

**IMPLEMENTACIÓN DE UN MODELO PREDICTIVO PARA LA ESTIMACIÓN DE
UN INDICADOR DE BOREHOLE QUALITY, MEDIANTE LA METODOLOGÍA
MACHINE LEARNING, EN LOS POZOS PERFORADOS DURANTE LA
CAMPAÑA 2017-2020 DE DOS CAMPOS DE HIDROCARBUROS DE LA CUENCA
DEL VALLE MEDIO DEL MAGDALENA**

**JUAN FELIPE ANAGUANO BUCHELI
WILLIAM SNEYDER URREGO QUIMBAY**

**Proyecto integral para optar por el título de
Ingeniero de Petróleos**

**Orientador
SEBASTIÁN ALEJANDRO GÓMEZ ALBA
Ingeniero de Petróleos**

**FUNDACIÓN UNIVERSIDAD DE AMÉRICA
FACULTAD DE INGENIERÍAS
PROGRAMA DE INGENIERÍA DE PETRÓLEOS
BOGOTÁ D. C**

2021

NOTA DE ACEPTACIÓN

Director
Ing. Sebastián Gómez Alba

Jurado 1
Ing. Camilo Guerrero Martinez

Jurado 2
Ing. Yasin Naman Medina

Bogotá D.C., julio de 2021

DIRECTIVOS DE LA UNIVERSIDAD

Presidente de la Universidad y Rector del Claustro

Dr. MARIO POSADA GARCIA PEÑA

Consejero Institucional

Dr. LUIS JAIME POSADA GARCÍA-PEÑA

Vicerrectora Académica y de Investigaciones

Dra. ALEXANDRA MEJÍA GUZMÁN

Vicerrector Administrativo y Financiero

Dr. RICARDO ALFONSO PEÑARANDA

Secretaria General

Dra. JOSÉ LUIS MACÍAS RODRIGUEZ

Decano Facultad de Ingeniería

Ing. JULIO CÉSAR FUENTES ARISMENDI

Director Programa de Ingeniería de Petróleos

Ing. JUAN CARLOS RODRÍGUEZ ESPARZA

DEDICATORIA

Quiero dedicar este proyecto a Dios, a mi madre, que en su cariño y apoyo encontré la inspiración para iniciar y terminar mis estudios. A mis abuelos y mi tío Vla, sus grandes enseñanzas y apoyo en todo sentido fueron fundamentales. A mi hermanito Carlos que desde el cielo siempre me acompaña y tantas cosas de la vida me enseñó. A mis mejores amigos Jorge, Adri y Juanfer por su amistad y cariño. Por último y no menos importante, a mi compañero y por sobre todas las cosas mi gran amigo Willi, gracias por su dedicación y compromiso con este proyecto.

Juan Felipe Anaguano Bucheli

DEDICATORIA

Quiero dedicar este trabajo a mis padres, a mi madrina, a mi abuela, a mi abuelo que me cuida desde el cielo, a mis docentes de la Universidad de América que me nutrieron durante este largo camino, a la ingeniera Adriana Valderrama que nos apoyó como directora de tesis, al docente Sebastián como tutor de tesis y fiel acompañante, a mis amigos que siempre estuvieron ahí, a las personas especiales que me alentaron en los momentos difíciles, a mi compañero y amigo incondicional Juan Felipe que me colaboró y aguantó durante la realización de este proyecto.

William Urrego Quimbay

AGRADECIMIENTOS

Agradecemos a la ingeniera Adriana Valderrama por su apoyo y disposición para gestionar los trámites dentro de la compañía Ecopetrol S. A.

A la Fundación Universidad de América por ser el lugar donde tuvo lugar nuestra formación como profesionales e inculcarnos los valores que hacen a una persona correcta.

Al ingeniero Sebastián Gómez por su total e incondicional apoyo, siendo una persona fundamental en el desarrollo del proyecto.

A la compañía Ecopetrol S.A. por brindarnos el apoyo, soporte técnico e información necesaria para llevar a cabo este proyecto de investigación.

Las directivas de la Universidad de América, los jurados calificadores y el cuerpo docente, no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores.

TABLA DE CONTENIDO

	pág.
RESUMEN	16
INTRODUCCIÓN	17
1. MARCO TEÓRICO	20
1.1 Perforación	20
1.2 Parámetros de perforación	20
1.3 Calidad de pozo	22
1.4 Machine learning	23
<i>1.4.1 Tipos de machine learning</i>	23
<i>1.4.2 Análisis exploratorio de datos (AED)</i>	25
2. METODOLOGÍA Y DATOS	27
2.1 Obtención y selección de parámetros	29
<i>2.1.1 Cálculo de variables implícitas</i>	29
<i>2.1.1 Importación de librerías en Python y cálculo de la TVD</i>	31
<i>2.1.3 Caracterización de las variables</i>	33
2.2 Análisis exploratorio de datos (AED)	34
2.3 División del dataset en datos de entrenamiento y prueba	40
2.4 Implementación y validación del modelo predictivo	41
<i>2.4.1 Ajuste de hiper parámetros de los modelos</i>	42
<i>2.4.2 Validación de los modelos predictivos</i>	43
<i>2.4.3 Generación de mapas de calor</i>	43
<i>2.4.4 Consolidación de rangos operativos</i>	43
2.5 Construcción del indicador de calidad (WQS)	44
2.6 Análisis financiero del indicador de calidad (WQS)	45
3. ANÁLISIS DE RESULTADOS	46
3.1 Entrenamiento de los modelos predictores	46
3.2 Implementación y ajuste de modelos predictivos	47
<i>3.2.1 Ajuste de hiper parámetros (Tunning)</i>	49
3.3 Importancia de las variables	51
3.4 Validación de los modelos predictivos	53

3.4 Generación de mapas de calor	54
3.4.1 <i>Análisis de los mapas de calor</i>	59
3.5 Rangos óptimos para los parámetros de perforación	63
3.6 Indicador de calidad de pozo (WQS)	63
3.7 Generación de líneas base de MSE para los campos en estudio	70
3.7.1 <i>Caracterización estadística MSE</i>	70
3.8 Análisis financiero del WQS	74
4. CONCLUSIONES	75
BIBLIOGRAFÍA	77
ANEXOS	81

LISTA DE FIGURAS

	pág.
Figura 1. WQS para un pozo en particular	22
Figura 2. Diagrama de proceso del aprendizaje supervisado	24
Figura 3. Modelamiento de datos con aprendizaje no supervisado	24
Figura 4. Aprendizaje por refuerzo	25
Figura 5. Histograma	26
Figura 6. Diagrama de caja y bigotes	26
Figura 7. Diagrama de flujo del proyecto	28
Figura 8. Significado operativo de DDI	31
Figura 9. Código encargado de interpolar la inclinación	33
Figura 10. Caracterización de las variables en un pozo	34
Figura 11. Metodología del análisis exploratorio de datos	35
Figura 12. Histogramas de parámetros de perforación (Sección 12 ¼")	35
Figura 13. Histogramas de parámetros de perforación (Sección 8 ½")	36
Figura 14. Diagrama de caja para los parámetros de perforación (Sección 12 ¼")	37
Figura 15. Diagrama de caja para los parámetros de perforación (Sección 8 ½")	37
Figura 16. Matriz de correlación entre variables (Sección 12 ¼")	38
Figura 17. Matriz de correlación entre variables (Sección 8 ½")	39
Figura 18. Metodología a seguir para el modelo Random Forest	41
Figura 19. Metodología para la creación de WQS	44
Figura 20. Clasificación de pozos	45
Figura 21. Entrenamiento de los modelos predictivos sección (12 ¼")	46
Figura 22. Entrenamiento de los modelos predictivos sección (8 ½")	47
Figura 23. Precisión de los modelos predictivos (Sección 12 ¼")	48
Figura 24. Precisión de los modelos predictivos (Sección 8 ½")	48
Figura 25. Código para ajustar hiper parámetros	50
Figura 26. Métricas de desempeño para los modelos predictivos	51
Figura 27. Matriz de clasificación para la sección 12 ¼"	52
Figura 28. Matriz de clasificación para la sección 8 ½"	52
Figura 29. Validación ROP sección 12 ¼"	53

Figura 30. Validación ROP sección 8 ½”	54
Figura 31. Mapa de calor ROP, sección 12 ¼”	55
Figura 32. Mapa de calor ROP, sección 8 ½”	56
Figura 33. Mapa de calor MSE sección 12 ¼”	57
Figura 34. Mapa de calor MSE, sección 8 ½”	58
Figura 35. Porcentaje de coincidencia de variables, sección 12 ¼”	64
Figura 36. Porcentaje de coincidencia de variables, sección 8 ½”	65
Figura 37. Peso de variables en WQS, sección 12 ¼”	66
Figura 38. Peso de variables en WQS, sección 8 ½”	67
Figura 39. WQS para los pozos en estudio	68
Figura 40. Análisis estadístico MSE, sección 12 ¼”	70
Figura 41. Análisis estadístico MSE, sección 8 ½”	71
Figura 42. Mapa de calor MSE, sección 12 1/4"	72
Figura 43. Mapa de calor MSE, sección 8 1/2"	73
Figura 44. Mapa de calor para RPM_FON, sección 12 ¼”	82
Figura 45. Mapa de calor para TORQUE, sección 12 ¼”	83
Figura 46. Mapa de calor para WOB, sección 12 ¼”	84
Figura 47. Mapa de calor para DDI, sección 12 ¼”	85
Figura 48. Mapa de calor para RPM_FON, sección 8 ½”	86
Figura 49. Mapa de calor para TORQUE, sección 8 ½”	87
Figura 50. Mapa de calor para WOB, sección 8 ½”	88
Figura 51. Mapa de calor para DDI, sección 8 ½”	89
Figura 52. Validación para el modelo predictivo de DDI, sección 12 ¼”	90
Figura 53. Validación para el modelo predictivo de WOB, sección 12 ¼”	91
Figura 54. Validación para el modelo predictivo de TORQUE, sección 12 ¼”	92
Figura 55. Validación para el modelo predictivo de RPM_FON, sección 12 ¼”	93
Figura 56. Árbol de decisión ROP, sección 12 ¼”	94
Figura 57. Árbol de decisión TORQUE, sección 12 ¼”.	95
Figura 58. Árbol de decisión WOB, sección 12 1/4".	96
Figura 59. Árbol de decisión para la RPM_FON , sección 12 1/4".	97
Figura 60. Árbol de decisión DDI, sección 12 1/4".	98

Figura 61. Árbol de decisión ROP, sección 12 1/4"

99

LISTA DE TABLAS

	pág.
Tabla 1. Parámetros de perforación a analizar.	21
Tabla 2. Librerías de Python necesarias para el proyecto	32
Tabla 3. Rangos operativos de las variables de entrada (Sección 12 ¼”)	40
Tabla 4. Rangos operativos de las variables de entrada (Sección 8 ½”)	40
Tabla 5. Primera exactitud de cada modelo predictivo (Sección 12 ¼”)	42
Tabla 6. Primera exactitud de cada modelo predictivo (Sección 8 ½”)	42
Tabla 7. Proporciones utilizadas para los modelos predictivos	49
Tabla 8. Rangos óptimos variables sección 12 ¼”	63
Tabla 9. Rangos óptimos variables sección 8 ½”	63
Tabla 10. Descripción cualitativa de los pozos	69
Tabla 11. Líneas base de MSE	73
Tabla 12. Comparación de costos y duración de perforación	74

LISTA DE ECUACIONES

	pág.
Ecuación 1. Cálculo de la MSE	21
Ecuación 2. Along Hole Displacement (AHD)	29
Ecuación 3. Tortuosidad	30
Ecuación 4. Índice de dificultad direccional	30
Ecuación 5. Cálculo para TVD	32

LISTA DE ABREVIATURAS

INCL: inclinación

ROP: tasa de penetración

WOB: peso sobre la broca

RPM_FON: revoluciones por minuto en fondo

RPM_SUP: revoluciones por minuto en superficie

WQS: indicador de calidad de pozo (Wellbore Quality Scorecard)

MD: profundidad medida

TVD: profundidad vertical verdadera

ML: Machine Learning

MSE: energía mecánica específica

AHD: desplazamiento a lo largo del pozo

DDI: índice de dificultad direccional

AED: análisis exploratorio de datos

RESUMEN

El documento presenta la implementación de un modelo predictivo para la estimación de un Borehole Quality, dicho proyecto de investigación nació a partir de la necesidad de tener valores o rangos de parámetros estandarizados como óptimos para evaluar la efectividad en las operaciones de perforación para dos campos de Valle Medio del Magdalena. En primer lugar, se generó una base de datos con base en los datos de los pozos perforados en el periodo 2017-2020 con perfil direccional vertical, dicha base de datos fue depurada mediante un análisis exploratorio de datos con el fin de prepararla para la división y prueba del modelo predictivo. Una vez validada la base de datos se implementó un algoritmo de aprendizaje automático supervisado como lo es el Random Forest Regressor el cual fue evaluado con el error cuadrático medio, teniendo como variables de entrada revoluciones por minuto en superficie y fondo (ROP_SUP), (ROP_FON) respectivamente, tasa de penetración (ROP), torque (TORQUE), peso de la broca (WOB), índice de dificultad direccional (DDI), dichas variables de entrada fueron también las variables resultantes con el fin de tener los rangos óptimos de cada una, siendo estas parte de la ecuación de la energía mecánica específica (MSE). Los rangos óptimos de operación fueron establecidos mediante la representación gráfica de las mismas en mapas de calor; teniendo en cuenta que el algoritmo arrojó exactitudes entre (72.18%-95.41%), y (70.35%-94%) en las secciones 12 ¼” y 8 ¾” respectivamente. En segundo lugar, el indicador de calidad (WQS) fue validado al comparar los datos predichos con los datos reales, obteniendo puntajes entre (6-9.5) y (7-13) para las secciones 12 ¼” y 8 ¾” respectivamente con base en la matriz de validación. En tercer lugar, se generó la línea base de energía mecánica específica (MSE) para los dos campos de hidrocarburos utilizados como objeto de estudio. Finalmente, se realizó un análisis financiero con base en la tasa de penetración.

PALABRAS CLAVE: Machine Learning, Borehole Quality (BQ), Random Forest, tasa penetración (ROP), peso sobre broca (WOB), indicador direccional dificultad (DDI), energía mecánica específica (MSE).

INTRODUCCIÓN

En la Cuenca del Valle Medio Del Magdalena hay dos Campos de Hidrocarburos operados por la empresa Ecopetrol S.A, en los cuales se han desarrollado operaciones en el periodo de tiempo de 2017-2020 sin tener un análisis detallado de la calidad del hueco, bajo la premisa de que dicho ítem carece de estandarización en la empresa, estableciendo así las secciones 12 ¼” y 8 1/2” como objetivo primordial puesto que en estas se presenta la mayor cantidad de operaciones. Dicho esto, hay una variabilidad notable en los parámetros de perforación, no obstante, la profundidad final y el perfil direccional son muy similares, lo cual deja una incertidumbre al determinar cuál ha sido el mejor desempeño alcanzado y por consiguiente vacíos estadísticos al momento de analizar la eficiencia de las operaciones

Debido a que la calidad de hueco no se encuentra estandarizada en el ámbito técnico, este proyecto propone la implementación de un indicador de Borehole Quality mediante la metodología de análisis de datos Machine Learning, cuyo objetivo es evaluar el desempeño obtenido en perforaciones concluidas y contrastarlo con un modelo de predictibilidad; dando a conocer los posibles resultados que podría alcanzar las operaciones al tener en cuenta el control de parámetros de perforación tales como (WOB, ROP, Torque, DDI entre otros), siendo así, un proyecto que entregara una herramienta digital que a futuro podrá tener un impacto en la eficiencia operativa así, generando un valor agregado al análisis de parámetros de perforación en la industria

Este análisis se ha implementado en ocasiones pasadas ,tal como se demuestra en el trabajo de grado titulado “Evaluación técnico-financiera de las operaciones de perforación para la disminución de los tiempos no productivos durante los viajes en los pozos seleccionados del campo Castilla”, de los autores Juan Sebastián Robayo Perdomo y Lina Paola Rodríguez Peña, donde después de realizar una descripción del campo, su localización, vías de acceso y características geológicas del petróleo presente en el mismo; procedieron a elaborar una selección de pozos representativos, con ciertas características, que permitan una evaluación detallada de sus tiempos no productivos, sus causas, soluciones y disminución a futuro. Dando como resultado que las fallas operacionales son las situaciones que generan en mayor cantidad la presencia de NPT, con porcentajes respectivamente de: 74,47 %, 55,75 %, 90,21 %, 60,23 %, 73,64 % para los pozos

Castilla 232, Castilla 234, Castilla 348, Castilla 359 y Castilla 469. Además, el mantenimiento o cambio de equipos con anticipación, como las bombas de superficie, garantiza una disminución aproximada del 15 % para los futuros pozos del campo. Por otro lado, la tesis denominada Aplicación de una metodología propuesta para el diagnóstico y el análisis de las causas raíz de los problemas operacionales durante la perforación de los pozos en el campo Castilla, de los autores Paula Alejandra Espinoza Franco y Andrés Felipe Góngora Sarmiento, implementa una metodología propuesta por la Shell International Exploration para el análisis y prevención de causas y problemas operacionales. Haciendo uso de esta metodología, se llevó a cabo el método propuesto por la compañía y arrojó los siguientes resultados: Los NPT representan aproximadamente el 45 % (9 días) de una perforación normal. En toda la campaña de perforación, el problema de pega de tubería fue el que tuvo más recurrencia con un porcentaje del 41 % equivalente a 5425 horas de NPT, por lo que es el más crítico. Seguido a este se encuentra los problemas de hueco con un 33 % (4324,4 horas) y el último problema de alta criticidad está relacionado con las herramientas en el subsuelo con una participación del 18 % (2429 horas). Lo anterior arroja un cálculo aproximado de 27'125.000 USD anuales, los cuales corresponden a las 5425 horas de NPT provenientes de las pegas de tubería, teniendo en cuenta que el costo promedio de 1 hora de taladro es de 5000 USD; por esta razón es importante un análisis de los posibles problemas operacionales y una ejecución de estrategias que ayuden a mitigar el surgimiento de NPT a lo largo de las actividades de perforación. Y Por Ultimo en el estudio “ Borehole Quality Design and Practices to Maximize Drill Rate Performance” de los autores F.E Dupriest, W. C. Elks donde hicieron una descripción detallada del nuevo proceso instaurado para el manejo y análisis de la calidad del hueco , evaluándolo con los NPTS y el control de la MSE, junto a la estandarización de unos limitantes de estabilidad del hueco para así llegar a unos resultados de reducción en NPT en aproximadamente un 90% y un incremento de la rata de penetración en un 63% promedio en todos los campos utilizados para dicho estudio

El objetivo general de este proyecto de investigación es Implementar un modelo predictivo para la estimación de un indicador de Borehole Quality mediante la metodología Machine Learning en los pozos perforados durante la campaña 2017-2020 en dos campos de hidrocarburos de la Cuenca del Valle Medio del Magdalena. Como objetivos específicos se definieron:

- Diseñar una matriz de clasificación, a partir de una base de datos, que permita la identificación de las variables críticas que afectan el desempeño óptimo de la perforación.
- Construir un modelo de predictibilidad en el lenguaje de programación Python y calibrarlo con una muestra representativa de las variables críticas estimadas.
- Ejecutar el modelo predictivo sobre la muestra no evaluada en la etapa de calibración.
- Establecer el indicador de Borehole Quality y su impacto financiero mediante un análisis de los indicadores económicos VPN y TIR.

