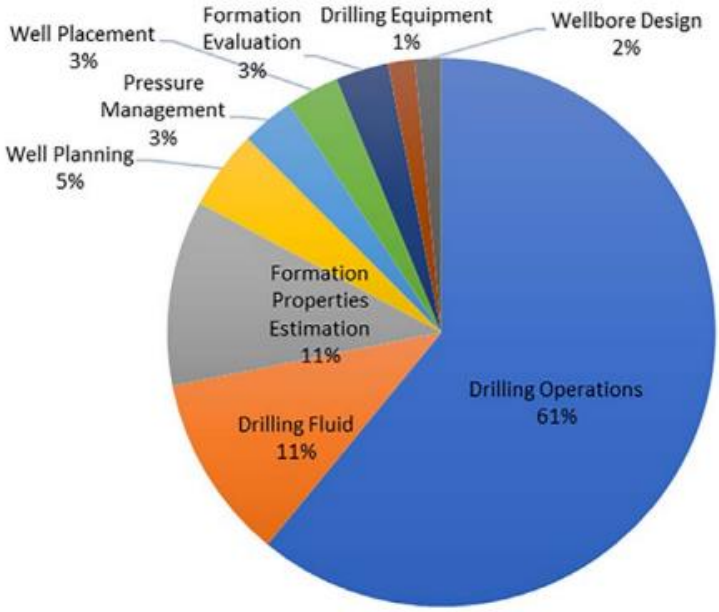

Figura 1: Porcentaje de aplicación de aprendizaje de máquina en la industria petrolera.

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) indican 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.

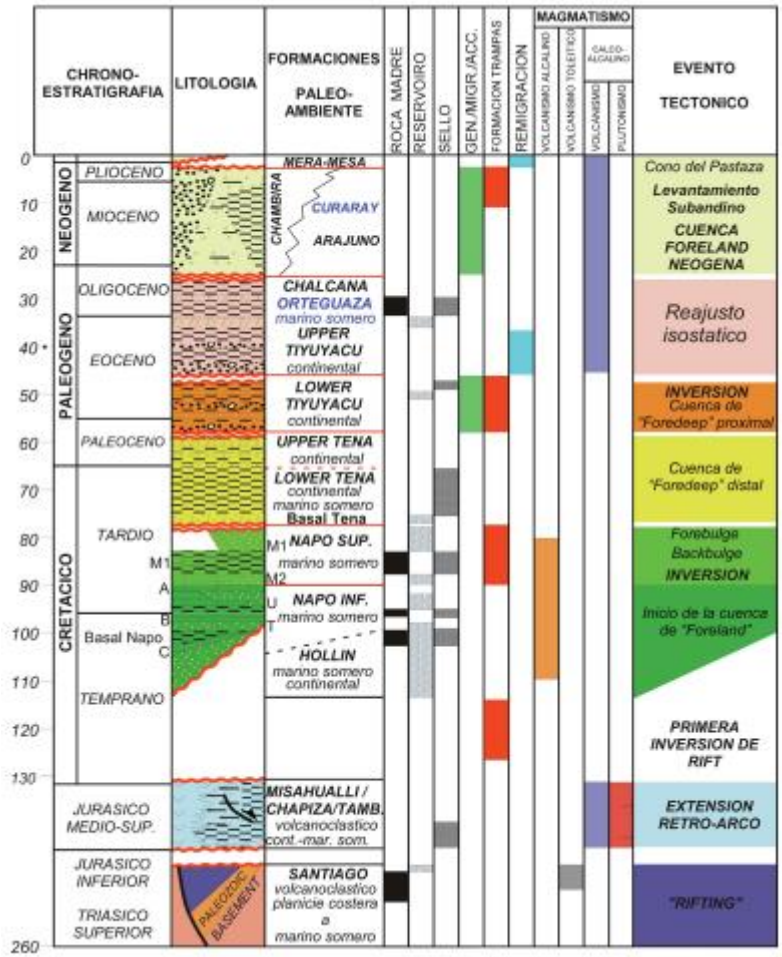

Figura 2: Columna estratigráfica Cuenca Oriente.

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 y un incorrecto diseño de lodo de perforación puede comprometer el éxito de las operaciones, con posibles consecuencias negativas, como la pérdida del pozo, incapacidad para alcanzar el objetivo geológico previsto, y reventones en situaciones más graves, 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).

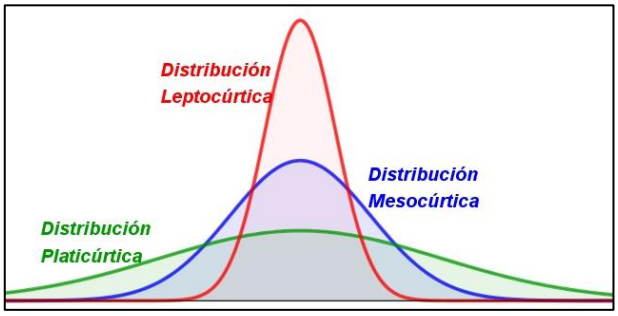

Figura 3: Curva de Curtosis

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).

[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.

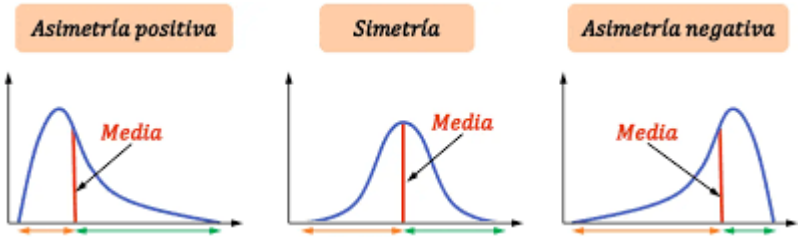

Figura 4: Curva de Asimetría

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

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).

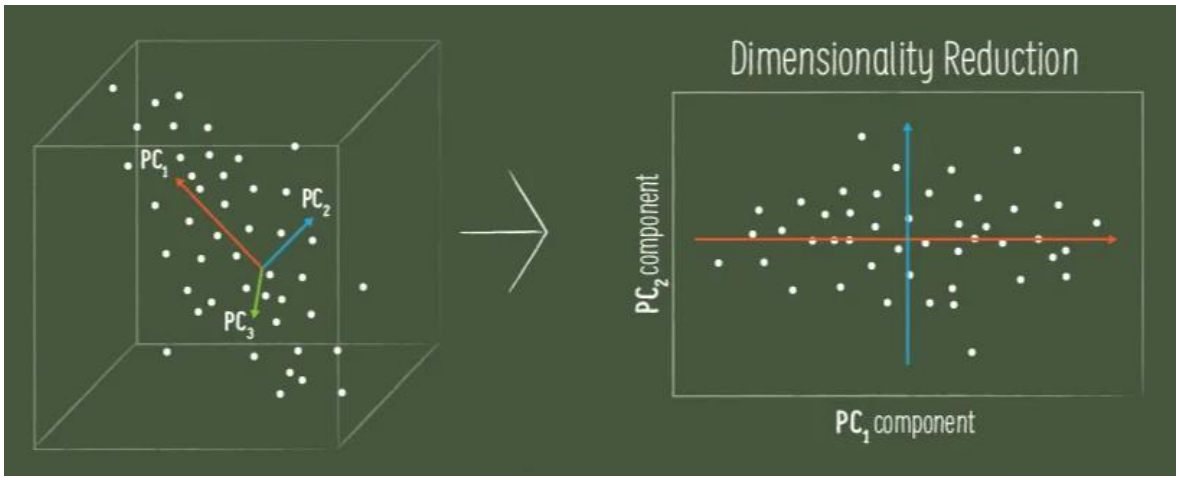

Figura 5: Análisis de Componentes Principales

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).

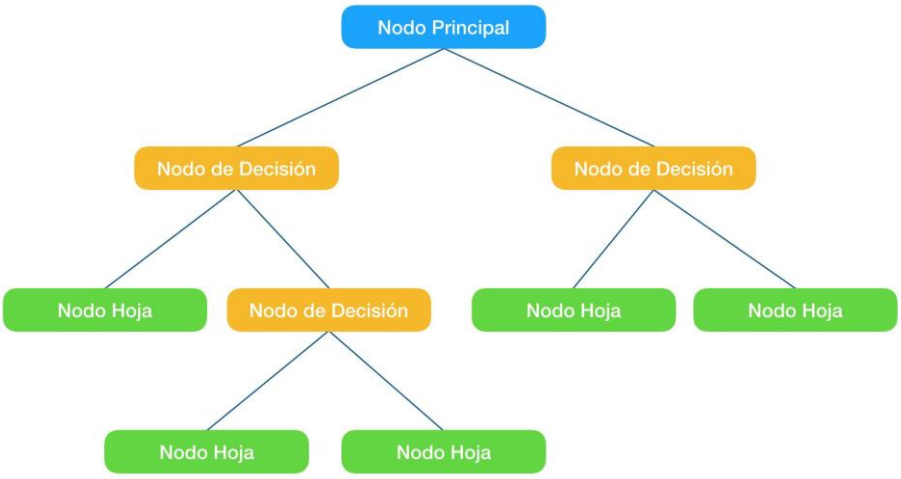

Figura 6: Estructura de un Árbol de Decisiones

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.

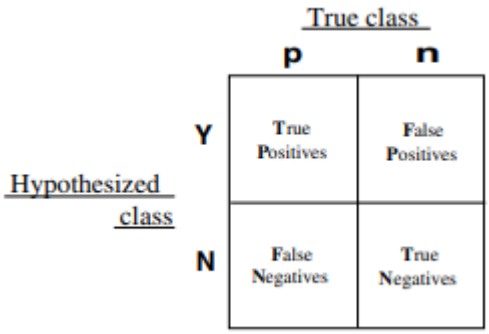

Figura 7: Matriz de Confusión

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).

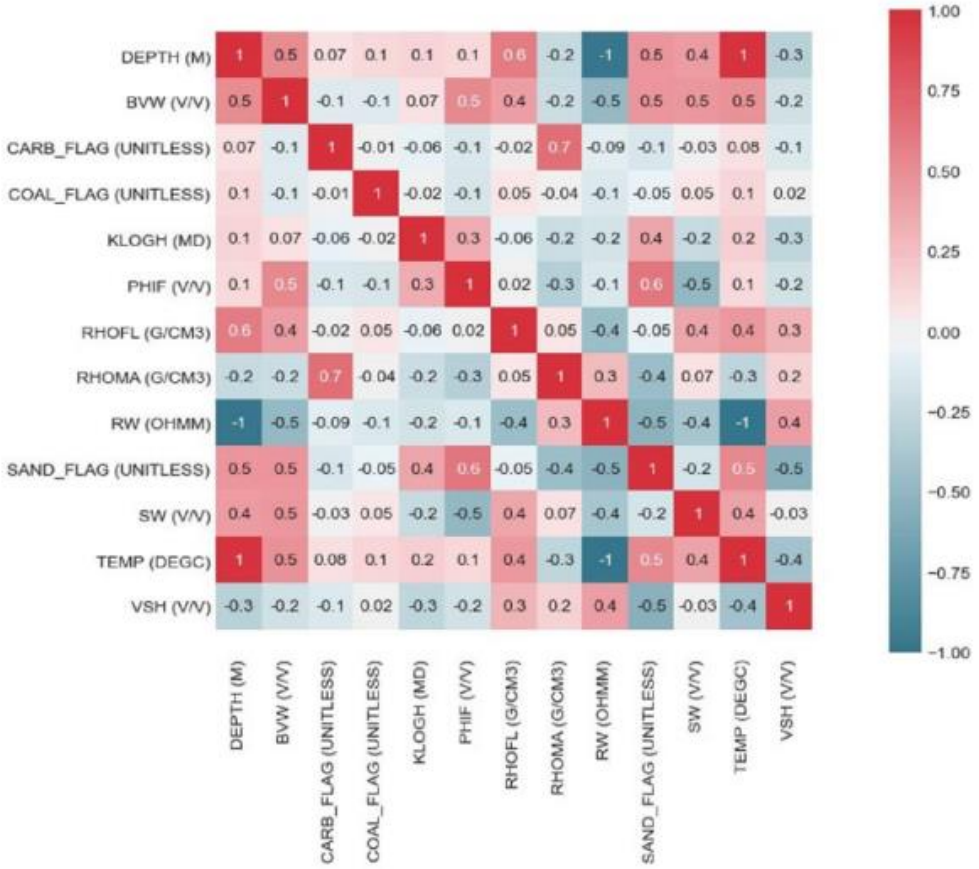

Figura 8: Mapa de Calor

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 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. Las variables que fueron utilizadas como parámetros de entrada para el desarrollo de la base de datos se enlistan en la Tabla 4.

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
Densidad	den	lpg
Viscosidad	vis	sec/qt

[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
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

2.1 Limpieza y procesamiento de datos:

- Los problemas operacionales considerados para el desarrollo de la base de datos son pega mecánica/empaquetamiento, embolamiento, pérdida de circulación, taponamiento de jets e influjos. Los problemas operacionales en tiempos planos no fueron considerados.
- Dado que los reportes de perforación no reportan las propiedades de lodo ni litología a todas las profundidades, fue necesario recurrir a diferentes fuentes de información. Por un lado, los reportes finales de fluidos permitieron rellenar los datos incompletos de parámetros de lodo con parámetros promedio por sección a perforar, mientras que, los masterlogs permitieron completar el porcentaje litológico en filas vacías.
- Las variables categóricas de sección de pozo fueron ingresadas de forma numérica con valores de 16 para sección superficial, 12.25 para sección intermedia y 8.5 para sección de producción.
- Los masterlogs no proporcionaban topes ni bases formacionales previas a Orteguaza. Por ende, se utilizó el término “Neógeno/Cuarternario” como edad geológica para categorizar las formaciones Mera, Mesa, Chambira, Curaray, Arajuno y Chalcana (Baby, Rivadeneira, & Barragán, 2014).

- En las profundidades deslizadas, los parámetros de torque y RPM fueron rellenados con un valor de cero, dado que la sarta no rota en esta condición.
- Las profundidades donde se reportaban problemas operacionales, pero no se presentaban datos completos de parámetros de perforación o propiedades de lodo, fueron rellenadas con el promedio de los valores de dichas variables del problema en cuestión.
- Aquellas profundidades reportadas como condición normal, pero que no presentaban datos completos de parámetros de perforación o propiedades de lodo, fueron rellenadas con el promedio de las variables de pozos vecinos con profundidades similares. Si el pozo vecino está ubicado en una zona geológica similar, es probable que el mismos tipo de litología y características subsuperficiales estén presentes en el pozo nuevo. Utilizar parámetros similares puede ser una estrategia efectiva para enfrentar los mismos desafíos geológicos y anticipar las condiciones de perforación.