1. MARCO TEÓRICO

Este capítulo hace referencia a los aspectos teóricos necesarios para el entendimiento y desarrollo de este proyecto de investigación. A continuación, se presentarán los conceptos básicos de las actividades de perforación, calidad de pozo y el manejo de Big Data por medio del aprendizaje automático para desarrollar un eficiente indicador de Borehole Quality.

1.1 Perforación

Se conoce como perforación a la actividad de realizar un agujero desde superficie hasta una cierta profundidad, con el fin de explorar y/o explotar un yacimiento de hidrocarburos. [1]

Existen muchas formas de vigilar el proceso de una actividad de perforación. La más usada es por medio del control de los parámetros que hacen parte de ella. Estos son aspectos técnicos fundamentales al momento de entender y saber optimizar la perforación de un pozo en particular.

1.2 Parámetros de perforación

La tabla 1 muestra una breve descripción de los parámetros de perforación mencionados anteriormente.

Tabla 1.

Parámetros de perforación a analizar.

Parámetro	Unidades	Descripción
MSE (Mechanical specific energy)	Psi	Cantidad de energía empleada para remover un volumen de roca
Torque	Klbs-ft	Fuerza proveniente de la sarta de perforación debido a la rotación de esta
WOB (Weight on bit)	Klbs	Peso que experimenta la broca durante la perforación
ROP (Rate of penetration)	Ft/hr	Velocidad a la que se profundiza durante la perforación
RPM (Revoluciones por minuto)	Rpm	Razón a la que la broca gira en la perforación
Tortuosidad	°/100ft	Medida en que el pozo se desvía de su trayectoria originalmente planeada

Nota. Descripción de los parámetros de perforación, analizados en este proyecto, con sus respectivas abreviaciones y unidades.

La ecuación 1 muestra la implicación de los parámetros de perforación en el cálculo de la MSE.

Ecuación 1.

Cálculo de la MSE

$$MSE = \frac{480 (\text{Torque}) (\text{RPM})}{D^2 (\text{ROP})} + \frac{4 (\text{WOB})}{\pi D^2}$$

Donde:

MSE = Energía mecánica específica, psi

Torque = Torque, Klbs-ft

WOB = Peso sobre la broca, Klbs

RPM = Revoluciones por minuto, Rpm

ROP = Tasa de penetración, Ft/hr

D = Diámetro de la broca, in

1.3 Calidad de pozo

Durante años ha crecido la necesidad de unificar la importancia de la correcta optimización de los parámetros de perforación en un solo término que los resuma de una manera técnica y satisfactoria para el estudio, planeación y desarrollo de un pozo. Ese término es la calidad de pozo (Borehole Quality). Desde que se empezó a trabajar con fuerza sobre este término, se considera que un pozo de calidad está asociado con uno direccionalmente perfecto con respecto a la trayectoria planeada, que es liso, de calibre y tiene un mínimo espiralamiento a lo largo de su trayectoria.[2]

En 2005, Mason y Chen, desarrollaron una metodología para la estimación de la calidad de un pozo mediante, lo que ellos llamaron, Wellbore Quality Scorecard (WQS). El sistema funciona marcando individualmente las respuestas de perforación, último viaje de tubería a superficie y la corrida del casing en cierto pozo. Se calcula la puntuación total que está en el rango de 0 a 20. Siendo cero un pozo muy mal perforado y completado y 20 el pozo perfecto en términos de calidad. La WQS se aprecia de la siguiente manera, como muestra la figura 1.

Figura 1.

WQS para un pozo en particular

Drilling Response (5 points maximum)	Score
Severe drilling problems	
– stuck pipe	0
– near stuck pipe incident	1
Transient drilling problems	
– poor hole cleaning with high cuttings bed	2
– severe pack-off	2
– severe loss circulation	2
– erratic torque and drag response	3
Torque and drag response	
– all parameters follow smooth trend	4
– lower than expected torque and drag	5

Nota. WQS para un pozo en su fase de perforación. Tomado de: C. J. Mason and D. C. Chen, "The perfect wellbore!", *SPE Annual Technical Conference and Exhibition*, 2005. [En línea]. Disponible en: <https://n9.c1/d083> [Acceso octubre 20, 2020]

El hecho de entender a cabalidad los aspectos fundamentales de la perforación, sus parámetros e importancia de la calidad de un pozo, es de gran utilidad al momento de desarrollar e implementar un modelo de predictibilidad para la correcta optimización de estos. Para ello, se tratará los conceptos relacionados al aprendizaje automático (machine learning) y diferentes elementos dentro de esta práctica tan utilizada en el mundo del manejo de Big Data.

1.4 Machine learning

“El aprendizaje automático o machine learning es el diseño y estudio de las herramientas informáticas que utilizan la experiencia pasada para tomar decisiones futuras; es el estudio de programas que pueden aprender de los datos. El objetivo fundamental del machine learning es generalizar, o inducir una regla desconocida a partir de ejemplos donde esa regla es aplicada.”[1]

1.4.1 Tipos de machine learning

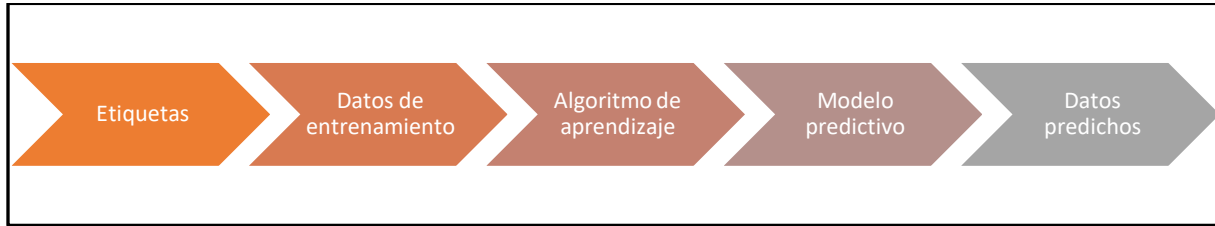
<<El Machine Learning tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica. Pero para poder abordar cada uno de estos temas es crucial en primer lugar distinguir los distintos tipos de problemas de Machine Learning con los que nos podemos encontrar.>> [3]

1.4.1.i Aprendizaje supervisado. En los problemas de aprendizaje supervisado, se entrena al modelo predictivo con datos que ya vienen etiquetados de acuerdo a ciertas categorías. En otras palabras, se le dice al modelo qué representa cada tipo de datos y con base a ello, realiza las predicciones futuras necesarias. [3]

La figura 2 muestra una representación gráfica del procedimiento interno de un aprendizaje supervisado.

Figura 2.

Diagrama de proceso del aprendizaje supervisado



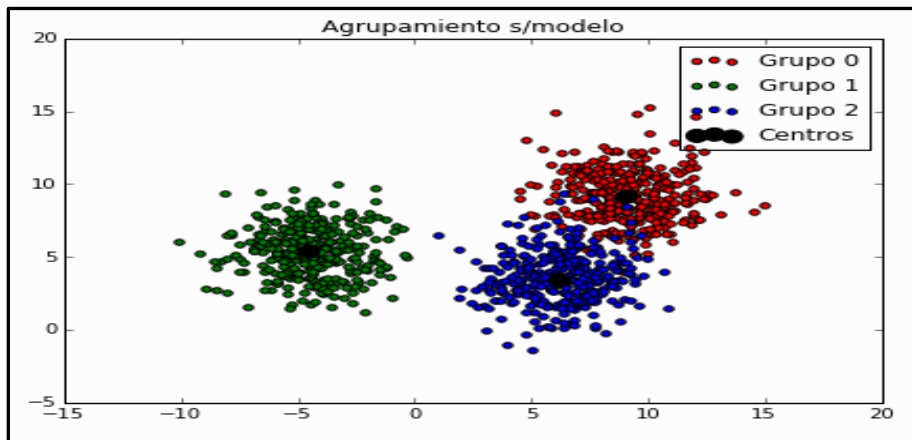
Nota. Proceso interno llevado a cabo por un programa con aprendizaje supervisado

1.4.1.ii Aprendizaje no supervisado. “En los problemas de aprendizaje no supervisado el algoritmo es entrenado usando un conjunto de datos que no tiene ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo pueda encontrar por si solo patrones que ayuden a entender el conjunto de datos.” [3]

A continuación, se muestra un ejemplo de aprendizaje no supervisado:

Figura 3.

Modelamiento de datos con aprendizaje no supervisado



Nota. Clasificación y agrupamiento de datos por medio de aprendizaje no supervisado.

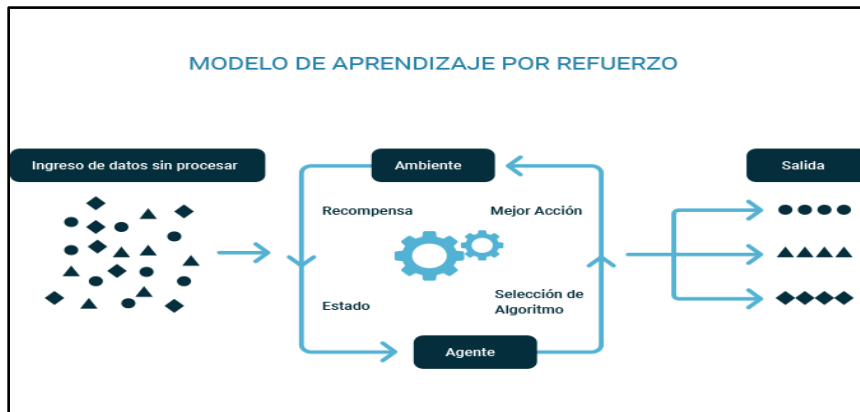
Tomado de: R.E, Lopez-Briega. 2015. Machine Learning. IAAR. [En línea]. Disponible en: <https://iaarbook.github.io/ML/> [Acceso: febrero 17, 2021]

1.4.1.iii Aprendizaje por refuerzo. “En los problemas de aprendizaje por refuerzo, el algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error.” [3]

La figura 4 muestra un ejemplo de aprendizaje por refuerzo aplicado en el lenguaje de programación Python:

Figura 4.

Aprendizaje por refuerzo



Nota. Modelo a seguir de un aprendizaje por refuerzo. Tomado de: Medium.

Tipos de aprendizaje. Aprendizaje por refuerzo. <https://n9.cl/3u8wq>

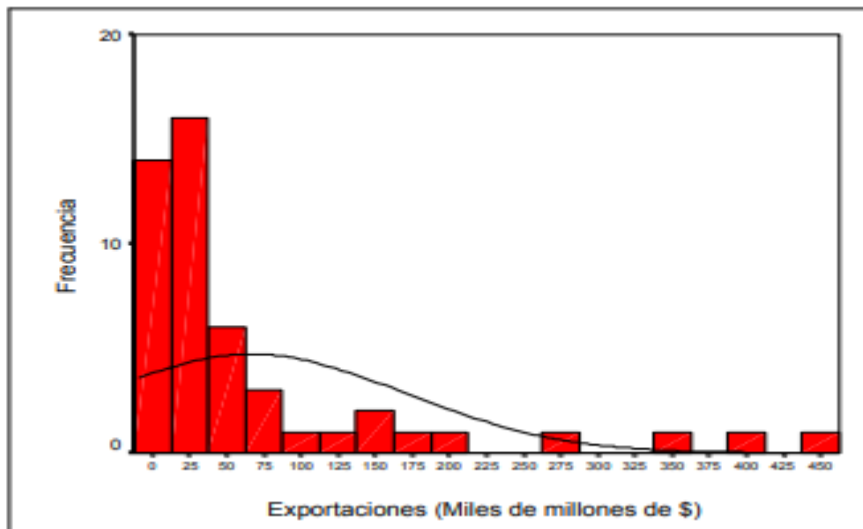
1.4.2 Análisis exploratorio de datos (AED)

“El Análisis Exploratorio de Datos (A.E.D.) es un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.” [3]

Existen diferentes formas de visualizar y analizar datos, dos de ellas son los histogramas y diagramas de caja y bigotes. Estos son gráficos estadísticos que van a representar de una manera simple y entendible los datos que se requieran manejar en el proyecto. A continuación, en las figuras 5 y 6 se podrá visualizar estos diagramas.

Figura 5.

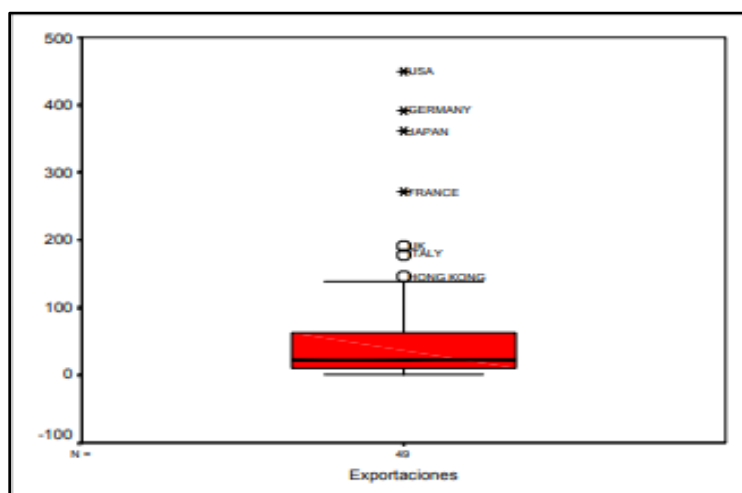
Histograma



Nota. Ejemplo de un histograma con distribución normal sesgada a la izquierda para el valor de las exportaciones de una ciudad. Tomado de: Figueras, S. and M. Gargallo, 2003. "Análisis exploratorio de datos," pp. 18. Disponible en: <https://n9.cl/zgz7>

Figura 6.

Diagrama de caja y bigotes



Nota. Diagrama de caja y bigotes para el ejemplo presentado en la figura 5. Tomado de: Figueras, S. and M. Gargallo, 2003. "Análisis exploratorio de datos," pp. 18. Disponible en: <https://n9.cl/zgz7>

2. METODOLOGÍA Y DATOS

En este capítulo se expone la metodología a emplear para el desarrollo satisfactorio del proyecto.

Esta da inicio con la recopilación de información de los pozos perforados en la campaña 2017-2020 para los campos en estudio. A partir de dicha obtención, se muestra el proceso para la creación de una base de datos consistente que contenga valores de los parámetros de perforación a analizar. Luego es necesario realizar un análisis exploratorio de datos (AED) para estudiar el comportamiento estadístico de las variables de entrada en la base de datos, esto permitirá un mejor entendimiento de la información y su posterior desarrollo. Después se realizará la división del Dataset en datos de entrenamiento y prueba. Llegados a este punto, se implementarán modelos de aprendizaje supervisado automático *Decision Tree Regressor* con proporciones 70/30, 80/20 y 90/10 y verificar cuál es la que arroja mejores resultados de exactitud; para posteriormente validar la exactitud de los modelos y verificar la necesidad de hacer mejoras o no (Tunning). Después, se procede a la validación de los modelos implementados, por medio de graficas TVS vs cada variable predicha. Posteriormente, se calculan numéricamente las variables más importantes al momento de predecir determinada variable. Esto con el fin de generar mapas de calor y observar cuales son los rangos óptimos para los parámetros de perforación. Luego, se hará la consolidación del indicador de calidad de pozo mediante la metodología Wellbore Quality Scorecard (WQS). Con los resultados de los mapas de calor, se procede a crear las líneas base de MSE para su utilización en los dos campos de estudio. Finalmente se realiza un breve análisis financiero para comprobar la eficacia del indicador de calidad WQS.

A continuación, se presenta la metodología explicada paso a paso para dar cumplimiento a los objetivos del proyecto. (Figura 7)

Figura 7.

Diagrama de flujo del proyecto



Nota. Diagrama de flujo para la realización del proyecto “Implementación de un modelo predictivo para la estimación de un indicador de Borehole Quality, mediante la metodología machine learning, en los pozos perforados durante la campaña 2017-2020 de dos campos de hidrocarburos de la cuenca del Valle Medio del Magdalena”.

2.1 Obtención y selección de parámetros

La compañía Ecopetrol S. A. fue la encargada de suministrar la información requerida para la realización del proyecto. Dicha información consta de informes finales de perforación, mud Logging y surveys. De los reportes finales de perforación se extraen los estados mecánicos de los pozos con el fin de identificar la base y tope de cada sección a estudiar. Los informes de mud Logging contienen datos correspondientes a profundidad medida (MD), revoluciones por minuto en superficie (RPM_SUP), revoluciones por minuto en fondo (RPM_FON), Torque y peso sobre la broca (WOB); cabe resaltar que la data de mud logging está registrada pie a pie en. Por último, los surveys proporcionan inclinación (INCL), dog leg severity (DLS), coordenadas (norte, este) y profundidad vertical verdadera (TVD). Toda la data proporcionada por los surveys está dada cada 100 pies, por lo que más adelante será necesaria una conversión para estandarizar todos los valores con referencia pie a pie.

2.1.1 Cálculo de variables implícitas

Las variables implícitas son aquellas que no son dadas de forma directa en los registros e informes suministrados, por esta razón es necesario determinarlas a partir de la información inicial.

2.1.1.i Along Hole Displacement (AHD). La variable AHD hace referencia al desplazamiento horizontal que tiene el pozo proyectado sobre un plano de 180° medido en pies[4]. A continuación, se presenta la ecuación pertinente para su cálculo.

Ecuación 2.

Along Hole Displacement (AHD)

$$AHD = \sqrt{North^2 + East^2}$$

Donde: AHD = Along hole displacement, ft

North = Coordenada al norte, ft

East = Coordenada al este, ft

2.1.1.ii Tortuosidad (T). Esta variable es un indicador de la desviación de un pozo con respecto a la vertical. Se mide en grados cada 100 pies ($^{\circ}/100 \text{ ft}$)[4]. La ecuación para calcularla es la siguiente:

Ecuación 3.

Tortuosidad

$$T = \sum_{i=1}^n DLS$$

Dónde: T = Tortuosidad, $^{\circ}/100 \text{ ft}$

DLS = Dog Leg Severity, $^{\circ}/100 \text{ ft}$

2.1.1.iii Índice de dificultad direccional (DDI). Es una variable que mide la dificultad relativa que puede ser encontrada al momento de perforar un pozo[4]. Es un parámetro adimensional que es calculado de la siguiente forma:

Ecuación 4.