2.2 Codificación de Variables Categóricas

En el contexto de algoritmos de aprendizaje de máquina, es frecuente que los modelos necesiten que las etiquetas de clase se representen como valores numéricos. La función de LabelEncoder, perteneciente a la biblioteca de Sckit Learn, permitió convertir las variables categóricas a numéricas, de forma que también puedan ser tomadas en cuenta para la predicción de los problemas operacionales. En las Tabla 5 y Tabla 6 se muestran los resultados de la codificación de variables.

Tabla 5: Variable categórica "Formación" a variable numérica

Formación	Valor Numérico
Basal Tena	0
Caliza A	1
Caliza B	2
Caliza C	3
Caliza M1	4
Caliza M2	5
Conglomerado Inf. Tiyuyacu	6
Conglomerado Sup. Tiyuyacu	7
Hollín Inferior	8
Hollín Superior	9

Formación	Valor Numérico
Napo	10
Neógeno/Cuarternario	11
Orteguaza	12
T Inferior	13
T superior	14
Tena	15
Tiyuyacu	16
U Inferior	17

Elaborado por: Mateo Romero

Tabla 6: 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

Elaborado por: Mateo Romero

2.3 Estandarización de variables

Este procedimiento es importante para garantizar una escala uniforme entre todas las variables, lo que facilita la interpretación de gráficas y evita que ciertas variables ejerzan un dominio excesivo en el set de datos. Además, contribuye a prevenir el sobreajuste del modelo y mejora el rendimiento general del algoritmo, favoreciendo su convergencia en un tiempo más reducido.

Dado que la estandarización utiliza la media y la desviación estándar de cada variable, fue necesario realizar este procedimiento por sección de perforación, puesto que en cada sección los parámetros de operación y propiedades de lodo son distintos.

2.4 Aplicación de Análisis de Componentes Principales

Como se explicó previamente, el PCA es una técnica de reducción de dimensionalidad que transforma las variables originales en un nuevo conjunto de variables (las componentes principales) que son combinaciones lineales de las variables originales.

Al tener un conjunto de datos con muchas variables, resulta difícil representarlo en un gráfico. Al aplicar PCA y reducir la dimensionalidad, se pueden visualizar los datos en un espacio más manejable (bidimensional o tridimensional) para entender mejor su estructura y agrupación.

Las Figura 9, Figura 10, Figura 11, Figura 12, Figura 13, Figura 14 y Figura 15, muestran la distribución de los puntos de forma individual por problema en un PCA de 2 dimensiones con el fin de resaltar patrones estructurales más significativos en el set de datos.

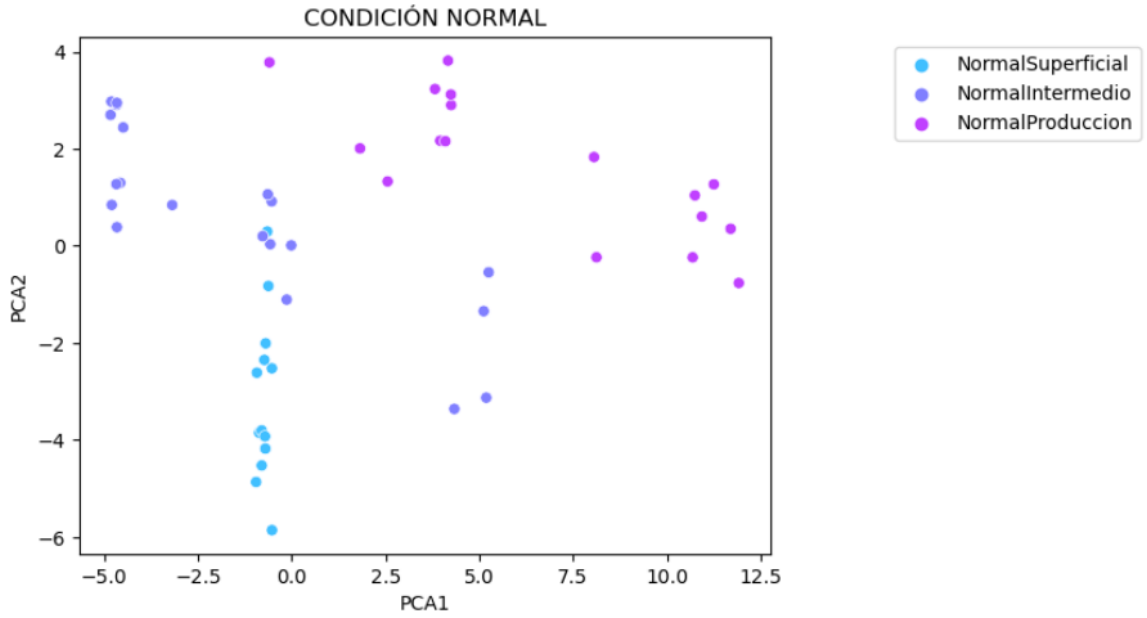

Figura 9: Análisis de Componentes Principales: Condición Norma (50 datos)

Elaborado por: Mateo Romero

Es importante destacar que, al construir el PCA de la Figura 9, se requirió emplear un conjunto de 50 datos aleatorios de condiciones normales, debido a que el conjunto original, que constaba de 1074 datos, generaba ruido en la creación del modelo, lo que afectaba negativamente su rendimiento. Por lo tanto, la selección de un subconjunto más pequeño y representativo de datos permitió una mejor agrupación de los mismos, para así obtener un modelo más preciso y con alto desempeño. El escoger una muestra mayor a 50 datos generaba problemas de distribución y confusión con otros problemas operacionales, mientras que, una muestra menor a 50 datos presentaba no lograba ser suficientemente significativa para la detección de esta condición.

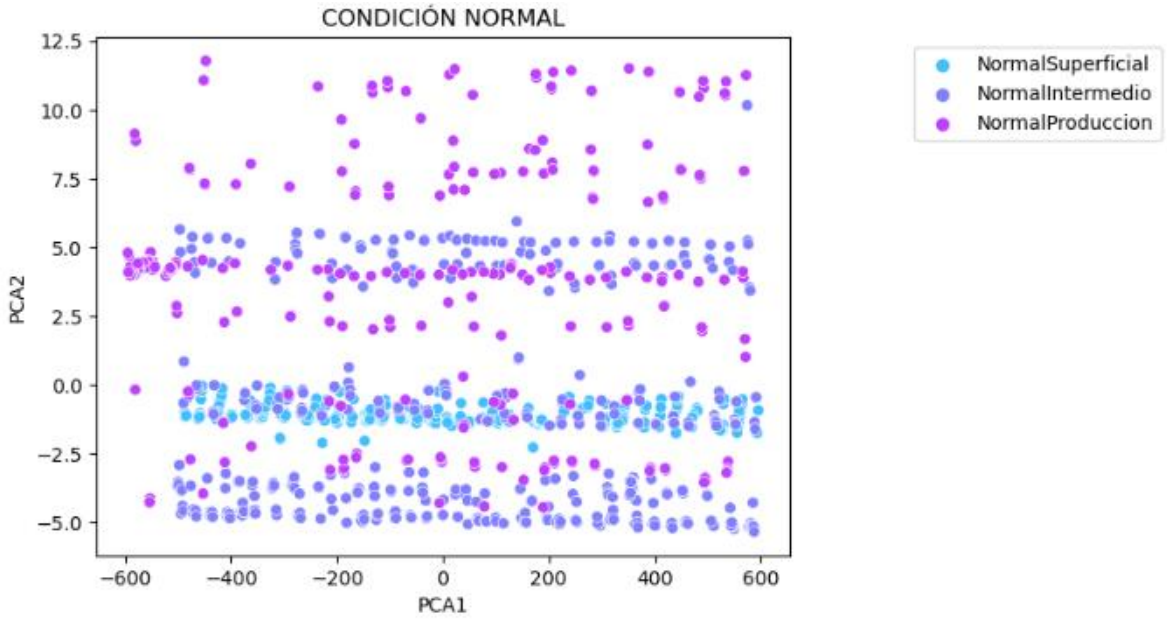

Figura 10: Análisis de Componentes Principales: Condición Normal (1074 datos)

Elaborado por: Mateo Romero

En la Figura 10 se muestra una distribución de los 1074 datos de todas las condiciones normales. En la misma se observa que no existe una agrupación realmente identificable por sección de perforación. Como se mencionó, un conjunto de datos representado de esta manera genera ruido con los demás problemas operacionales. El ruido puede agregar complejidad innecesaria al modelo y afectar su capacidad para encontrar patrones y relaciones verdaderas entre las variables, lo que resulta en predicciones menos precisas y confiables.

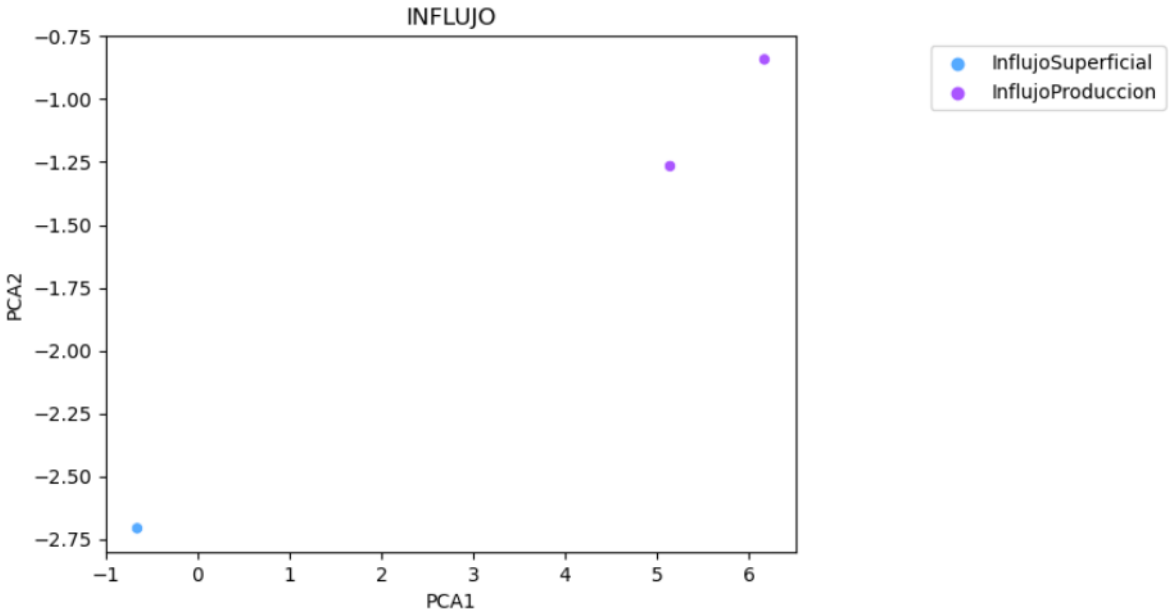

Figura 11: Análisis de Componentes Principales: Influjo

Elaborado por: Mateo Romero

En los reportes de perforación proporcionados, únicamente se encontraron 3 problemas de influjo, 2 de ellos en la sección de producción y el restante en la sección superficial (Figura 11). El uso de solo 3 datos para realizar una predicción en un modelo de aprendizaje de máquina es muy pobre debido a la falta de representatividad y escasez de información. Los modelos de aprendizaje automático generalmente necesitan una cantidad de datos lo suficientemente grande para identificar patrones y relaciones complejas entre las variables. Con solo 3 datos, el modelo no puede capturar la diversidad y variabilidad de los casos en el mundo real, lo que resulta en una predicción poco confiable. Por ende, se puede prever que la generalización del modelo a nuevos datos o situaciones será deficiente, resultando en una predicción errónea, ya que la muestra es demasiado pequeña para comprender la complejidad del problema.

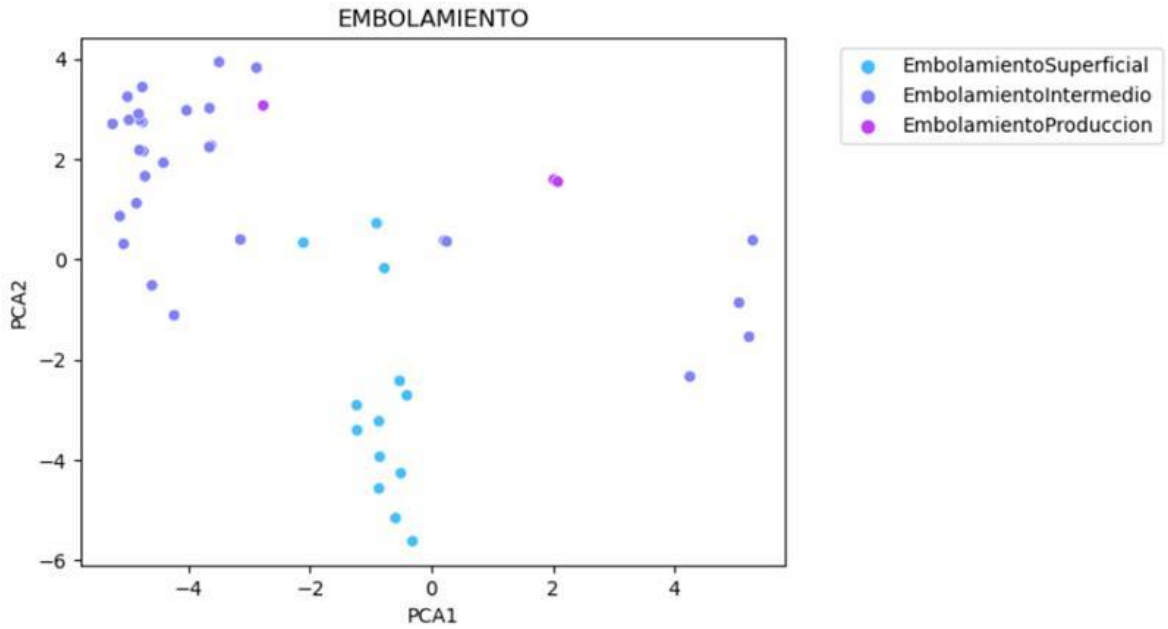

Figura 12: Análisis de Componentes Principales: Embolamiento

Elaborado por: Mateo Romero

La Figura 12 muestra una agrupación identificable de datos de embolamientos en la sección superficial (parte central inferior) y embolamientos en la sección intermedia (parte superior izquierda). Los embolamientos en la sección de producción presentan el mismo problema que los influjos, pues se cuenta únicamente con 3 datos. Por ende, se puede prever que es posible que no existan predicciones correctas para embolamientos en la sección de producción.