Índice de dificultad direccional

$$DDI = \text{Log}_{10} \left(\frac{MD}{TVD} * AHD * T \right)$$

Donde: DDI = Índice de dificultad direccional

MD = Profundidad medida, ft

TVD = Profundidad vertical verdadera, ft

AHD = Desplazamiento a lo largo del pozo, ft

T = Tortuosidad, $^{\circ}/100 \text{ ft}$

El índice de dificultad direccional (DDI) tiene una connotación física importante, ya que es el parámetro que caracteriza a un pozo de una forma breve y concisa de acuerdo a los inconvenientes que existen a lo largo de la perforación de este. Por esta razón, el DDI tiene un grupo de rangos con su respectivo significado.

Figura 8.

Significado operativo de DDI

Dificultad	Tipo de pozo
DDI < 6	Pozos relativamente cortos / perfiles simples con baja tortuosidad
6.0 < DDI < 6.4	Pozos cortos con alta tortuosidad / pozos más largos con baja tortuosidad
6.4 < DDI < 6.8	Pozos más largos con trayectorias relativamente tortuosas
DDI > 6.8	Pozos largos con perfiles tortuosos con un alto grado de dificultad

Nota. Rangos de significado físico para DDI. Tomado de: ALISTAR.W, “*The Directional Difficulty Index – A New Approach to Performance Benchmarking*”, SPE, 59196, 2019, [En línea] <https://n9.cl/05efn> [Acceso: marzo 2, 2021]

2.1.1 Importación de librerías en Python y cálculo de la TVD

Una librería es un conjunto de implementaciones funcionales, codificadas en un lenguaje de programación, que ofrece una interfaz bien definida para la funcionalidad que se invoca. Dicho en otras palabras, son bases de datos que almacenan códigos con los cuales se puede crear, ejecutar y entrenar un modelo. En la tabla 2 se presentan las librerías importadas y utilizadas para el desarrollo de este proyecto.

Tabla 2.*Librerías de Python necesarias para el proyecto*

Librería	Descripción
Matplotlib	Matplotlib es una biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación Python
Seaborn	Seaborn es una librería de visualización de datos para Python desarrollada sobre matplotlib. Ofrece una interfaz de alto nivel para la creación de atractivas gráficas.
Pandas	Pandas es una biblioteca de software escrita como extensión de NumPy para manipulación y análisis de datos para el lenguaje de programación Python. En particular, ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.
ScikitLearn	Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. Además, presenta la compatibilidad con otras librerías de Python como Numpy y matplotlib.
Numpy	Numpy es una biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.

Nota. Mención y descripción de las librerías a utilizar para la realización de este proyecto.

Como se mencionó anteriormente, es necesario calcular algunas variables que no están explícitas en la información entregada por la compañía Ecopetrol S. A. Una de ellas es la profundidad vertical verdadera (TVD). Es de vital importancia el cálculo de esta variable porque va a ser el parámetro base sobre el cual las demás variables estarán referenciadas. La ecuación que calcula la TVD es la siguiente:

Ecuación 5.

Cálculo para TVD

$$TVD_i = (MD_i - MD_{i-1}) * \cos(INCL_i) + TVD_{i-1}$$

Dónde: TVD = Profundidad vertical verdadera, ft

MD = Profundidad medida, ft

INCL = Inclinación, °

Cabe recordar que la información suministrada por el survey está dada cada 100 ft de, con lo que al momento de realizar el cálculo de la **Ecuación 5**, existen algunos valores de MD para los cuales no hay un dato de inclinación. Por esta razón, surge la necesidad de interpolar el valor de inclinación para que así esté registrado pie a pie. La figura 9 muestra el código utilizado para generar la interpolación necesaria.

Figura 9.

Código encargado de interpolar la inclinación

```
1 import pandas as pd
2 df=pd.read_excel('C:/Users/juana/Desktop/DATACASI/Y187.xlsx', sheet_name='Hoja1')
3 print(df)
4 df=df.interpolate(method='linear',axis=0).bfill()
5 print(df)
6 df.to_excel('C:/Users/juana/Desktop/DATACASI/DDIY187.xlsx')
```

Nota. La línea 2 contiene el código encargado de importar el archivo de Excel que alberga el dataset para el pozo Y187. La línea contiene la instrucción para interpolar la inclinación en las celdas donde no existe ningún valor. Por último, la línea 6 exporta el DataFrame a un archivo de Excel. Se resalta que en este caso es necesaria la importación de la librería “pandas” porque se crean e importan objetos DataFrames.

Hecho el cálculo de la TDV para todos los pozos en estudio, se procede a crear un dataset para cada pozo y analizar con mayor facilidad si existen valores nulos a lo largo de las columnas.

2.1.3 Caracterización de las variables

La caracterización de variables es un paso indispensable en el proceso para crear un dataset consistente. Indica el tipo de dato de cada una de las variables presentes. Para los parámetros: TORQUE, ROP, RPM_SUP, RPM_FON, WOB, SECTION y DDI se busca un tipo de dato float ya que son variables que contienen decimales, para MD se asigna dato tipo int y para la variable Well_Name un tipo de dato categórico nominal (Object). Así pues, se muestra en la figura 9 un ejemplo de un dataset para un pozo en particular.

Figura 10.