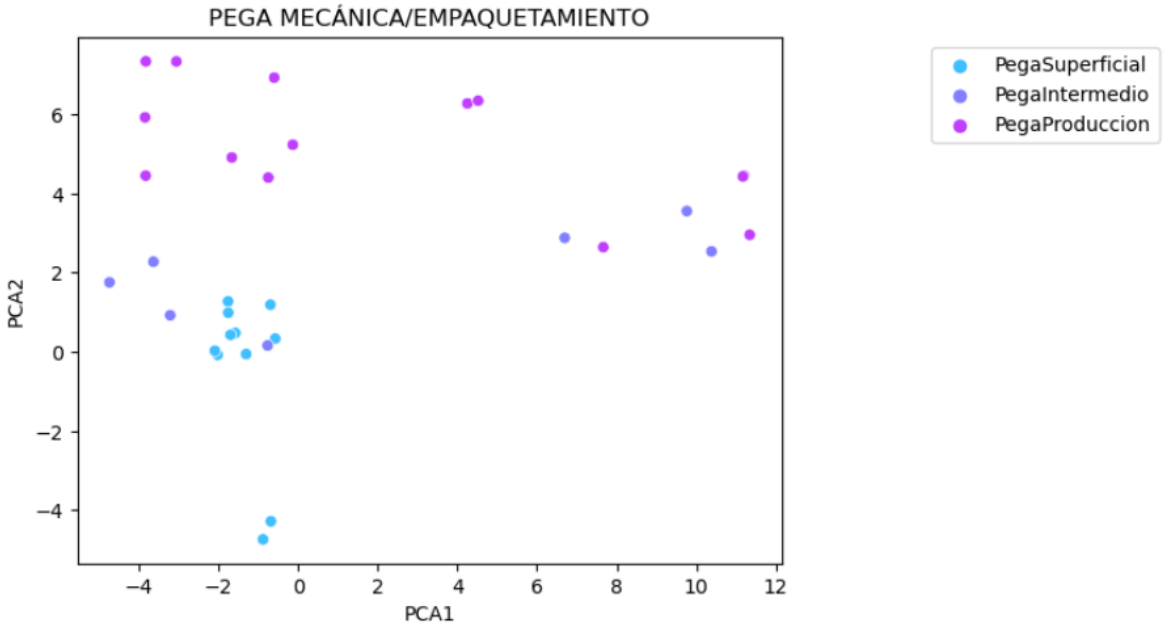

Figura 13: Análisis de Componentes Principales: Pega Mecánica/Empaquetamiento

Elaborado por: Mateo Romero

La Figura 13 muestra un conjunto de datos claramente distintivos. Por un lado, las pegas de la sección superficial se encuentran en la parte central izquierda (2 datos aislados), mientras que las pegas de la sección de producción se encuentran agrupadas en la parte superior izquierda (3 datos aislados). Finalmente, las pegas de la sección intermedia muestran una particularidad, y es que presentan 2 formas de agruparse. Mientras 4 datos se agrupan en la parte lateral izquierda, los 3 restantes se agrupan en la parte lateral derecha, aunque con una menor claridad de distribución. Dado que no existe una agrupación clara de este set de datos en esta sección, es posible que una mayor cantidad de información sobre este problema operacional sea necesaria para su correcta predicción.

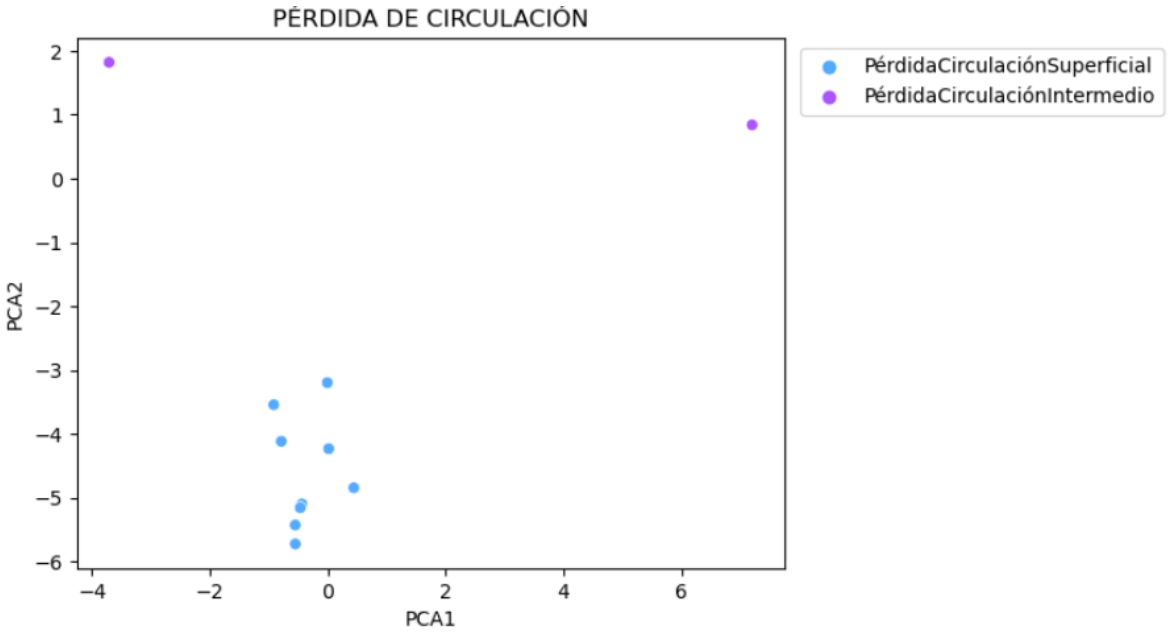

Figura 14: Análisis de Componentes Principales: Pérdida de Circulación

Elaborado por: Mateo Romero

La Figura 14 muestra una distribución del conjunto de datos totalmente distinguible. No existe ningún dato de pérdida de circulación en la sección superficial que genere ruido. No obstante, dado que se cuenta con únicamente 2 datos de pérdida de circulación en la sección intermedia, es probable que, si el modelo tuviera que predecir esta condición, resulte en respuestas incorrectas.