Caracterización de las variables en un pozo

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7680 entries, 0 to 7679
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  -
0   SECTION    7680 non-null   float64
1   WELL_NAME  7680 non-null   object
2   MD          7680 non-null   int64
3   TVD         7680 non-null   float64
4   RPM_SUP    7680 non-null   float64
5   RPM_FON    7680 non-null   float64
6   TORQUE     7680 non-null   float64
7   WOB        7680 non-null   float64
8   ROP        7680 non-null   float64
9   DDI        7680 non-null   float64
dtypes: float64(8), int64(1), object(1)
memory usage: 600.1+ KB
```

Nota. Caracterización de las variables en el pozo ejemplo, se puede observar 8 variables tipo float64, una variable tipo int64 y una tipo Object. Para un total de 7680 entradas del DataFrame correspondientes a la longitud total medida en MD. Las referencias non-null muestran la cantidad de valores no nulos en el dataset.

2.2 Análisis exploratorio de datos (AED)

Una vez consolidada la base de datos a implementar en el modelo de predicción, se procede a realizar una descripción estadística de las variables en cuestión, con el fin de describir su comportamiento, eliminar valores atípicos y establecer correlaciones entre las mismas. Para esto se empleará la metodología mostrada en la figura 11.

Figura 11.

Metodología del análisis exploratorio de datos



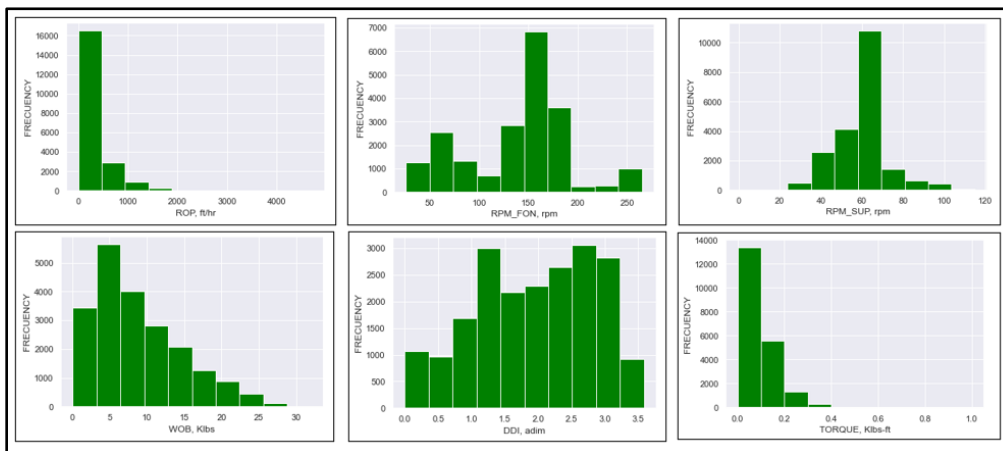
Nota. Descripción de los pasos a seguir para realizar el estudio estadístico de las variables contenidas en el dataset a implementar en el modelo predictivo

Para iniciar, se generarán los histogramas para cada una de las variables. Estos muestran la distribución de frecuencias que tienen los valores de cada variable. Por esta razón, son gráficos muy útiles para el entendimiento estadístico de los parámetros de estudio.

A continuación, se presentan las figuras 12 y 13, las cuales contienen los histogramas generados para los pozos a analizar de la secciones 12 ¼” y 8 ½” respectivamente

Figura 12.

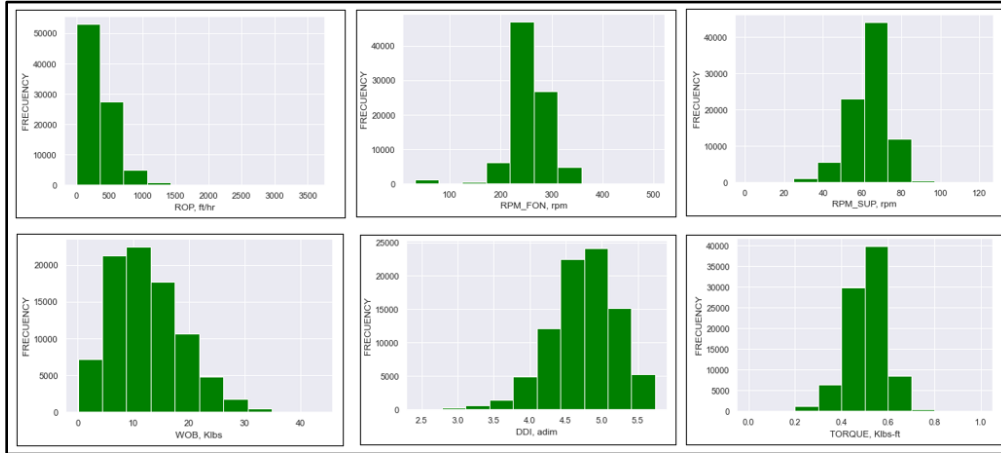
Histogramas de parámetros de perforación (Sección 12 ¼”)



Nota. Generación de histogramas para las variables en la sección 12 ¼”.

Figura 13.

Histogramas de parámetros de perforación (Sección 8 ½”)



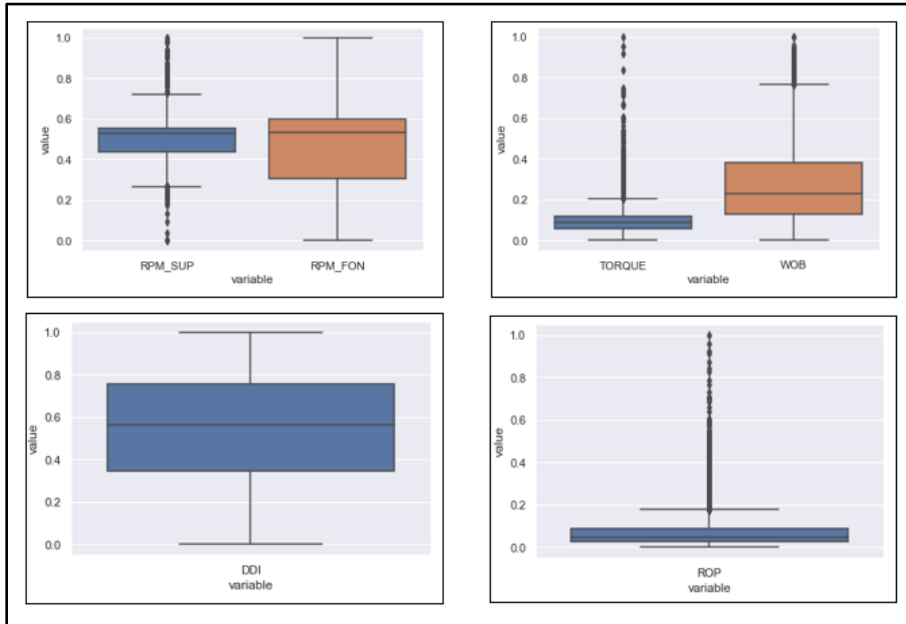
Nota. Generación de histogramas para las variables en la sección 8 ½”

Los histogramas muestran información que valida los datos iniciales. Por ejemplo, la variable ROP evidencia un sesgo a la izquierda; esto quiere decir que la mayoría de sus valores a lo largo de los pozos son bajos y se debe a que la sección de mayor longitud (8 ½”) es difícil de perforar con ROP’s altas debido a su complejidad estructural.

Así como los histogramas fueron generados, los diagramas de caja lo serán, con una diferencia. La variables que sean de la misma especie (iguales unidades de medición) serán graficadas en un solo plano cartesiano; de la misma manera, las variables que tengan diferente especie estarán individualmente graficadas. Para esto fue necesario hacer uso de las librerías Pandas, NumPy, Seaborn y Matplotlib dado que se trabaja con DataFrames, variables numéricas y visualización de datos en plots.

Figura 14.

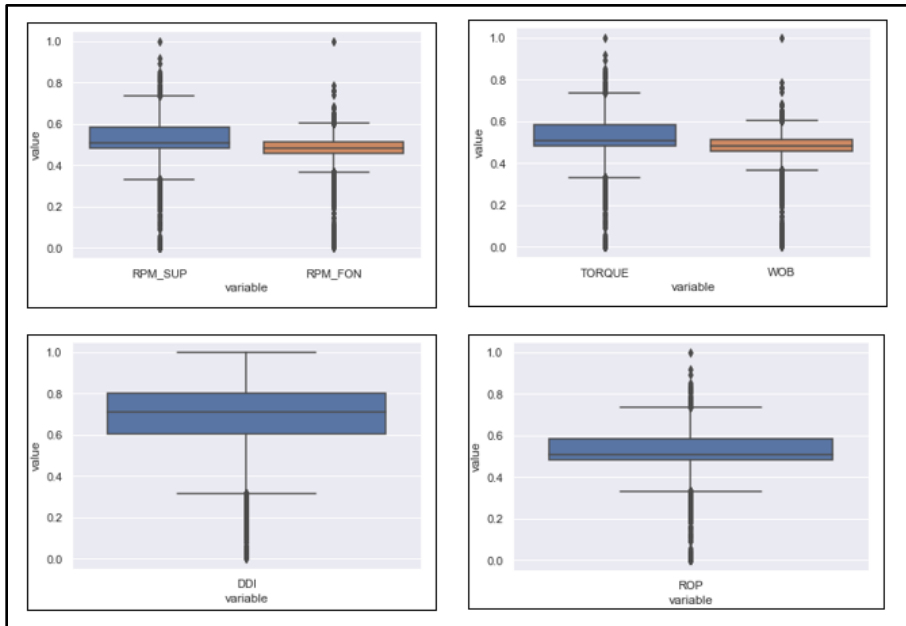
Diagrama de caja para los parámetros de perforación (Sección 12 ¼”)



Nota. Visualización de los parámetros de perforación por medio de diagramas de caja separados por especie de variable

Figura 15.

Diagrama de caja para los parámetros de perforación (Sección 8 ½”)

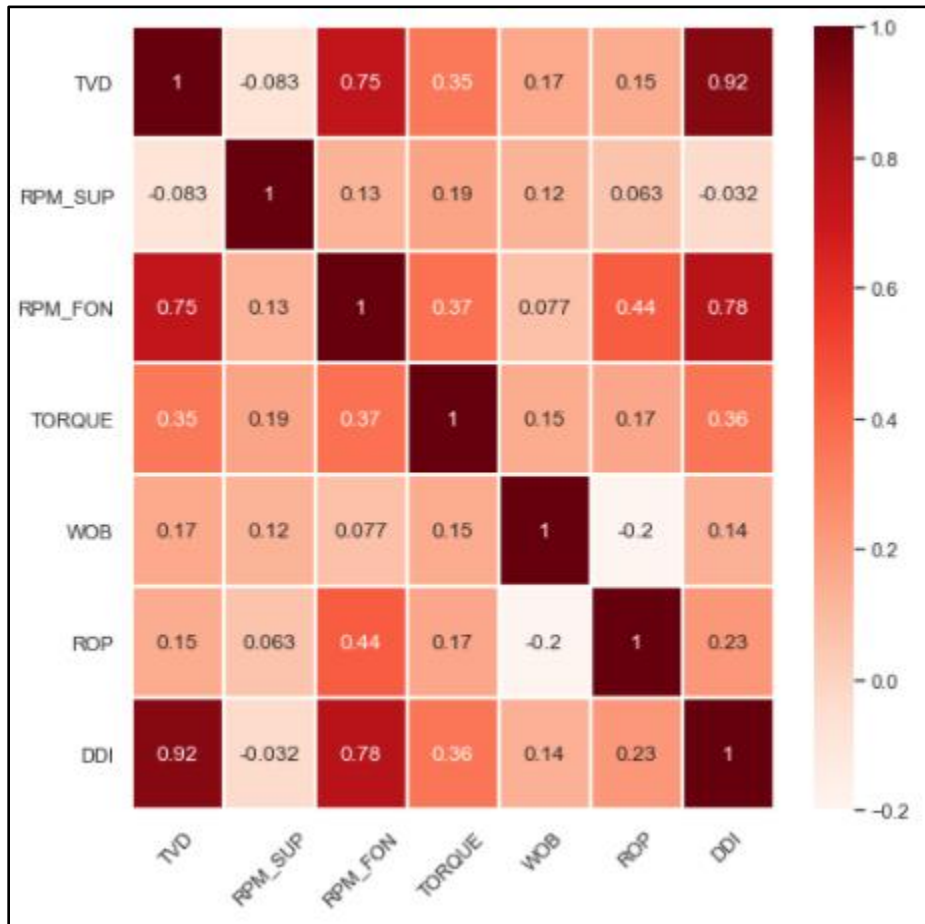


Nota. Visualización de los parámetros de perforación por medio de diagramas de caja separados por especie de variable

Seguido a esto, se procede a la generación de una matriz de correlación entre las variables. Este gráfico ayuda al entendimiento del grado de implicación que tiene un parámetro sobre otro y así poder observar cuáles de ellos son los que más afectan a la información de entrada y por lo tanto al modelo predictivo. Las figuras 16 y 17 muestran lo mencionado anteriormente.

Figura 16.

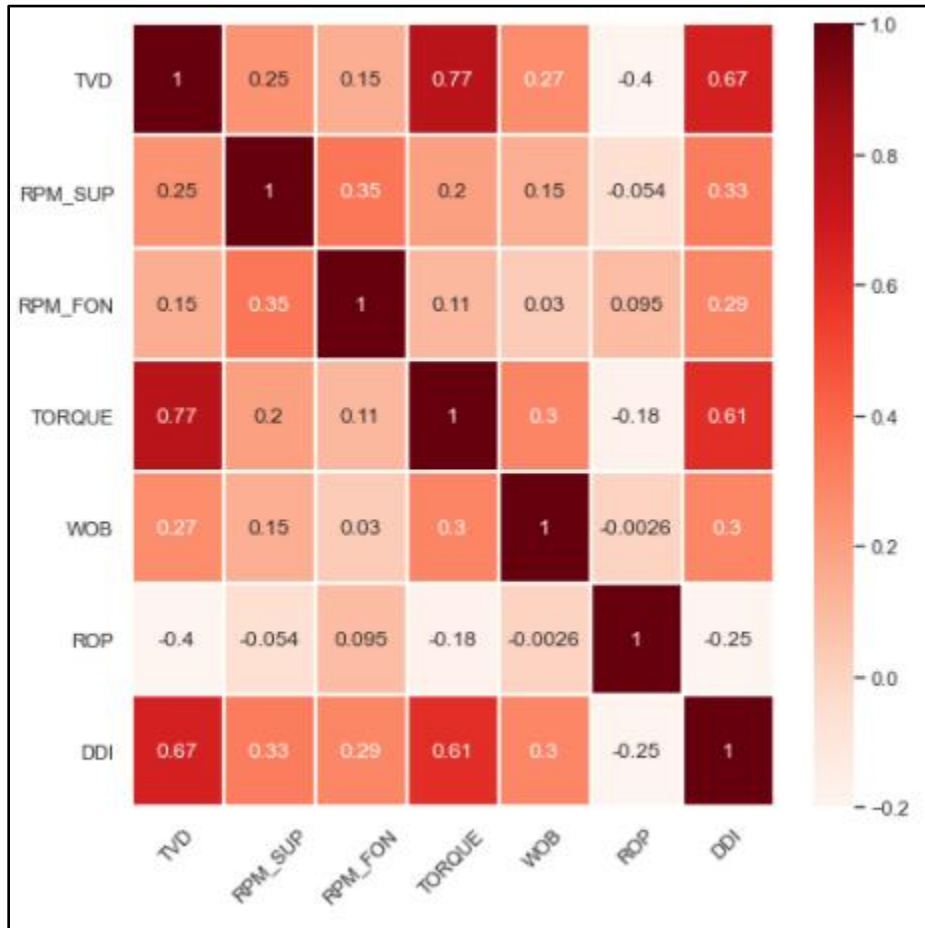
Matriz de correlación entre variables (Sección 12 ¼”)



Nota. Los valores negativos de cada casilla de la matriz indican una proporcionalidad inversa entre las variables correspondientes a ella, de la misma forma sucede con los valores positivos.

Figura 17.

Matriz de correlación entre variables (Sección 8 ½”)



Nota. Los valores negativos de cada casilla de la matriz indican una proporcionalidad inversa entre las variables correspondientes a ella, de la misma forma sucede con los valores positivos.

Si se tiene en cuenta la **Ecuación 1.**, se puede observar que las variables TORQUE, RPM y WOB son directamente proporcionales y la ROP es inversamente proporcional a ellas. Esto se evidencia en las figuras 16 y 17, la casillas que corresponden a ROP y las demás variables, contienen valores negativos; en cambio las casillas que corresponden entre TORQUE, RPM y WOB solo hay valores positivos.

Este procedimiento de estudiar estadísticamente los parámetros de perforación es realizado, entre otras cosas, para generar los rangos operativos de estos y así poder compararlos con las rangos

óptimos obtenidos del modelo predictivo a implementar. Las tablas 3 y 4 muestran los rangos operativos iniciales mencionados anteriormente.

Tabla 3.

Rangos operativos de las variables de entrada (Sección 12 ¼”)

Parámetro	Rango
RPM_SUP, rpm	1-115
RPM_FON, rpm	26-266
TORQUE, Klbs-ft	0-15
WOB, Klbs	0-32
ROP, ft/hr	10-900
DDI, adim	1-3.6

Nota. Rangos operativos para los parámetros de perforación al momento de alimentar el modelo predictivo.

Tabla 4.

Rangos operativos de las variables de entrada (Sección 8 ½”)

Parámetro	Rango
RPM_SUP, rpm	1-120
RPM_FON, rpm	30-500
TORQUE, Klbs-ft	0-32
WOB, Klbs	0-44
ROP, ft/hr	10-800
DDI, adim	2-5.8

Nota. Rangos operativos para los parámetros de perforación al momento de alimentar el modelo predictivo.

2.3 División del dataset en datos de entrenamiento y prueba

La división del dataset es un procedimiento que consiste en realizar una partición de la información contenida en él. Un porcentaje es utilizado para entrenamiento y el restante para prueba del modelo. Es importante resaltar que, debido a la cantidad de variables a predecir, es necesario realizar una partición diferente para cada sección y parámetro. Es decir, se obtendrán diferentes proporciones para la división del dataset según la variable que en ese momento se vaya a predecir.

Las proporciones a usar para la implementación y validación del modelo predictivo de cada variable dependen de la exactitud que arroje cada proporción. Ahora bien, en dado caso de que las exactitudes de todas las proporciones posibles (60/40, 70/30, 80/20 y 90/10), para la división de un dataset en el momento de predecir cierta variable, sean demasiado bajas (<60%) se procede a realizar un ajuste de hiper parámetros del modelo predictivo. Esto será tratado con detalle más adelante.

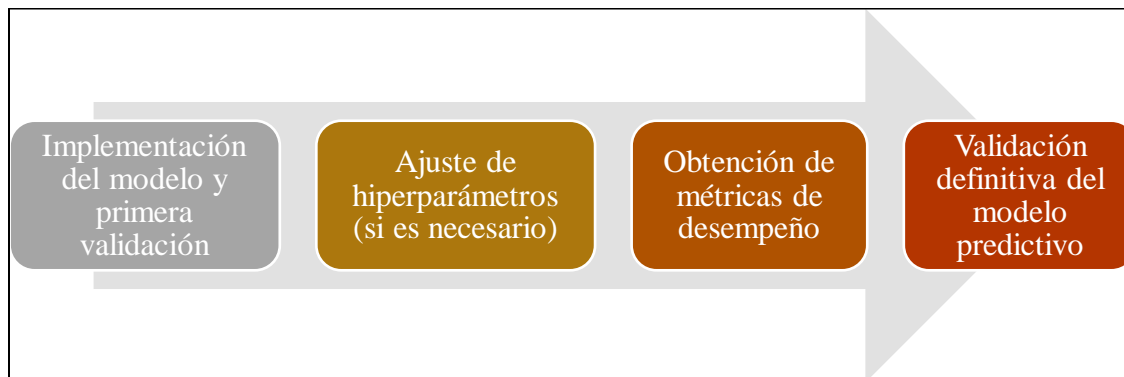
La librería Scikit-learn proporciona la función *Train_Test_Split*, la cual genera de forma automática la división del dataset correspondiente.

2.4 Implementación y validación del modelo predictivo

Teniendo en cuenta la división de los dataset para cada una de las variables a predecir, se procede a crear, implementar y validar el modelo predictivo de aprendizaje automático *Random Forest*. Para ello, se presenta el procedimiento a seguir en la **Figura 17**.

Figura 18.

Metodología a seguir para el modelo Random Forest



Nota. Pasos a seguir para predecir cada uno de los parámetros de perforación al momento de implementar cada modelo predictivo se procede a verificar su exactitud, esto con el fin de comprobar si es necesario realizar un ajuste de hiper parámetros o no. Este criterio será basado en el hecho de que una exactitud mayor o igual al 70% se considera aceptable. Para ello, se genera una tabla que resume dichos resultados para cada variable a predecir con las respectivas proporciones de división para los dataset.

Tabla 5.*Primera exactitud de cada modelo predictivo (Sección 12 ¼”)*

Parámetros	70/30	80/20	90/10
ROP	79.03	79.72	81.59
RPM_FON	95.41	95.13	95.39
TORQUE	71.64	72.63	69.27
WOB	70.56	72.18	70.82
DDI	92.82	92.86	92.91

Nota. Se presentan los porcentajes de exactitud de los modelos predictores para cada parámetro. Las celdas resaltadas en rojo son las proporciones con mayor exactitud para su correspondiente variable, por lo tanto, son las que serán utilizadas para las predicciones.

Tabla 6.*Primera exactitud de cada modelo predictivo (Sección 8 ½”)*

Parámetros	70/30	80/20	90/10
ROP	66.93	68.37	70.34
RPM_FON	82.88	84.55	84.10
TORQUE	92.96	93.80	93.99
WOB	68.10	69.77	71.55
DDI	84.14	84.61	84.30

Nota. Se presentan los porcentajes de exactitud de los modelos predictores para cada parámetro. Las celdas resaltadas en rojo son las proporciones con mayor exactitud para su correspondiente variable, por lo tanto, son las que serán utilizadas para las predicciones.

2.4.1 Ajuste de hiper parámetros de los modelos

Con el fin de obtener modelos de predicción optimizados y lo más posiblemente exactos, se realizará un ajuste de hiper parámetros que consiste en la combinatoria de estos junto con un

proceso iterativo para obtener sus valores óptimos y de esa manera mejorar la precisión de los modelos que lo requieran.

Seguido a esto, se generarán matrices de clasificación que representan la importancia de las variables predictoras sobre la variable objetivo en cada modelo de predicción. Se seleccionarán las 3 que mayor influencia tengan.

2.4.2 Validación de los modelos predictivos

Por medio de gráficas que representen los datos reales (base de datos) contra los generados en los modelos para cada variable, se realizará la validación de los modelos de predicción. Una buena precisión de cada modelo será evidenciada en la poca variabilidad que tendrán los dos conjuntos de datos previamente mencionados.

2.4.3 Generación de mapas de calor

Los mapas de calor son una representación gráfica en colores, de un par de variables con la combinación de sus valores utilizados para predecir una variable de salida []. Este tipo de gráficos son generados a partir de un dataset creado con los rangos de entrada de las variables predictoras de un *target*. Dichos rangos se pueden visualizar en las tablas 3 y 4 para cada sección de estudio. Cada variable predicha será generada mediante 3 mapas de calor que combinan las 3 variables que más incidencia tengan en ella; esto se hará mediante los resultados de la matriz de clasificación.

2.4.4 Consolidación de rangos operativos

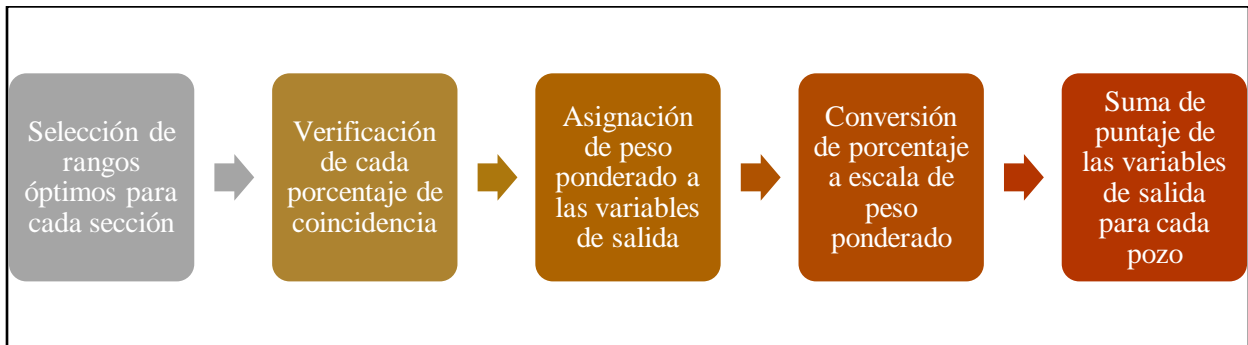
Los 3 mapas de calor generados para cada variable a predecir se corresponden en una única escala que mide el rango de resultados. Este corresponde al rango operativo óptimo para cada parámetro de salida. Dichos resultados serán utilizados para establecer el indicador de calidad (WQS)

2.5 Construcción del indicador de calidad (WQS)

Una vez se generen los rangos óptimos para las variables deseadas, se procede a generar el indicador de calidad WQS. Este procedimiento se llevará a cabo como se muestra en la figura 19.

Figura 19.

Metodología para la creación de WQS



Nota. La figura muestra la metodología a ejecutar en el momento de generar el WQS.

Los rangos óptimos se obtienen de las escalas unificadas para cada grupo de mapas de calor. Posteriormente se procede a verificar qué porcentaje de coincidencia tienen los rangos óptimos en cada pozo de determinada sección. Esto se logra dividiendo el número de datos coincidentes entre el número de datos totales para el pozo en una sección.

La asignación de un peso ponderado a cada variable se hace teniendo en cuenta la influencia general que tienen con respecto a la perforación de un pozo. La sumatoria de los pesos ponderados debe ser igual a 20 porque así es la escala manejada en el paper *The Perfect Wellbore*, el cual se utilizó para el desarrollo de esta metodología.

Figura 20.

Clasificación de pozos

RANGOS DE CLASIFICACIÓN	DESCRIPCIÓN
$0 < WQS \leq 2$	Pega de tubería
$2 < WQS \leq 6$	Pozo de calidad baja
$6 < WQS \leq 10$	Pozo de calidad media
$10 < WQS \leq 14$	Pozo de calidad alta
$14 < WQS < 20$	Pozo de excelente calidad
$WQS = 20$	Pozo perfecto

Nota. Rangos utilizados para la clasificación de pozos según su WQS. Tomado de: C. J. Mason and D. C. Chen, "The perfect wellbore!", *SPE Annual Technical Conference and Exhibition*, 2005. [En línea]. Disponible en: <https://n9.c1/d083> [Acceso octubre 20, 2020]

Teniendo como resultado los pesos ponderados para las variables, se procede a realizar una conversión del porcentaje de coincidencia a la escala de estos. Este procedimiento se puede hacer mediante una simple regla de tres que relacione los dos elementos a comparar. Posteriormente se ejecuta la sumatoria de estos resultados obteniendo una calificación del pozo en el rango presentado en la figura 20.

2.6 Análisis financiero del indicador de calidad (WQS)

Con el indicador de calidad (WQS) ya consolidado, es necesario realizar un breve estudio financiero para comprobar que la aplicabilidad de los rangos óptimos generados por el modelo predictivo resulta en una baja de costos para la compañía. La comparación se realizará entre las VPN y TIR en ambos escenarios (datos reales y aplicando rangos óptimos).

3. ANÁLISIS DE RESULTADOS

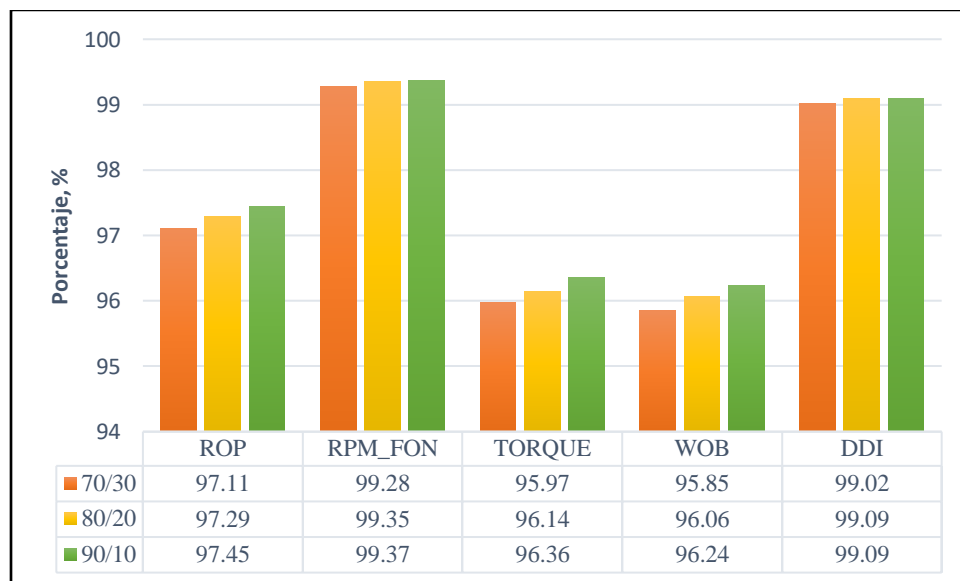
En este capítulo se indican los resultados de entrenamiento e implementación de los modelos predictivos junto con los respectivos rangos óptimos generados. También se presentarán los consolidados del indicador de calidad de pozo (WQS), sumado al estudio financiero del proyecto.

3.1 Entrenamiento de los modelos predictores

El proceso de entrenamiento de cada modelo predictor consiste en la división de su respectivo dataset en diferentes proporciones (70/30, 80/20 y 90/10) con el fin de que más adelante sea capaz de predecir valores aceptables para la variable deseada. A continuación, se muestran los resultados del entrenamiento.

Figura 21.

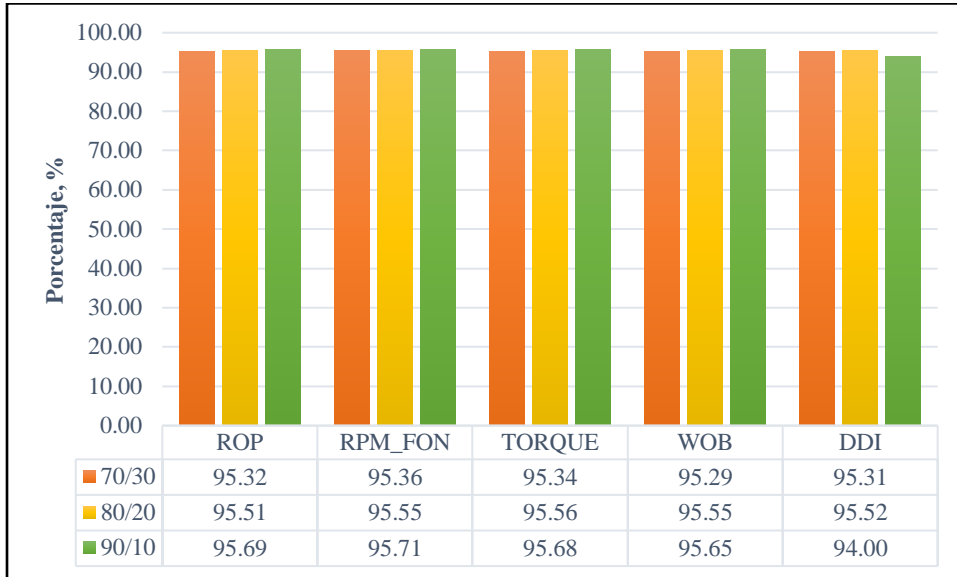
Entrenamiento de los modelos predictivos sección (12 ¼")



Nota. La gráfica representa los porcentajes de exactitud en el entrenamiento para los modelos predictores de cada variable

Figura 22.

Entrenamiento de los modelos predictivos sección (8 ½”)



Nota. La gráfica representa los porcentajes de exactitud en el entrenamiento para los modelos predictores de cada variable

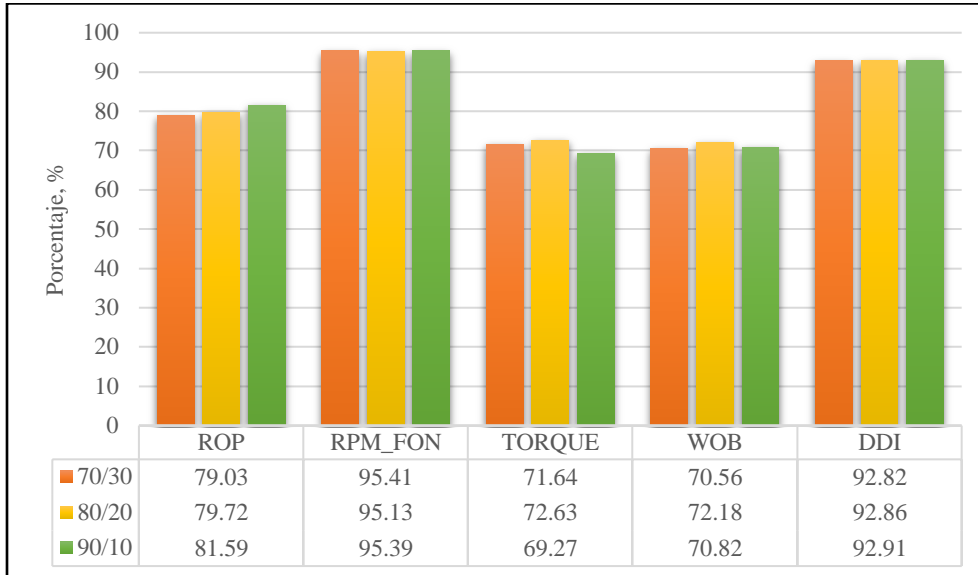
Es evidente que los porcentajes de acierto son excelentes, esto se debe a que se entrena el modelo con la misma información de entrada y lo que hace es generar un patrón en ella. Dichos resultados muestran que el modelo aprendió bien y está listo para generar buenas predicciones de una variable en específico.

3.2 Implementación y ajuste de modelos predictivos

Para iniciar la implementación de modelos predictivos es necesario tener la información de cuáles proporciones son las más adecuadas para cada variable objetivo. Para ello, las figuras 23 y 24 muestra los porcentajes de exactitud de los modelos con las diferentes proporciones para los respectivos dataset en cada sección de estudio.

Figura 23.

Precisión de los modelos predictivos (Sección 12 ¼”)



Nota. Porcentajes de exactitud de los modelos predictivos para predecir cada variable

Figura 24.

Precisión de los modelos predictivos (Sección 8 ½”)



Nota. Porcentajes de exactitud de los modelos predictivos para predecir cada variable

Los modelos para predecir RPM_FON y DDI son los que más exactitud arrojan, con porcentajes que rondan entre 95% y 92% en ambas secciones, respectivamente. El modelo para la variable TORQUE arroja una exactitud de 94% con una proporción 90/10 en la sección 8 ½”, esto se debe a que en dicha sección no suele haber muchos problemas de pega de tubería ni atascamientos. La tabla 7 muestra la selección de proporciones de una manera más visible y limpia.

Tabla 7.

Proporciones utilizadas para los modelos predictivos

Parámetro	Proporción	
	12 1/4"	8 1/2"
ROP	90/10	90/10
RPM_FON	70/30	80/20
TORQUE	80/20	90/10
WOB	80/20	90/10
DDI	90/10	80/20

Nota. Selección de proporciones de división para los modelos predictivos

Como se puede observar, la relación 90/10 es la más adecuada para la mayoría de los modelos. Esto se debe a que la información de entrada es confiable y por lo tanto es útil entrenarlos con el mayor porcentaje posible de ella.

3.2.1 Ajuste de hiper parámetros (Tunning)

Aunque todas las precisiones son aceptables al estar por encima del 70%, se hizo un ajuste de hiper parámetros para el modelo que predice TORQUE en la sección 12 ¼”. El procedimiento consiste en ejecutar el siguiente código.

Figura 25.

Código para ajustar hiper parámetros