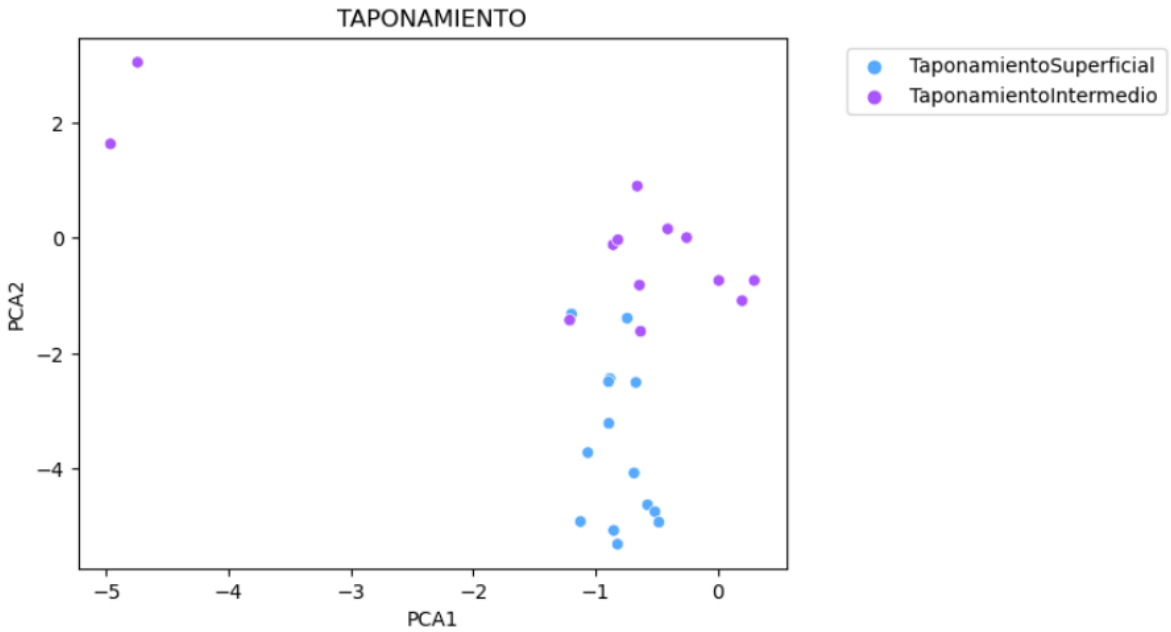

Figura 15: Análisis de Componentes Principales: Taponamiento

Elaborado por: Mateo Romero

La Figura 15 muestra una fácil identificación de la agrupación del conjunto de datos, pues solo se observan 2 puntos aislados en el caso de taponamientos intermedios.

Una vez los problemas operacionales fueron visualmente identificables a lo largo del gráfico bidimensional de análisis de componentes principales, se procedió a desarrollar el modelo de aprendizaje de máquina. El modelo fue desarrollado con el 70% de los datos, los cuales sirvieron como base de entrenamiento. De esa forma, el 30% restante se utilizará como base de prueba para comprobar su precisión y rendimiento real.

A continuación se detallan los hiperparámetros utilizados y cómo influyeron en el rendimiento y la precisión del modelo desarrollado:

- *n_estimators:* Determina la cantidad de árboles que serán utilizados en el modelo. Un valor más elevado puede resultar en una mejora del rendimiento del modelo, pero también puede implicar un tiempo de entrenamiento más prolongado y un mayor riesgo de sobreajuste.
- *learning_rate:* Controla la tasa de aprendizaje. Un valor de tamaño reducido puede ocasionar que el modelo tenga una velocidad de aprendizaje lenta y necesite muchas iteraciones para alcanzar la convergencia, mientras que un valor excesivamente grande podría provocar que el modelo no converja o incluso que se sobreajuste a los datos de entrenamiento.
- *max_depth:* Indica la profundidad máxima que se aplicará a cada árbol. Un valor mayor posibilita que el modelo aprenda relaciones más complejas en los datos, aunque también incrementa el riesgo de sobreajuste.
- *random_state:* Facilita la reproducibilidad del proceso, lo que significa que al volver a entrenar el modelo utilizando la misma semilla, se obtendrán resultados idénticos.

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

Las gráficas bidimensionales para el análisis de componentes principales (Figura 9, Figura 10, Figura 11, Figura 12, Figura 13, Figura 14 y Figura 15) permitieron observar gráficamente la distribución del set de datos para cada sección de los problemas operacionales utilizados en esta investigación. La Tabla 7 indica de forma numérica el rango en el cual es posible identificar claramente cada problema operacional en relación con los valores de PCA1 y PCA2.

Tabla 7 Rangos máximos y mínimos de PCA en los problemas operacionales

Problema Operacional	Sección	PCA1min	PCA1max	PCA2min	PCA2max
Condición Normal	Superficial	-0,93	-0,53	-5,86	-0,83
Condición Normal	Intermedio	-0,01	-4,83	0.01	2,97
Condición Normal	Producción	-0,59	11,92	-0,76	3,82
Embolamiento	Superficial	-2,1	-0,31	-5,62	0,72
Embolamiento	Intermedio	-5,24	-2,88	-1,55	3,94
Pega Mecánica/Empaquetamiento	Superficial	-2,09	-0,57	-0,07	1,27
Pega Mecánica/Empaquetamiento	Producción	-3,84	4,52	4,43	7,34
Pérdida Circulación	Superficial	-0,92	0,44	-5,72	-3,2
Taponamiento	Superficial	-1,2	-0,48	-5,3	-2,43
Taponamiento	Intermedio	-1,21	0,3	-1,62	0,9

El conjunto de datos correspondiente a las condiciones normales en la sección superficial muestra una agrupación en áreas claramente distintas en comparación con las otras secciones. Esta diferenciación se debe a la presencia de ambos componentes principales en valores negativos. En cuanto a las condiciones normales en la sección intermedia, se distinguen de la sección de producción al poseer un valor negativo en el primer componente principal a lo largo de todo el rango. En relación a los casos de embolamientos, los problemas detectados en la sección superficial tienden a formar un grupo más hacia la derecha en comparación con los de la sección intermedia, conforme se refleja en el PCA1. En lo que respecta a los incidentes de pegas mecánicas o empaquetamientos, los problemas encontrados en la sección de producción se agrupan por encima de los de la sección superficial, según lo indicado por el PCA2. Por último, en el contexto de los taponamientos en la sección intermedia, su agrupación tiende a posicionarse por encima de los taponamientos en la sección superficial, tal como lo señala el componente principal 2. Como se observa, cada problema presenta una forma distinta de agruparse, dependiendo de la sección en la que ocurrió. Esta distribución tan distinguible permitirá que el modelo clasifique los problemas operacionales de forma correcta.

Durante el entrenamiento del modelo XGBoost, se emplearon varios hiperparámetros fundamentales que influyeron de manera importante en el rendimiento y la precisión del

modelo. Estos se configuraron previamente al inicio del entrenamiento y afectaron la construcción de los árboles de decisión y el proceso de optimización. Es importante mencionar que cada hiperparámetro controla aspectos específicos del modelo y su elección puede tener un impacto significativo en los resultados obtenidos.

El modelo presenta las siguientes configuraciones: n_estimators=38, learning_rate=0.1, max_depth=15 y random_state = 42. Con ello, se obtiene la siguiente matriz de confusión.