```
1 #Tunning the RandomForest
2 from sklearn.model_selection import RandomizedSearchCV
3 n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
4 max_features = ['auto', 'sqrt']
5 max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
6 max_depth.append(None)
7 min_samples_split = [2, 5, 10]
8 min_samples_leaf = [1, 2, 4]
9 bootstrap = [True, False]
10 from pprint import pprint
11 random_grid = {'n_estimators': n_estimators,
12               'max_features': max_features,
13               'max_depth': max_depth,
14               'min_samples_split': min_samples_split,
15               'min_samples_leaf': min_samples_leaf,
16               'bootstrap': bootstrap}
17 pprint(random_grid)
18 rfr = RandomForestRegressor()
19 rfr_random = RandomizedSearchCV(estimator = rfr, param_distributions = random_grid, n_iter = 50,
20                               cv = 3, verbose=2, random_state=42, n_jobs = -1)
21 X=data.drop(["TORQUE"], axis=1)
22 y=data.TORQUE
23 rfr_random.fit(X_train, y_train)
24
25 rfr_random.best_params_
```

Nota. Instrucciones para realizar una optimización de hiper parámetros para la variable TORQUE en la sección 12 ¼”.

Las líneas 3 – 9 muestran condicionales que iteran los valores de hiper parámetros ingresados al modelo. La línea 25 indica al programa mostrar cuál es el mejor conjunto de hiper parámetros para que el modelo sea lo más exacto posible. En el momento que el programa arroje dicho conjunto, se procede a reemplazar estos valores en el código que crea el *RandomForest* y correr el modelo para obtener su exactitud.

El resultado de la implementación de este modelo en particular después de realizar el ajuste es de 74.04%. Si se compara con el porcentaje inicial (72.63%), es una mejora considerable.

La tabla 8 muestra los resultados de las métricas de desempeño para los modelos predictivos en cada sección de estudio.

Figura 26.

Métricas de desempeño para los modelos predictivos

Parámetro	Precisión, %	
	12 1/4"	8 1/2"
ROP	81.59	70.35
RPM_FON	95.41	84.55
TORQUE	74.04	94.01
WOB	72.18	71.56
DDI	92.91	84.62

Nota. Precisión en la predicción de las variables con los mejores modelos para su implementación

3.3 Importancia de las variables

Teniendo como información de entrada todos los valores de las variables generados por los modelos predictivos, se procede a generar una matriz de clasificación (objetivo 1) que da a conocer las 3 variables más importantes para cada predicción de otra variable (target). Dicha matriz es construida con el código “*rf.feature_importances_*”, el cual proporciona valores numéricos a cada variable predictora y así conocer las que más incidencia tengan al momento de la predicción de cierto parámetro.

En la sección 12 ¼” las variables más importantes para la predicción de ROP son: RPM_FON, TORQUE y DDI, en la predicción de RPM_FON son: ROP, RPM_SUP y DDI, en la predicción de TORQUE son: WOB, DDI y RPM_FON, en la predicción de WOB son: RPM_FON, ROP y DDI, por último, para la predicción de DDI son: RPM_FON, TORQUE y RPM_SUP. Es evidente que la RPM_FON es un parámetro que siempre está presente de una manera considerable para las predicciones de todos los modelos.

En la sección 8 ½” las variables más importantes para la predicción de ROP son: RPM_FON, TORQUE y DDI, en la predicción de RPM_FON son: ROP, RPM_SUP y DDI, en la predicción de TORQUE son: WOB, DDI y RPM_FON, en la predicción de WOB son: RPM_FON, ROP y DDI, por último, para la predicción de DDI son: RPM_FON, TORQUE y RPM_SUP. Cabe

resaltar que la variable RPM_SUP se utilizó como predictor, pero no es un target debido a su baja incidencia en las operaciones de perforación. A continuación, las tablas x resume de una manera más clara lo expuesto anteriormente.

Figura 27.

Matriz de clasificación para la sección 12 ¼”

ROP ->	RPM_FON TORQUE DDI	RPM_FON ->	ROP RPM_SUP DDI	TORQUE ->	WOB DDI RPM_FON
	WOB ->	RPM_FON ROP DDI	DDI ->	RPM_FON TORQUE RPM_SUP	

Nota. Variables con mayor importancia para cada variable objetivo

Figura 28.

Matriz de clasificación para la sección 8 ½”

ROP ->	RPM_FON TORQUE DDI	RPM_FON ->	ROP RPM_SUP DDI	TORQUE ->	WOB DDI RPM_FON
	WOB ->	RPM_FON ROP DDI	DDI ->	RPM_FON TORQUE RPM_SUP	

Nota. Variables con mayor importancia para cada variable objetivo

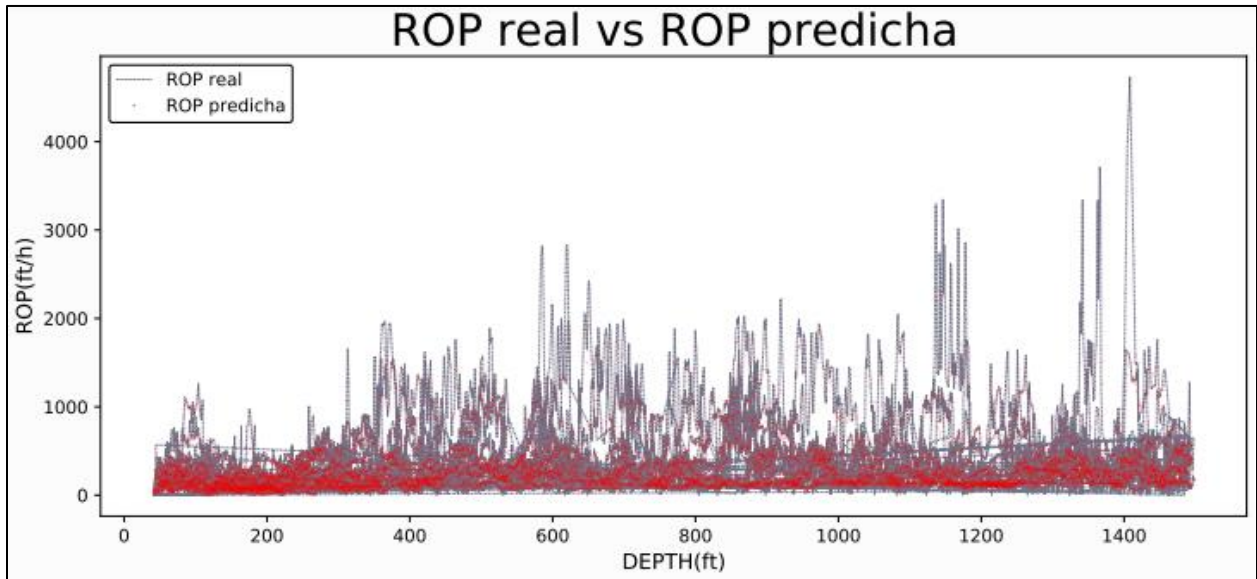
Haciendo un comparativo entre la figura 27 y la figura 14 se puede apreciar la consistencia en el desarrollo de la importancia de las variables. En la matriz de correlación es evidente el alto valor que poseen las variables RPM_FON, ROP y DDI relacionadas entre sí; esto se ve reflejado en que son las variables más recurrentes al momento de construir la matriz de clasificación para las dos secciones.

3.4 Validación de los modelos predictivos

La validación de los modelos predictivos se realiza a través de gráficos que comparan la variable objetivo generada por el modelo contra la misma con sus valores reales tomados en campo. Todo con referencia a la TVD. La figuras 29 y 30 muestran ejemplos de validación para el modelo predictivo de ROP para las dos secciones. Las demás graficas serán presentadas en la sección de anexos.

Figura 29.

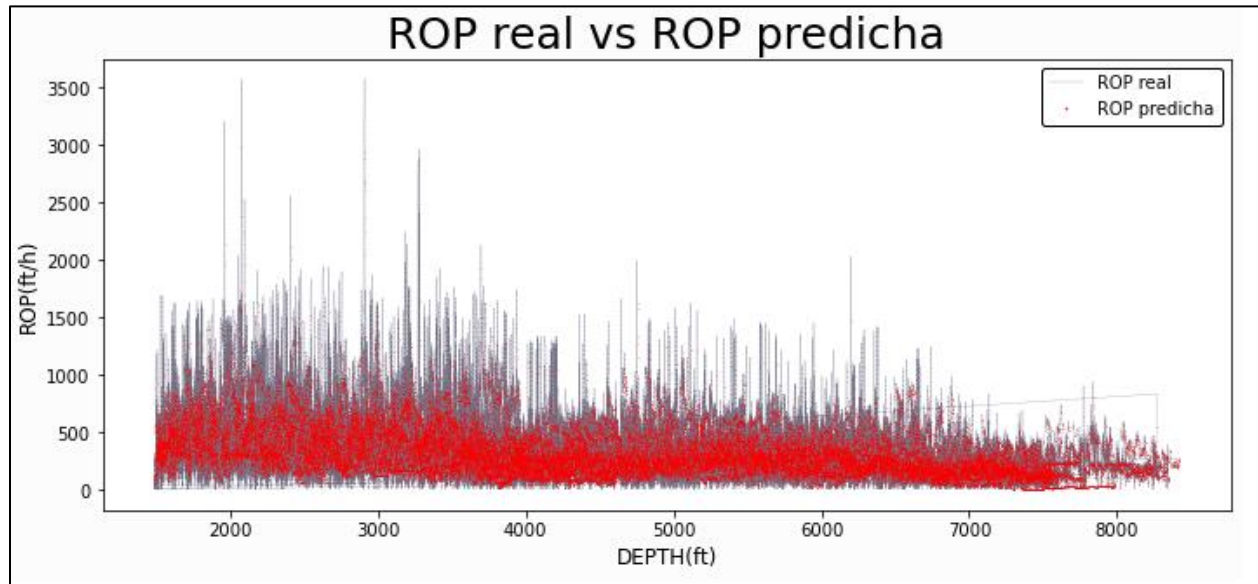
Validación ROP sección 12 ¼”



Nota. La figura muestra una validación aceptable para la ROP con algunos puntos dispersos debido a que no hay 100% de exactitud

Figura 30.

Validación ROP sección 8 ½”



Nota. La figura muestra una validación aceptable para la ROP con algunos puntos dispersos debido a que no hay 100% de exactitud

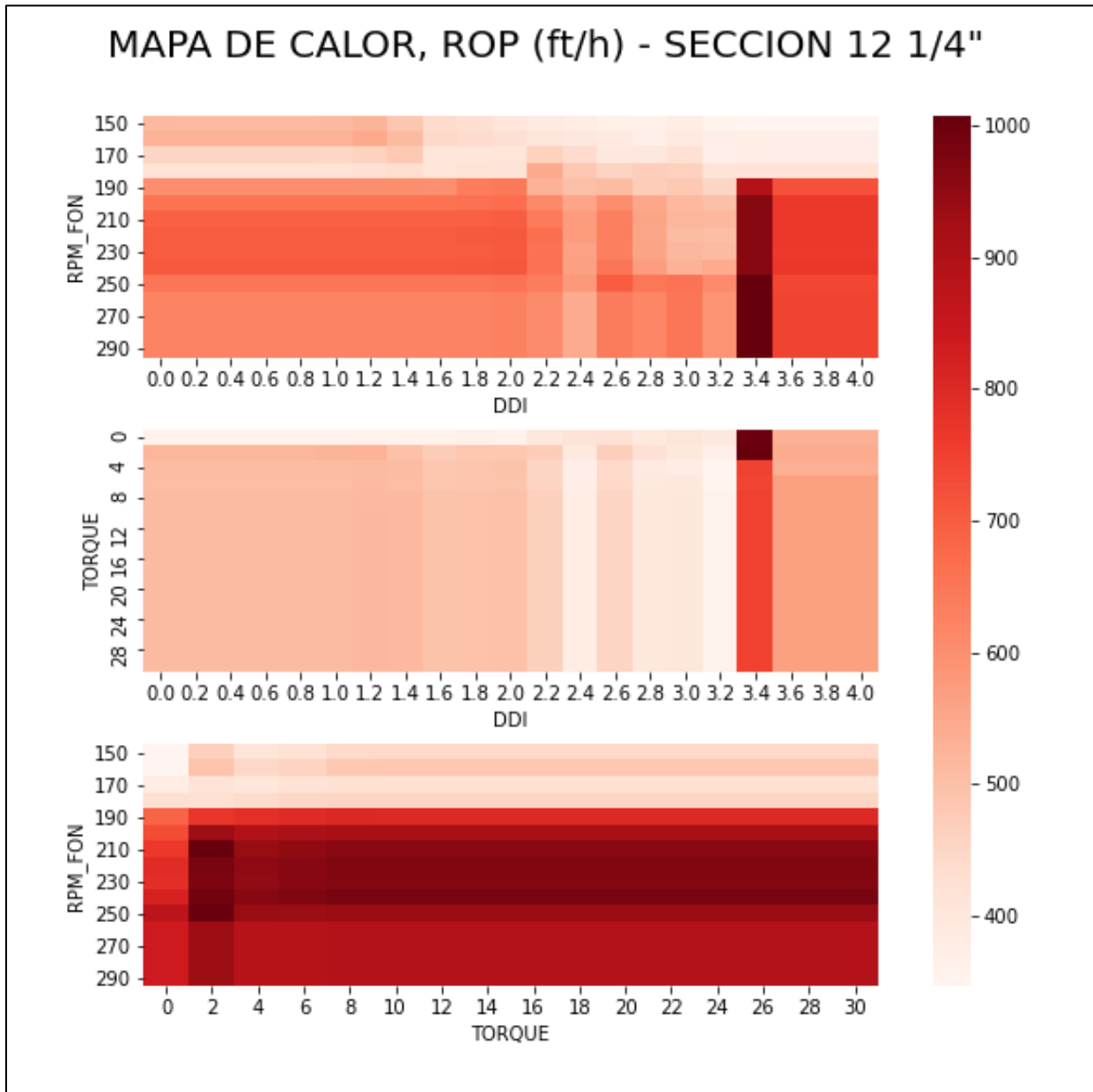
3.4 Generación de mapas de calor

Tomando como referencia la información contenida en las figuras 27 y 28 se procede a generar mapas de calor para cada variable. De ellos se extraen los rangos óptimos para los parámetros de perforación de las dos secciones en los pozos de estudio.

A continuación, se muestran los mapas de calor de ROP y MSE para las dos secciones. La sección de anexos contiene los mapas de calor para las demás variables contenidas en la ecuación que calcula la MSE.

Figura 31.

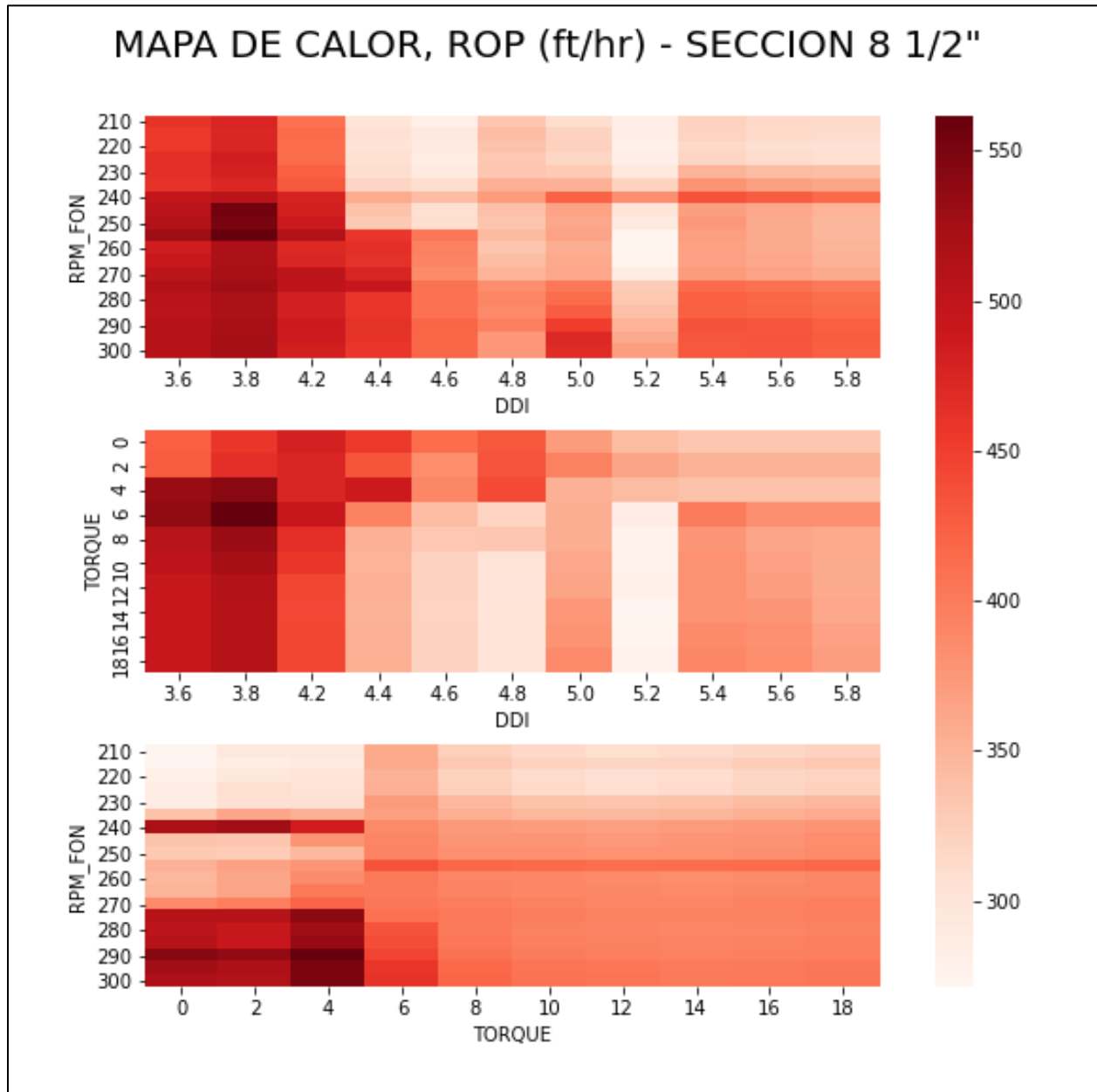
Mapa de calor ROP, sección 12 1/4"



Nota. La figura muestra los mapas de calor para la ROP en la sección 12 1/4"

Figura 32.

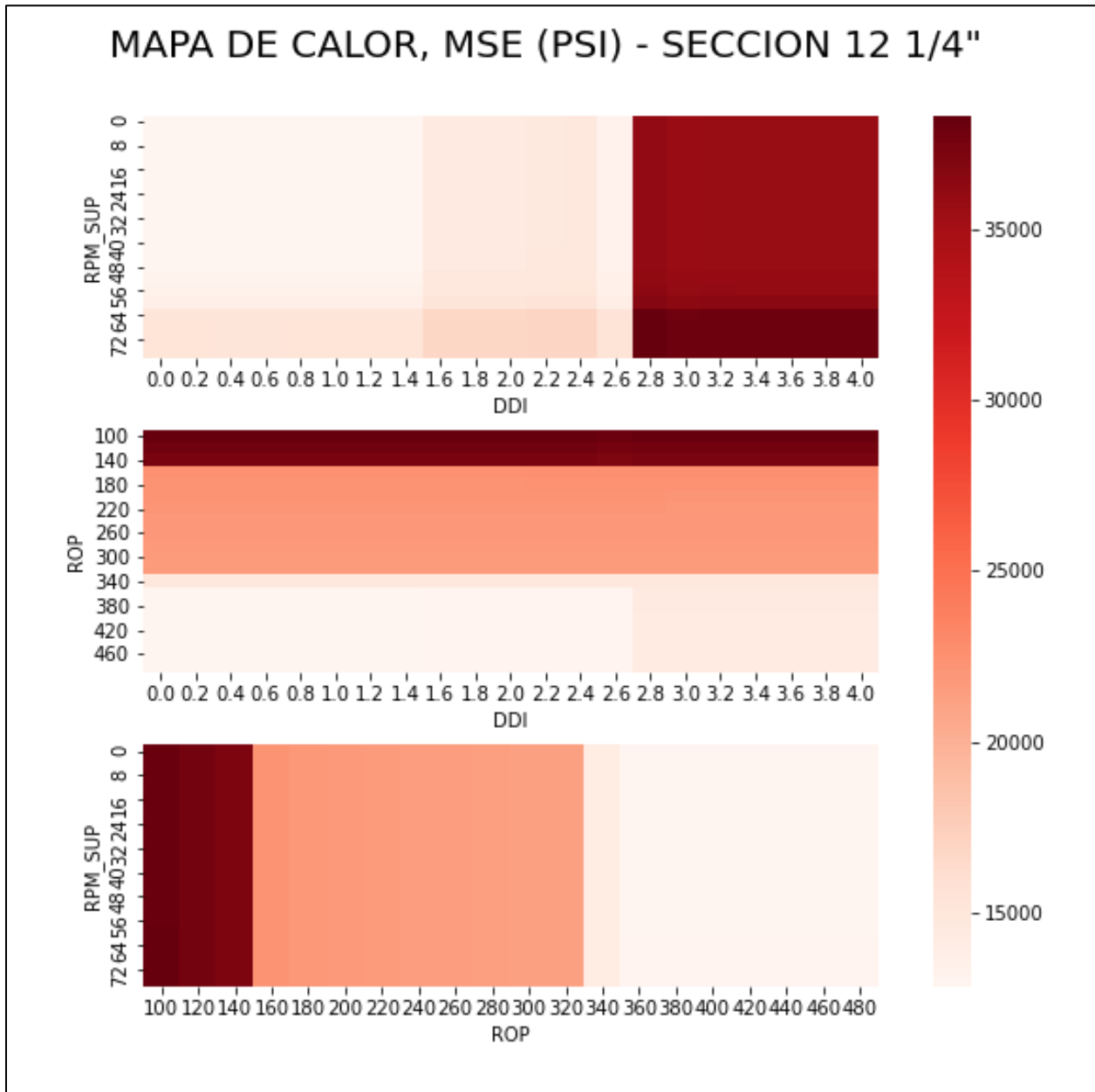
Mapa de calor ROP, sección 8 1/2"



Nota. La figura muestra los mapas de calor para la ROP en la sección 8 1/2"

Figura 33.

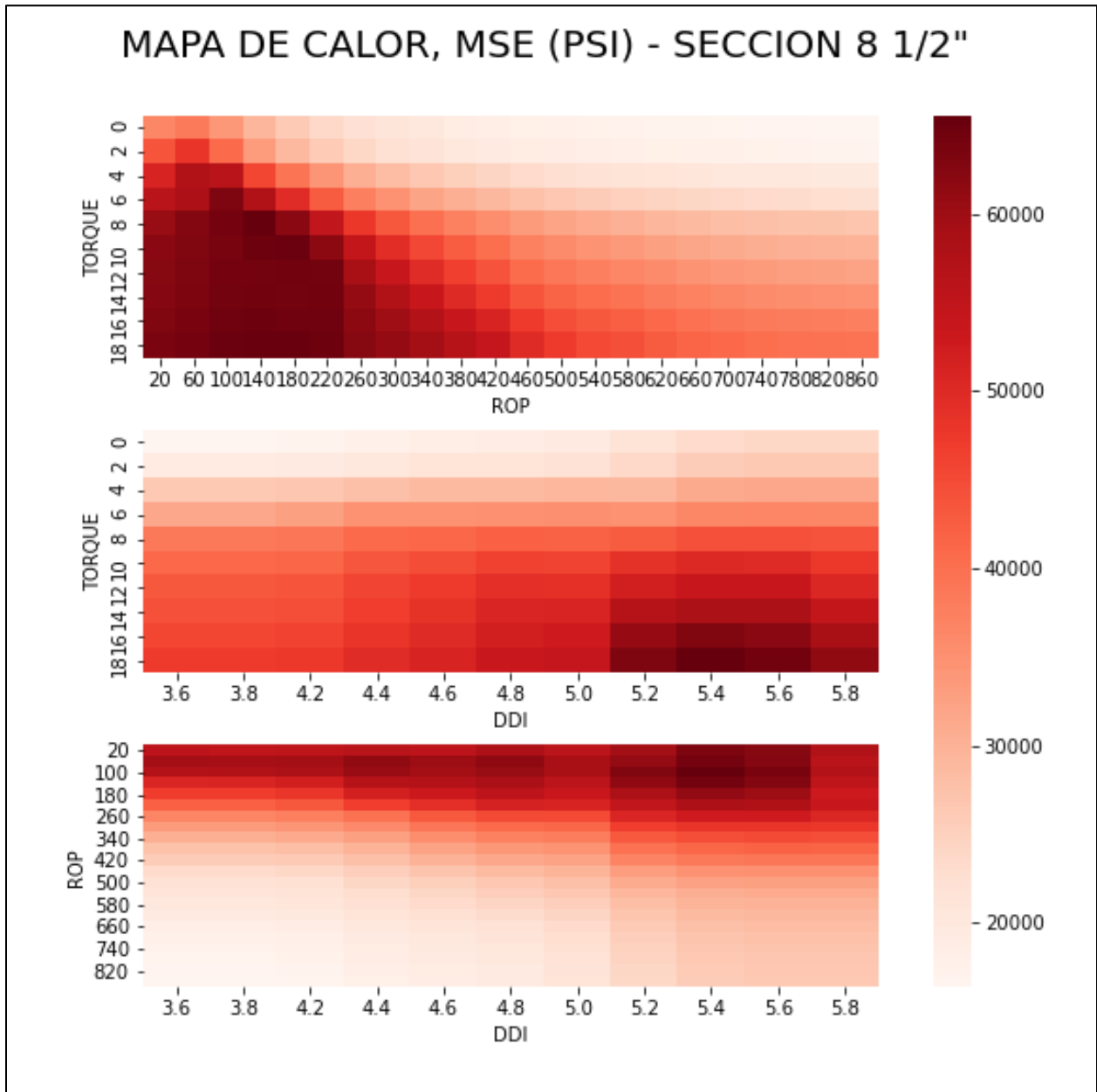
Mapa de calor MSE sección 12 1/4"



Nota. La figura muestra los mapas de calor para la MSE en la sección 12 1/4"

Figura 34.

Mapa de calor MSE, sección 8 ½"



Nota. La figura muestra los mapas de calor para la MSE en la sección 8 ½"

3.4.1 Análisis de los mapas de calor

3.4.1.i Sección 12 ¼". Una vez obtenidos los mapas de calor para las variables implícitas en la ecuación de la MSE tales como: RPM_FON, WOB, TORQUE, ROP junto al DDI se procederá a hacer el análisis correlativo, teniendo en cuenta los valores de rangos operativos óptimos arrojados por el modelo de predictibilidad.

La (RPM_FON) tiene como variables correlativas de mayor importancia (RPM_SUP, DDI y ROP), manteniéndose en rangos operativos de (110-160 rpm) con una relación directamente proporcional con la DDI alcanzando valores más altos en el rango de (3.3-4 unidades), puesto que la misma presenta una relación directa con la profundidad, por tanto, la remoción de estratos a mayor profundidad requiere una mayor cantidad de energía giratoria y por consiguiente mayores rpm, cabe resaltar que es pertinente no sobreexceder las rpm lo cual podría causar sobreesfuerzo en la broca y a corto plazo un desgaste muy elevado.[5]

Al correlacionar la RPM_FON y RPM_SUP se puede inferir que pueden ser directamente proporcionales hasta que en superficie se llegue a un rango de 56-64 rpm donde la RPM_FON podrán alcanzar valores de 160 rpm, lo cual está al límite del rango óptimo

En el caso de la correlación con la ROP se presenta lineal a lo largo del eje x debido a que en el rango de ROP (100-480) ft/h, no se reconoce cambio significativo a no ser que alcance valores superiores a 48 rpm en superficie, dando un incremento notable de 120 a 160 rpm en fondo.

El torque rotacional tiene como variables correlativas el peso sobre la broca (WOB), RPM_FON y el DDI, el rango optimo arrojado fue entre 1.5 y 3.25 klb-f, al representar como variable física la cantidad de fuerza en el momento del contacto con la roca, la correlación con el peso de esta es muy importante por lo cual los rangos en los que habrá una mayor transferencia de energía se presentan cuando el DDI está en un rango de 2.6 a 4 y WOB con valores entre 20 y 40 klb. La relación con la RPM_FON no se presenta dentro de los rangos óptimos puesto que se presentan los mayores torques (mayores a 3 klb-f) al sobrepasar los 190 rpm lo cual no coincide con el rango óptimo de la variable anteriormente mencionada. No obstante, la variabilidad del torque puede

depender de la dureza de la formación [6] por lo cual el aumento de las RPM no representa un descontrol en el manejo de la perforación.

El peso sobre la broca (WOB) arrojó valores óptimos entre 9 y 12 klb con RPM_FON, DDI y ROP como variables correlativas, teniendo los mejores valores (10-12 KLB) según el modelo en los rangos (220-480) ft/h y (2.2 a 3.2) para ROP y DDI respectivamente, el DDI representado en términos de profundidad, aunado a lo anterior, la mejor transferencia de energía a la broca y una tasa de penetración elevada serán registradas cerca a los 1100 pies.

El DDI tiene como variables correlativas el TORQUE , y las lecturas de revoluciones por minuto en superficie y fondo, el rango óptimo se predijo entre 2.5 y 3.1, no obstante esta variable al ser la representación de la dificultad presentada en el perfil direccional presenta correlaciones desfasadas con los rangos operativos, tales como torque entre 0 y 28 klb-f y RPM_SUP entre 40 y 76 , teniendo en cuenta que esta variable funciona como un parámetro externo de la transferencia de energía en el proceso de perforación, los resultados mencionados anteriormente no entorpecen la evaluación de la calidad del hueco.

Las predicciones arrojaron valores para la ROP un rango operativo entre 400 y 1000 ft/h con TORQUE, RPM_FON y DDI como variables correlativas de mayor importancia, se puede identificar que el mayor rango alcanzable durante la perforación(400-700ft/h), será con DDI menor a 3.2 y mayor a 3.6 , valores de torque entre 0 y 28 klb-f , no obstante los valores cercanos al límite superior (1000 ft/h) pueden verse reflejados en caso tal que las RPM_FON sobrepasen los 200 lo cual representa una correlación directamente proporcional desde ese punto en adelante, sin embargo, dicho incremento de RPM ocasionara un desgaste a todos los componentes de la sarta sin mencionar que hay formaciones poco consolidadas que requieren un paso de la operación de manera controlada[7]

La energía mecánica específica(MSE) es la representación conjunta de las variables previamente mencionadas a excepción del DDI, el rango óptimo arrojado por el modelo fue de 15000 a 35000 psi, las variables correlativas para este parámetro fueron la ROP, RPM_SUP y el DDI con lo cual se puede observar una relación inversamente entre la MSE y la ROP lo cual físicamente representa que a menor cantidad de energía requerida para la perforación habrá un mayor avance en la misma,

lo cual concuerda con el resultado del DDI, ya que a medida que el perfil direccional se torna más complicado la cantidad de energía requerida para perforar es mayor, y por ende, con valores mayores a 2.7 de DDI se presenta un incremento abrupto en la MSE, llegando al límite superior del rango óptimo (35000 psi), por ende en la fase final(1100-1500ft) de la sección 12 ¼” se presenta como la más difícil de perforar por factores direccionales y de cantidad de energía a emplear con base en los resultados predichos.

3.4.1.ii Sección 8 ¾”. El rango óptimo obtenido para el TORQUE fue de (3-10 klb-f), siendo las variables correlativas más importantes RPM_FON, RPM_SUP y DDI, para esta variable se presenta mucha variabilidad durante todo el rango de la sección, bien sea por que hay mayor cantidad de formaciones con descripción geológica compleja, o por cambios en el perfil direccional; junto al hecho de que con la construcción de esta sección se pretende alcanzar la formación objetivo. Con base en las predicciones se alcanzarán valores de torque con valores entre 260-280 y 30- 80 rpm en fondo y superficie respectivamente; y al mismo tiempo el DDI sobrepase 5.2 lo que indica direccionalmente hablando, un tramo de alta dificultad.

Se obtuvo un rango óptimo para la RPM_FON de (200-280rpm), con TORQUE, RPM_SUP, DDI como las variables correlativas de mayor importancia, en primer lugar, con un DDI superior a 4.6 las RPM_FON esperadas son superiores a 260 en cualquier rango de torque, por otro lado, durante todo el rango de DDI(3.6-5.8) a excepción de valores cercanos a 5.0 las RPM_FON esperadas serán cercanas al límite inferior 200 rpm, esto se debe principalmente a que en ese rango de profundidad se encuentra la formación colorado la cual presenta en su mayoría de composición geológica arcillas. [7]

Los valores para el rango óptimo del peso sobre la broca (WOB) fueron (10-18klb), con TORQUE, DDI y RPM_FON como variables correlativas de mayor importancia, se puede observar que los valores más altos de WOB se esperan con rangos de (8-18klb-f), (210-253) rpm y (4.9-5,1) para TORQUE, RPM_FON y DDI respectivamente. Aunado a lo descrito en el párrafo anterior, la cantidad de energía requerida para pasar por una formación compacta como lo es la Fm, Colorado se ve reflejada en las 3 variables, mayor cantidad de RPM_FON, por ende, mayor fuerza opuesta

(TORQUE), atravesar una formación compleja aumenta la posibilidad de creación de micro DLS (DDI) [8] y por ende la cantidad de energía a transmitir a la broca será mayor.

El rango óptimo para el DDI según la predicción fue (4.5-5.1), teniendo en cuenta que este parámetro depende de DLS, el AHD, y el azimut. Este rango óptimo se interpreta como el intervalo de valores donde las mejores condiciones operacionales pueden verse controladas con base en los demás rangos arrojados por el modelo predictivo, siendo RPM_FON, RPM_SUP y TORQUE las variables correlativas de mayor importancia, se puede observar que DDI, TORQUE y RPM bien sea de superficie o fondo tienen una relación directamente proporcional en la mayoría de su espectro, exceptuando un caso específico del mapa de calor donde con 60 rpm en superficie y un rango de (210-230rpm) en fondo se presentan valores bajos (4.2) de DDI.

Para la (ROP) el rango óptimo obtenido fue (250-600ft/h) , donde sus variables de importancia son RPM_FON, TORQUE y DDI, en contraste con la sección 12 ¼” la tendencia con el DDI no se ve reflejada totalmente plana, debido a la diferencia geológica y la necesidad de llegar a formación productora con parámetros controlados para evitar problemas de hueco, por ende, los rangos de ROP más altos 500-600 se ven para valores de DDI entre (3.6-4.2), esto se debe principalmente a que en este rango de profundidad se presentan las fases de construcción y una formación con tipo de grano fino. Con respecto al TORQUE no se ve una tendencia destacable, sin embargo, si se sobrepasan los 280 rpm se espera un torque bajo (1-5, klb-f) y un aumento considerable de ROP (>550 ft/h).

El rango óptimo para la MSE con base en los datos predichos fue (20000-60000)psi, teniendo como variables correlacionales de mayor importancia TORQUE, DDI y ROP, con base en el mapa de calor se puede identificar que en esta sección que los valores de ROP (>500 ft/h) tienen una relación inversa con respecto a la MSE, presentándose por todo el espectro del DDI (3.6-5.8) , esto puede verse representado principalmente a que hay mayor complejidad direccionalmente hablando, hay llegada a la zona productora , hay mayor cantidad de intercalaciones estratigráficas y a mayor profundidad. Junto al hecho que otras variables como presión de confinamiento y posibilidad de pegas aumentan, por lo cual controlar la tasa de penetración cobra mayor importancia, no obstante, la cantidad de energía necesaria para remover la roca aumenta

significativamente. En un caso puntual, se observa que hay tasas de penetración baja (20-180 ft/h) pero el torque está en sus más altos valores (8-18klb-f), por ende, la MSE está en su límite superior (<50000 psi), debido a la relación directa entre la cantidad de fuerza giratoria necesaria para remover la roca (>RPM_FON) por ende mayor TORQUE. De esta manera el modelo predictivo arrojó los parámetros óptimos para las secciones 12 ¼” y 8 ¾”

3.5 Rangos óptimos para los parámetros de perforación

Tabla 8.

Rangos óptimos variables sección 12 ¼”

Parámetros	ROP, ft/hr	DDI, adim	WOB, Klbs	TORQUE, Klbs-ft	RPM_FON, rpm	MSE, PSI
Rangos óptimos	400 - 1000	2.5 - 3.1	9.0 - 13.0	1.5 - 3.5	110 - 160	15000 - 35000

Nota. Rangos óptimos para los parámetros de perforación en la sección 12 ¼”

Tabla 9.

Rangos óptimos variables sección 8 ½”

Parámetros	ROP, ft/hr	DDI, adim	WOB, Klbs	TORQUE, Klbs-ft	RPM_FON, rpm	MSE, PSI
Rangos óptimos	250 - 600	4.5 - 5.1	10 - 18	3 - 10	200 - 280	20000 - 60000

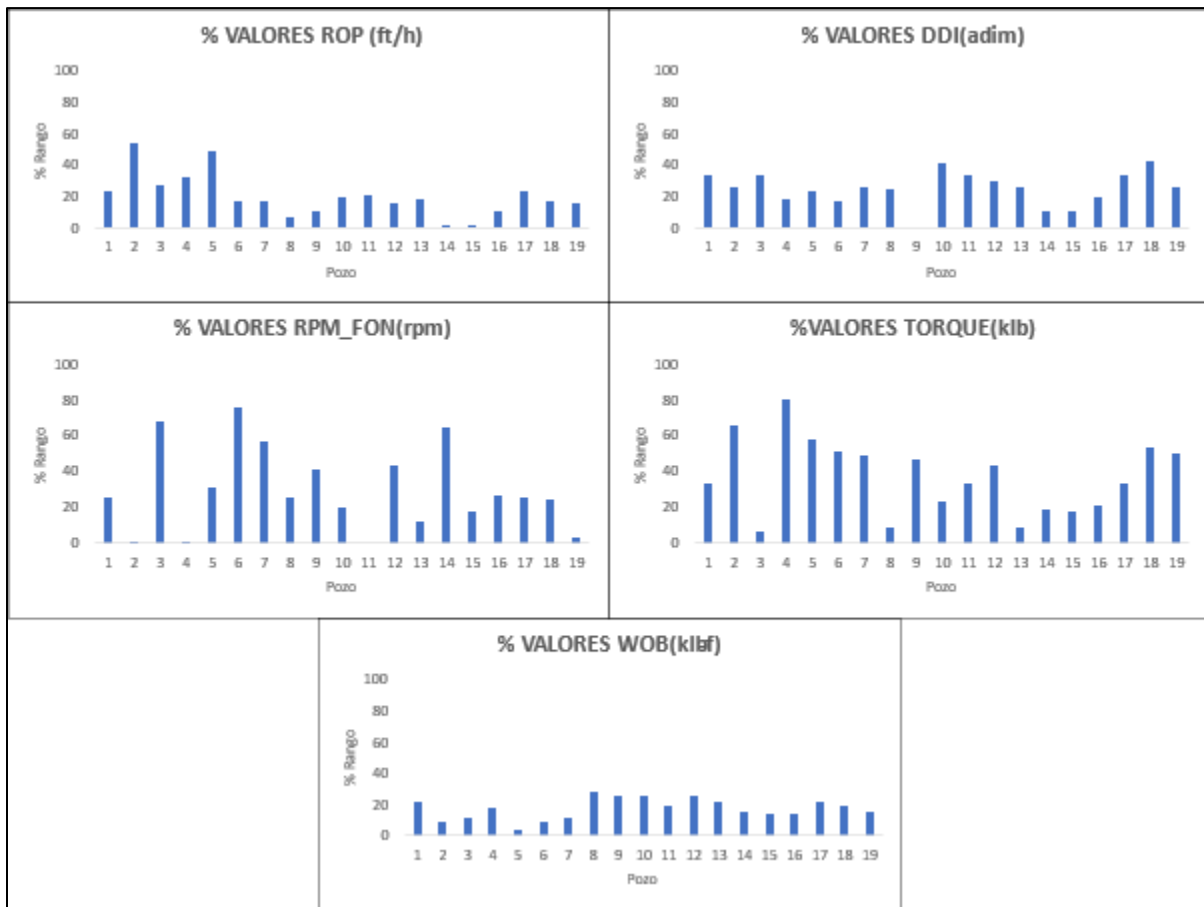
Nota. Rangos óptimos para los parámetros de perforación en la sección 8 ½”

3.6 Indicador de calidad de pozo (WQS)

Para la construcción del indicador de calidad de pozo se tomó como referencia los rangos óptimos de los parámetros de perforación. Con lo cual, en el dataset original donde están contenidos todos los datos de las dos secciones de estudio, se procede a calcular un porcentaje de cuántos datos de determinada variable de un pozo coinciden con el rango óptimo de esta. De este procedimiento se obtuvieron las siguientes gráficas, para la sección 12 ¼” y 8 ½”, respectivamente.

Figura 35.

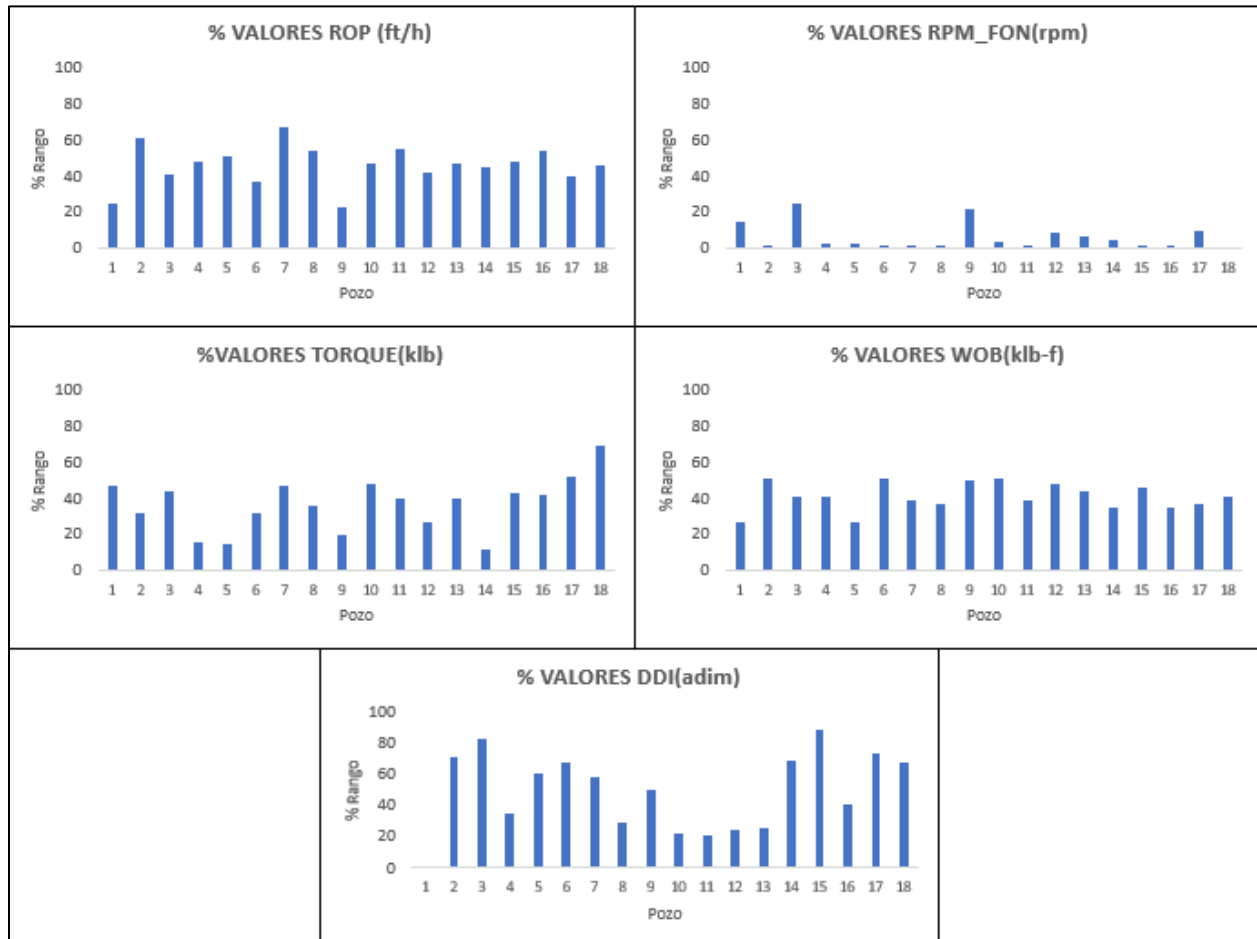
Porcentaje de coincidencia de variables, sección 12 ¼”



Nota. Porcentaje de coincidencia para cada parámetro de perforación en la sección 12 ¼”. Por temas de confidencialidad se asignó un valor numérico a cada pozo