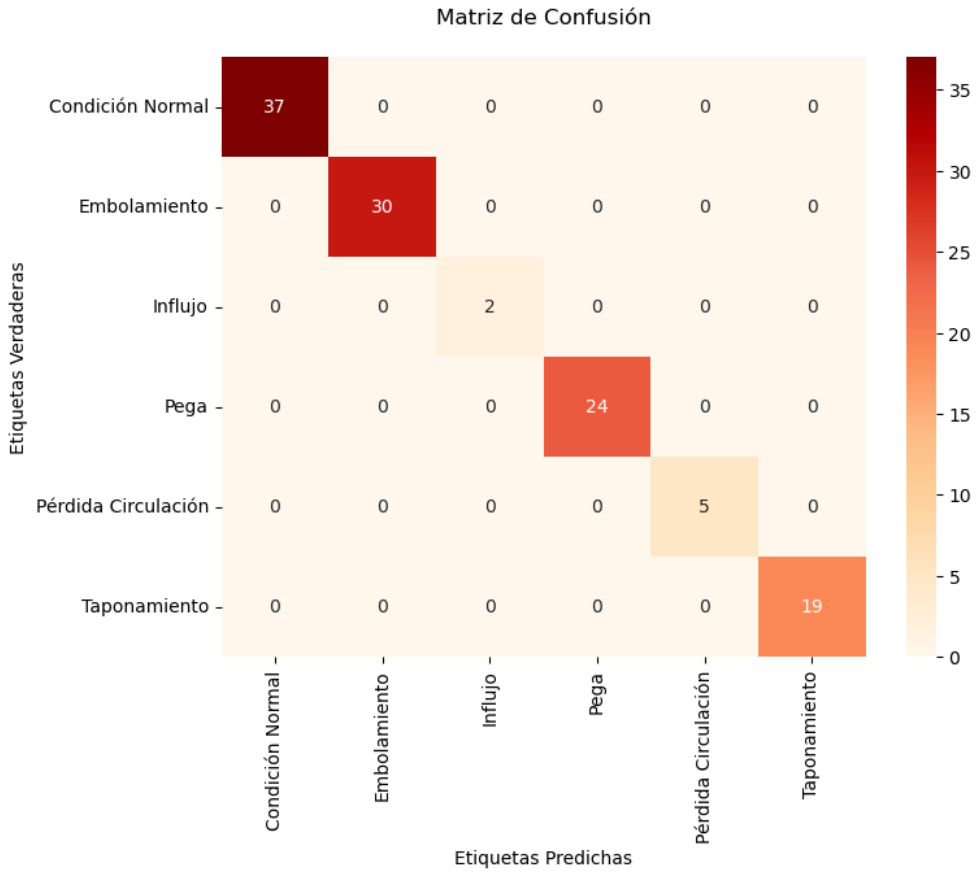

Figura 16: Matriz de Confusión (accuracy=1.0)

Elaborado por: Mateo Romero

La Figura 16 muestra que no existen errores de predicción de los problemas operacionales durante la perforación, incluidas las condiciones normales. Así, el modelo obtiene una precisión de 100%. No obstante, tener una precisión perfecta en el conjunto de entrenamiento es un indicador clásico de sobreajuste, ya que el modelo ha memorizado los datos de entrenamiento y no puede manejar correctamente nuevas instancias que no ha visto antes.

Para evaluar si el modelo está sobreajustado, es importante verificar su rendimiento en un conjunto de prueba que no se haya utilizado durante el entrenamiento. Si la precisión en el

conjunto de prueba es significativamente menor que en el conjunto de entrenamiento, es probable que exista sobreajuste. Un modelo con una precisión de 1 en el conjunto de entrenamiento puede ser sospechoso, y se debe realizar una evaluación más completa.

Por tal motivo, se ajustaron los hiperparámetros del modelo con el fin de reducir su precisión, pero para posiblemente garantizar un mejor rendimiento en la base de prueba.

El modelo ajustado presenta las siguientes configuraciones: n_estimators=20, learning_rate=0.05, max_depth=3 y random_state = 42. Con ello, se obtiene la siguiente matriz de confusión

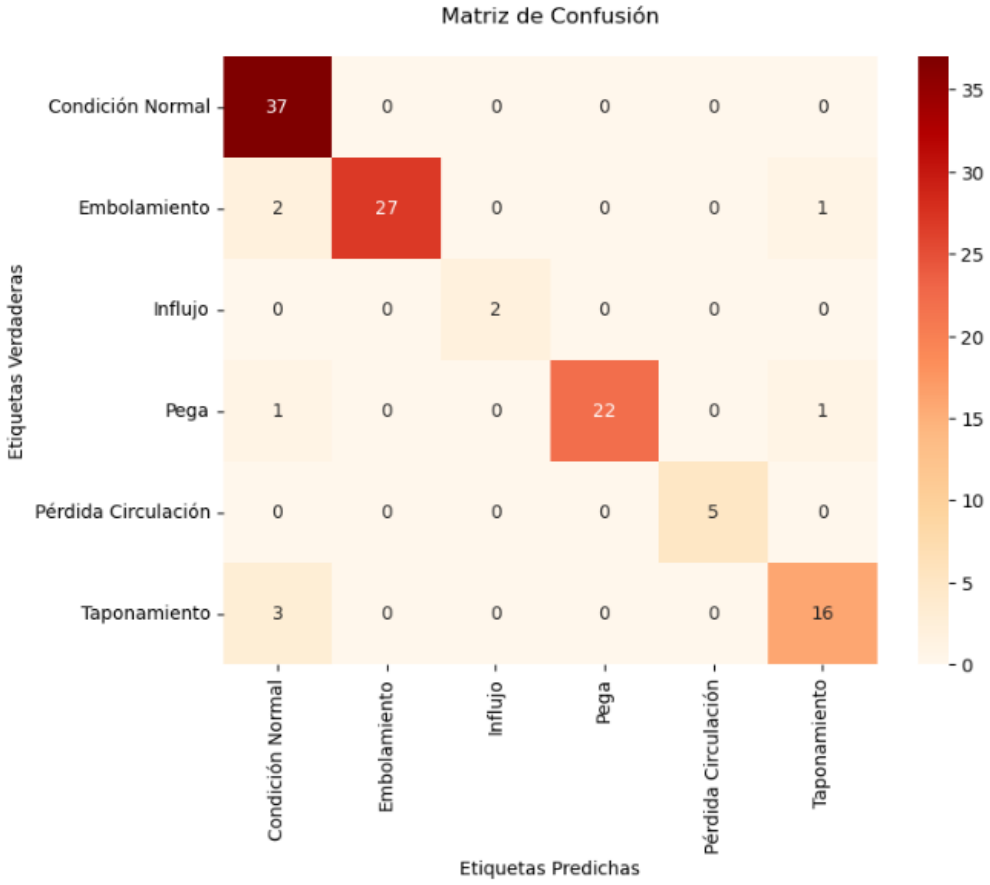

Figura 17: Matriz de Confusión (accuracy=0.93)

Elaborado por: Mateo Romero

La matriz de confusión de la Figura 17 tiene una precisión del 93%. Un modelo con una precisión de entrenamiento menor a 100% suele tener una mejor capacidad para generalizar a nuevos datos que no se han utilizado durante el entrenamiento. Es decir, el modelo no se ha ajustado excesivamente a los datos de entrenamiento y tiene una mayor probabilidad de realizar predicciones precisas en datos no vistos. Por ende, se puede prever que este modelo dará mejor resultado que el primero modelo presentado.

Sin embargo, no hay un valor específico al que se deba disminuir la precisión en la base de entrenamiento, debido a que esto depende de varios factores, como la complejidad del modelo, la cantidad de datos, la naturaleza del problema, los hiperparámetros a utilizar y los recursos del computador.

La reducción en la precisión en la base de entrenamiento puede variar según el caso. En general, un descenso en la precisión podría indicar que el modelo está capturando patrones más generales en lugar de simplemente memorizar los datos de entrenamiento. El objetivo principal al tratar con un modelo sobreajustado es alcanzar un equilibrio entre la precisión en el conjunto de entrenamiento y la capacidad del modelo para generalizar a nuevos datos. Se puede experimentar con diferentes ajustes en los hiperparámetros y técnicas de regularización hasta que se encuentre un punto en el que el modelo generalice de manera adecuada sin comprometer en exceso su capacidad para aprender patrones importantes.

Finalmente, se presenta la Figura 18, la cual representa las fronteras de decisión. Las fronteras de decisión son límites o líneas que separan diferentes clases o categorías en un modelo de aprendizaje de máquina. De esa forma, se puede conocer en qué zona se agrupa un determinado punto y comprobar si realmente está siendo predicho o no por el modelo.

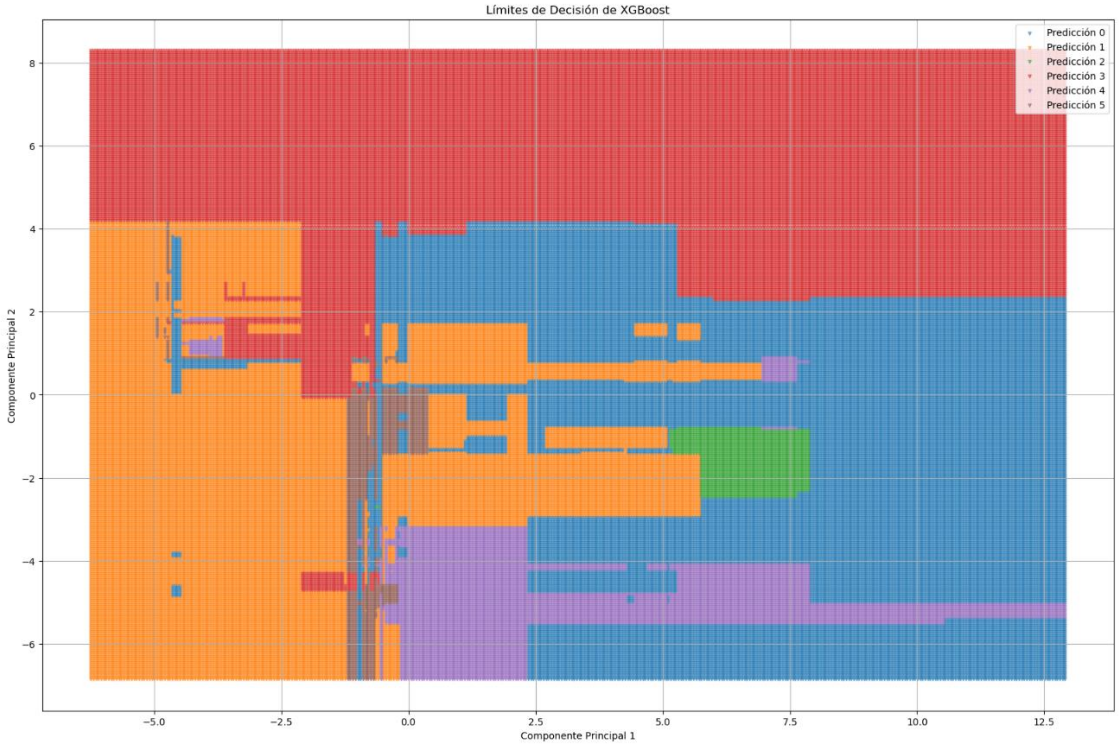

Figura 18: Límites de Decisión

Elaborado por: Mateo Romero

3.2 Conclusiones

- Se seleccionó pozos de distintas zonas en el campo de forma que exista diversidad de condiciones geológicas y geográficas, es decir, una mejor distribución de datos en el campo. La información fue validada de forma que los datos puedan tener calidad en la base desarrollada.

- Se generó una base con información de 104 pozos, incluyendo parámetros operacionales de perforación, geometría del hoyo, propiedades del lodo, características geológicas con el fin de detallar la condición operacional de la mejor manera posible.

- Los modelos de aprendizaje máquina más complejos, como el XGBoost, pueden ser difíciles de entrenar con pocos datos, ya que requieren más información para ajustar sus parámetros de manera efectiva. Los problemas operacionales como el influjo (3 datos), la pérdida de circulación en la sección intermedia (2 datos) o el embolamiento de la sección de producción (2 datos) no representan un patrón claro de agrupación. Por lo tanto, se puede prever que la generalización del modelo a nuevos datos o situaciones será deficiente, dado que el modelo no puede capturar la diversidad y variabilidad de los casos en el mundo real, lo que resulta en una predicción poco confiable.

- Al representar las 1074 condiciones normales para presentaba la base de datos desarrollada, no es posible identificar una agrupación claramente distinguible por sección de perforación. Un conjunto de datos organizado de esta manera introduce interferencias con otros problemas operacionales. Esta interferencia puede introducir complejidad innecesaria en el modelo y repercutir en su capacidad para distinguir patrones y relaciones auténticas entre las variables. Como resultado, las predicciones carecían de precisión y fiabilidad. Como resultado, se escogieron 50 datos aleatorios de condiciones normales para generar representatividad de este conjunto de datos.

- Se crearon 2 modelos de aprendizaje de máquina con precisiones de predicción distintos en el conjunto de entrenamiento, con 100% y 93% respectivamente. En el primer caso, los hiperparámetros utilizados permiten ahorrar tiempo y recursos del computador al encontrar la configuración óptima más rápidamente y evitar entrenamientos redundantes (utiliza el número de árboles necesarios para un correcto entrenamiento. En el segundo caso, se utiliza una tasa de aprendizaje más

baja, lo que ocasiona un tiempo de entrenamiento más prolongado, pero que en consecuencia, reduce el sobreajuste del modelo.

- Es normal que los modelos tengan un rendimiento mejor en el conjunto de entrenamiento que en el de prueba, pero la diferencia no debe ser excesiva. Sin embargo, el objetivo es encontrar un equilibrio para que el modelo generalice bien a datos nuevos y no se sobreajuste a los datos de entrenamiento.

3.3 Recomendaciones

- Recopilar la mayor cantidad de información posible. El rendimiento y la confiabilidad de los modelos de aprendizaje dependen en gran medida de la cantidad y diversidad de los datos en los que se entrenan. Un conjunto de datos sustancial permite que el modelo capture patrones, relaciones y matices complejos dentro de la información, lo que resulta en una mejor capacidad de generalización y predicción.

- Validar y limpiar la información recolectada para el desarrollo de la base de datos. Los datos sucios, incompletos o incorrectos pueden afectar significativamente la calidad y confiabilidad del modelo. La presencia de valores atípicos, datos faltantes o ruidosos puede llevar a resultados imprecisos y decisiones erróneas. Los datos limpios y bien estructurados permiten una mayor eficiencia durante el proceso de entrenamiento, se reduce el tiempo y los recursos necesarios para ajustar el modelo.

- Estandarizar el conjunto de datos. Si las características tienen escalas muy diferentes, las medidas de distancia pueden verse dominadas por las características con escalas más grandes, lo que afecta la capacidad del modelo para identificar patrones precisos. Los datos pueden converger más rápidamente cuando las características están en una escala similar.

- Comprender cómo cada hiperparámetro afecta el comportamiento y la complejidad del modelo de aprendizaje de máquina. Realizar ajustes incrementales en los hiperparámetros en lugar de cambios drásticos permitirá entender cómo cada ajuste afecta el rendimiento 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

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*.