Figura 36.

Porcentaje de coincidencia de variables, sección 8 ½”



Nota. Porcentaje de coincidencia para cada parámetro de perforación en la sección 8 ½”. Por temas de confidencialidad se asignó un valor numérico a cada pozo

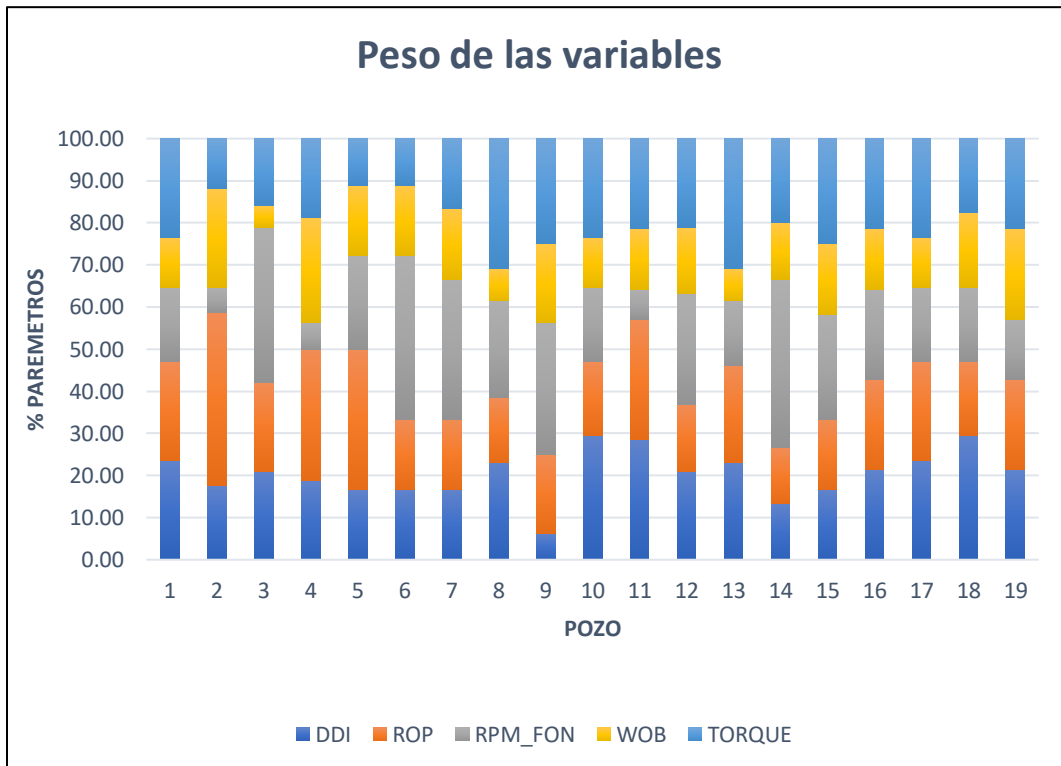
Como se puede observar, en la sección 12 ¼” de los pozos 9 y 11 para DDI y RPM_FON, respectivamente; no existe un porcentaje de coincidencia. Esto quiere decir que ningún dato de estos pozos coincide con el rango óptimo de las variables, influyendo negativamente en su calificación final (WQS).

Es evidentemente que la variable que menos coincidencia tiene es RPM_FON en la sección 8 ½”, debido al alto DDI promedio de la sección provocando una disminución en la rotación de la sarta.

Aplicando la conversión del porcentaje de coincidencia a escala de calificación de calidad de pozo (WQS) y sumando los valores de las variables de cada pozo para obtener su puntaje total, se obtuvo los siguientes resultados para la sección 12 ¼” y 8 ½”, respectivamente.

Figura 37.

Peso de variables en WQS, sección 12 ¼”

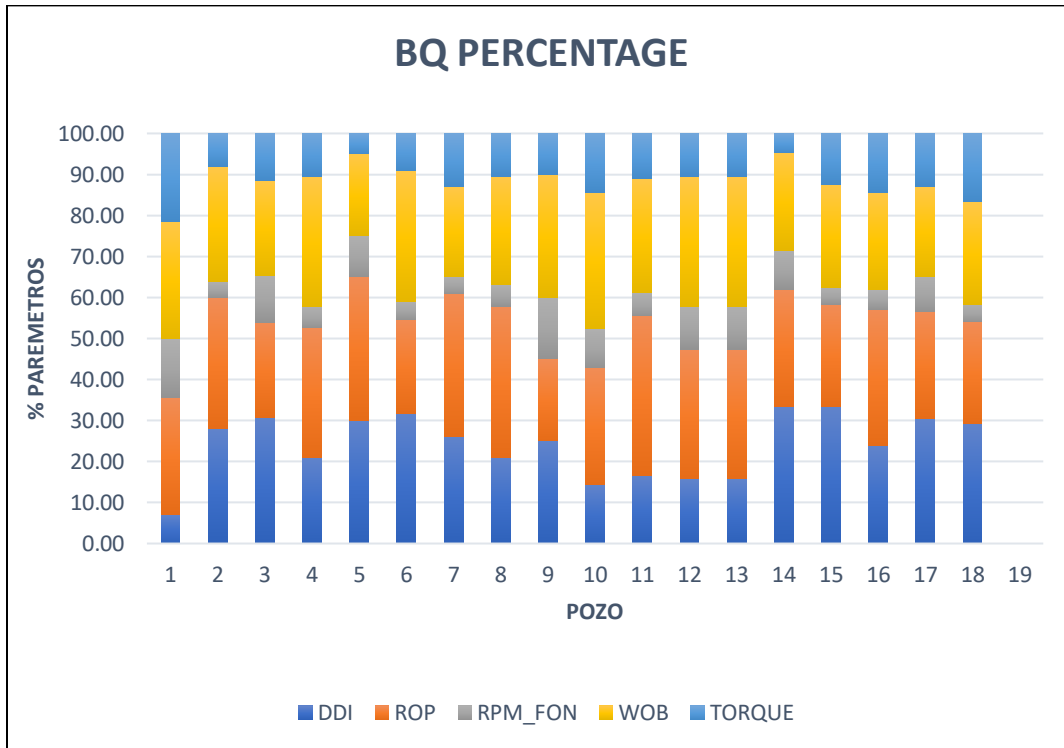


Nota. Influencia de los parámetros de perforación en la calificación final (WQS) de cada pozo para la sección 12 ¼”

En esta sección, el parámetro que generalmente más influye en la calificación final de cada pozo es la RPM_FON dado que los pozos tienen un espiralamiento considerable y esta variable es un muy buen indicador de la dificultad que tiene la sarta para girar y poder perforar de una manera eficiente.

Figura 38.

Peso de variables en WQS, sección 8 ½”



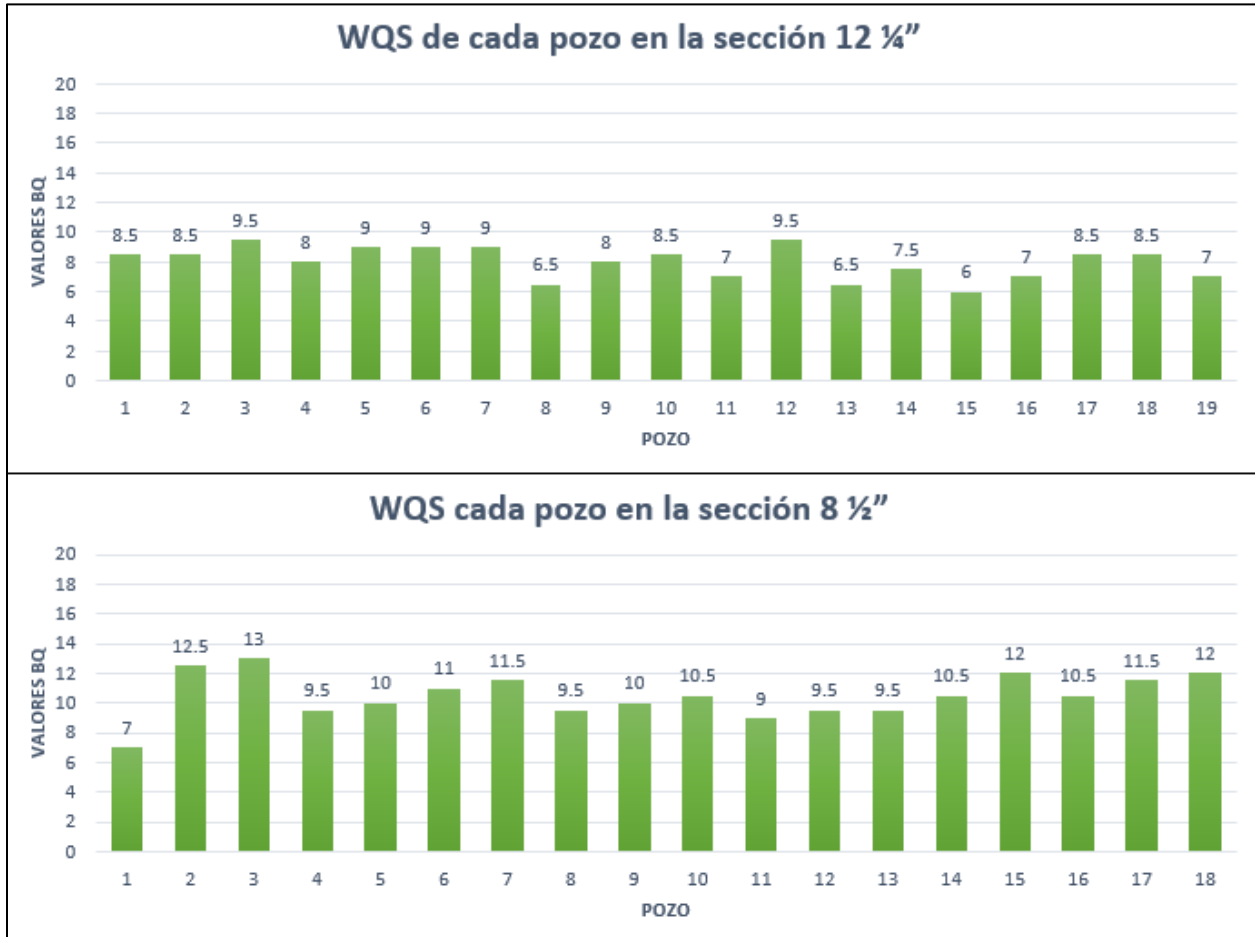
Nota. Influencia de los parámetros de perforación en la calificación final (WQS) de cada pozo para la sección 8 ½”

La influencia de las variables para la sección 8 ½” cambia. Para este caso en particular, el parámetro que más influye en la calificación final de cada pozo es la ROP. Esto se debe a que la sarta tiene mayor dificultad a la hora de perforar debido la índice de complejidad direccional (DDI) por lo que la ROP es un indicador certero de la capacidad que tiene la sarta para vencer el efecto del DDI en su contra.

Posteriormente, se realiza un consolidado de la calificación final de cada pozo estudiado en el proyecto para las secciones 12 ¼” y 8 ½” respectivamente.

Figura 39.

WQS para los pozos en estudio



Nota. Calificación de calidad de pozo (WQS) para los pozos en estudio de las secciones 12 1/4" y 8 1/2"

Aplicando los rangos cualitativos para la calificación final de un pozo (WQS) vistos en la figura 20, se asigna una descripción aproximada de los pozos estudiados. La siguiente tabla muestra los resultados obtenidos.

Tabla 10.*Descripción cualitativa de los pozos*

Pozo	12 ¼"	Descripción	8 ½"	Descripción
1	8.5	Pozo de calidad media	7	Pozo de calidad media
2	8.5	Pozo de calidad media	12.5	Pozo de calidad alta
3	9.5	Pozo de calidad media	13	Pozo de calidad alta
4	8	Pozo de calidad media	9.5	Pozo de calidad media
5	9	Pozo de calidad media	10	Pozo de calidad media
6	9	Pozo de calidad media	11	Pozo de calidad alta
7	9	Pozo de calidad media	11.5	Pozo de calidad alta
8	6.5	Pozo de calidad media	9.5	Pozo de calidad media
9	8	Pozo de calidad media	10	Pozo de calidad media
10	8.5	Pozo de calidad media	10.5	Pozo de calidad alta
11	7	Pozo de calidad media	9	Pozo de calidad media
12	9.5	Pozo de calidad media	9.5	Pozo de calidad media
13	6.5	Pozo de calidad media	9.5	Pozo de calidad media
14	7.5	Pozo de calidad media	10.5	Pozo de calidad alta
15	6	Pozo de baja calidad	12	Pozo de calidad alta
16	7	Pozo de calidad media	10.5	Pozo de calidad alta
17	8.5	Pozo de calidad media	11.5	Pozo de calidad alta
18	8.5	Pozo de calidad media	12	Pozo de calidad alta
19	7	Pozo de calidad media	10	Pozo de calidad media

Nota. Descripción cualitativa para los pozos de las dos secciones en estudio

La figura 40 es un elemento que confirma los antecedentes en los cuales se basó el proyecto desarrollado; en las campañas de perforación de los pozos para dos campos de hidrocarburos en la Cuenca del Valle Medio del Magdalena no se lleva un control riguroso de los parámetros de perforación y su implicación en la obtención de una buena calidad de pozo, es por esta razón que la mayoría de ellos tienen una calidad media. La falta de dicho control también se ve reflejada en los altos NPT's que poseen los campos en sus respectivos proyectos de perforación.

3.7 Generación de líneas base de MSE para los campos en estudio

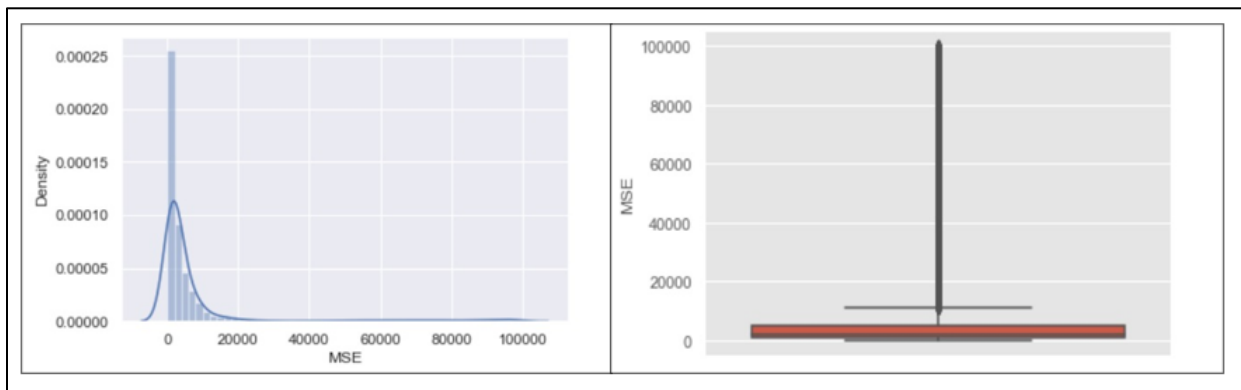
El procedimiento de creación de líneas base tiene la misma metodología de la predicción de parámetros de perforación; así, los rangos óptimos de esta variable serán las líneas base para cada sección de los pozos en los dos campos en estudio.

3.7.1 Caracterización estadística MSE

El inicio de la creación de líneas base tiene lugar en la caracterización estadística de la variable en cuestión. Para ello se generaron dos gráficos que resumen los datos de MSE con respecto a la profundidad vertical verdadera, ellos son: histogramas de distribución y diagramas de caja. Las siguientes figuras muestran los resultados obtenidos.

Figura 40.

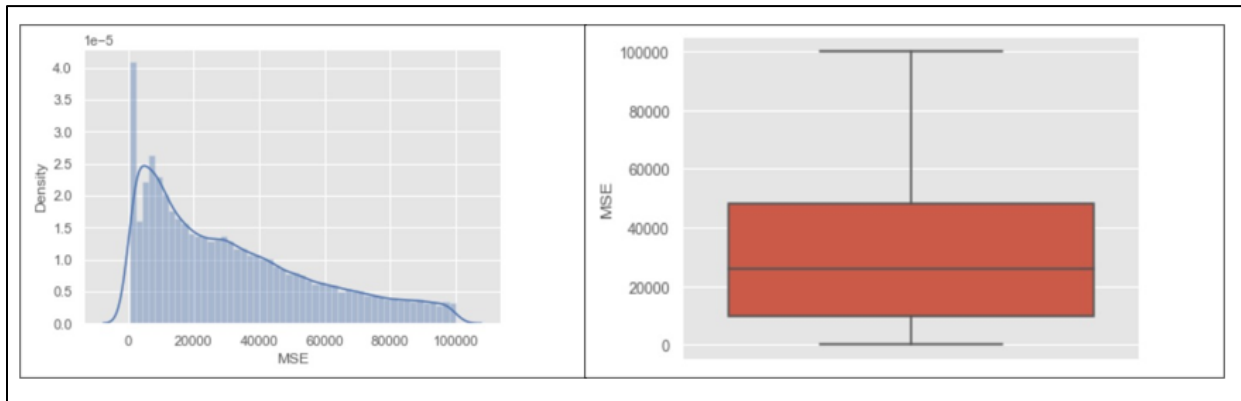
Análisis estadístico MSE, sección 12 ¼"



Nota. Diagramas de caja e histograma para MSE en la sección 12 ¼"

Figura 41.

Análisis estadístico MSE, sección 8 ½”



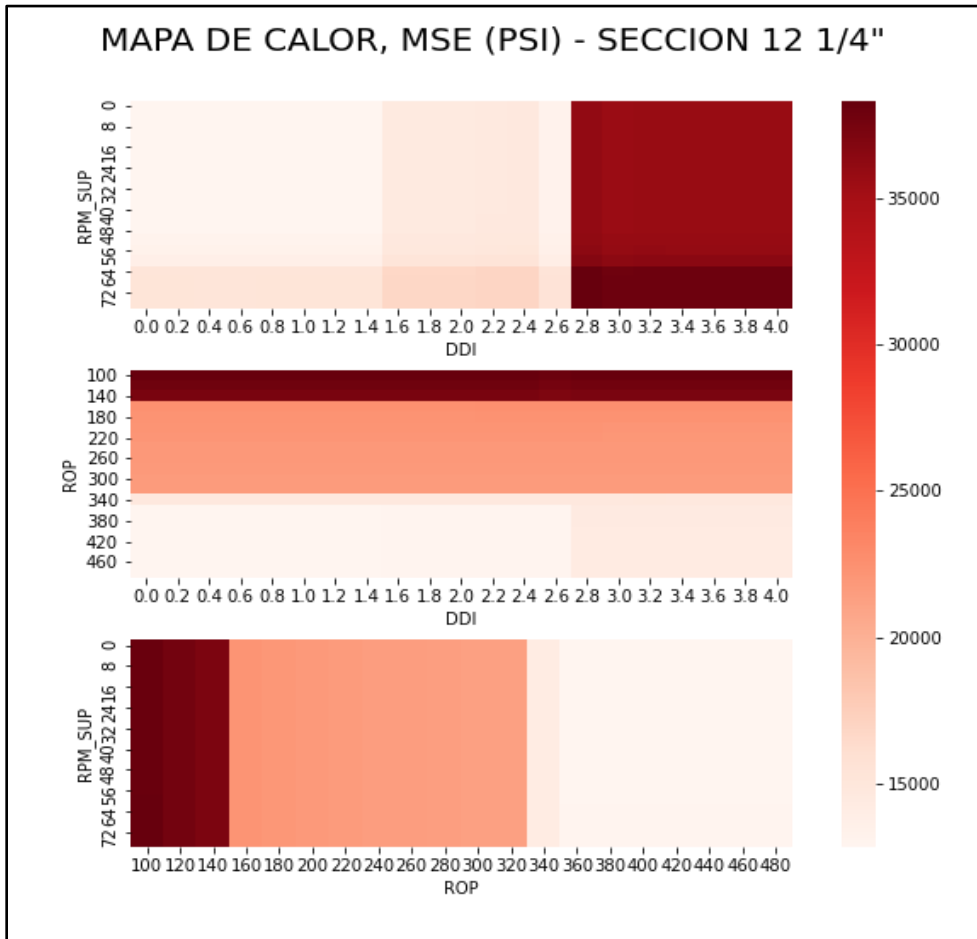
Nota. Diagramas de caja e histograma para MSE en la sección 12 ¼”

Como se puede observar en las figuras 41 y 42, el rango de entrada para la MSE es [0 - 8000] psi y [10000 - 50000] para la sección 12 ¼” y 8 ½”, respectivamente. Evidentemente el rango operativo de este parámetro es mayor en la segunda sección porque la MSE es una variable con tendencia a aumentar cada vez que se perfora en zonas de mayor profundidad.

Después de obtener los rangos de entrada se procede a crear un modelo predictivo *RandomForest* para generar datos nuevos y usarlos en la construcción de los mapas de calor respectivos. Se utilizó una proporción de 80 / 20 y 70 / 30 para la división del dataset en la sección 12 ¼” y 8 ½”, arrojando precisiones para el entrenamiento y prueba del modelo iguales a 83.73% y 95.33%, respectivamente. Al ser las precisiones mayores al 70% se procede a generar los mapas de calor. Aplicando el mismo procedimiento utilizado para las demás variables se obtuvo las figuras 33 y 34. Los rangos presentados en la tabla 11 serán considerados como las líneas base, teniendo en cuenta que dichos rangos se obtuvieron de los mapas de calor mostrados en las figuras 42 y 43, tanto para la sección 12 ¼” como para la sección 8 ½” en las cuales la compañía se debe basar al momento de perforar diferentes pozos en los dos campos para los que el proyecto fue desarrollado.

Figura 42.

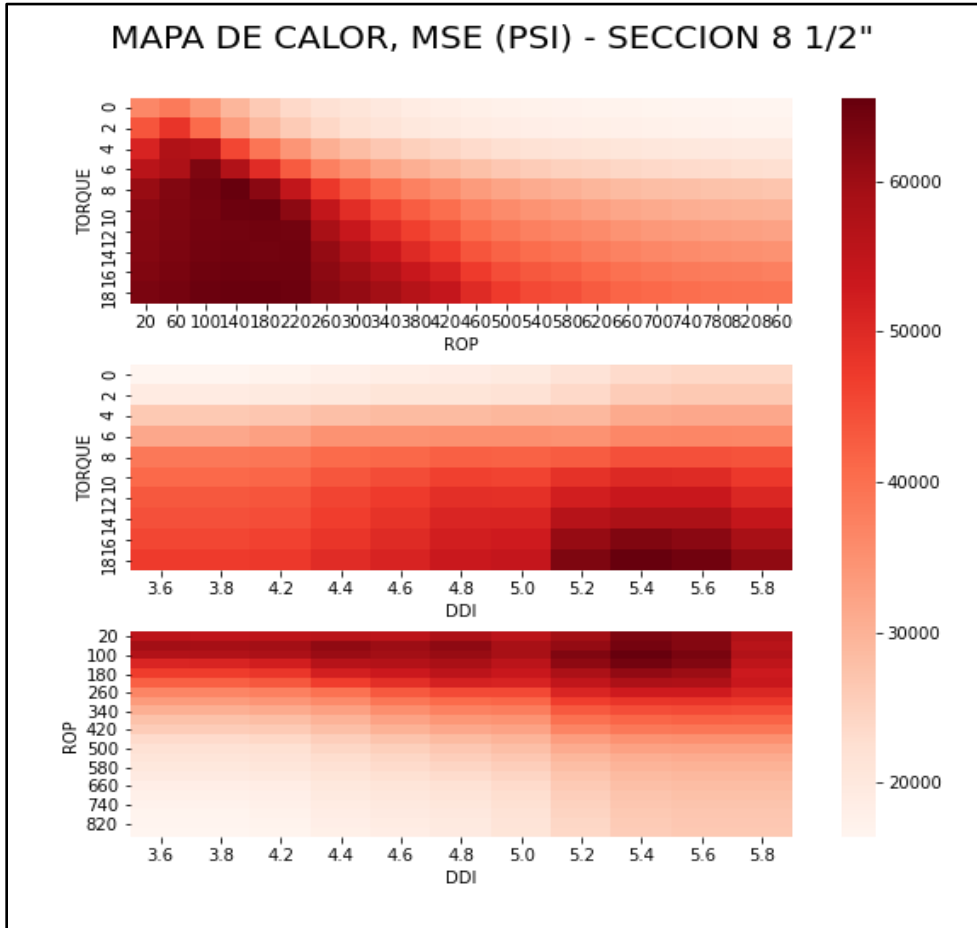
Mapa de calor MSE, sección 12 1/4"



Nota. Mapas de calor de MSE con su respectiva escala de medición para la sección 12 1/4"

Figura 43.

Mapa de calor MSE, sección 8 1/2"



Nota. Mapas de calor de MSE con su respectiva escala de medición para la sección 8 1/2"

Tabla 11.

Líneas base de MSE

Sección	Líneas base
12 1/4"	15000 - 35000
8 1/2"	20000 - 60000

Nota. Valores de las líneas base para las secciones en estudio

3.8 Análisis financiero del WQS

Debido a la confidencialidad que la compañía Ecopetrol S. A maneja en el aspecto de su información financiera, el análisis económico para el indicador de calidad (WQS) quedó limitado por los siguientes aspectos:

- Se manejará un costo promedio de pozo en USD, extrapolado para todos los pozos de un año.
- Se asumirá una ROP constante a lo largo de la perforación efectiva de un pozo.
- Los tiempos manejados en el análisis serán netos (sin NPT's y eventos aislados).
- La ROP a manejar para el análisis será la efectiva para los pozos.

La tabla 12 muestra los resultados de la comparación entre la duración y el costo de los pozos en los años 2019 y 2020.

Tabla 12.

Comparación de costos y duración de perforación

POZOS	VARIABLES	SECCIÓN 12 1/4"	SECCIÓN 8 1/2"
2020	Tiempo(hr)	17	108.2
	Profundidad(ft)	0-1490	1490-8800
	longitud Sección(ft)	1490	7310
	ROP promedio(ft/h)	87	68
	Costo Promedio Pozo(USD)	1660000	
2019	Tiempo(hr)	19	75
	Profundidad(ft)	0-1490	1490-8000
	longitud Sección(ft)	1490	6510
	ROP promedio(ft/h)	87	78
	Costo Promedio Pozo(USD)	1690000	
PREDICCIÓN (PERFORACIÓN)	Tiempo(hr)	5.5	34.8
	Profundidad(ft)	0-1490	1490-8800
	longitud Sección(ft)	1490	7310
	ROP promedio(ft/h)	400	300
	Costo Promedio Perforación (USD)	605576	

Nota. Comparación de costos y duración de perforación para los pozos de los años 2019 a 2020 con respecto a datos reales y datos de predicción.

4. CONCLUSIONES

El análisis exploratorio de datos (ADE) mediante la estadística descriptiva permitió consolidar los rangos operacionales para las variables de estudio, dichos valores son: (1-115) y (1-120) revoluciones por minuto en superficie, (26-266) y (30-500) revoluciones por minuto en fondo, (0-15) y (0-32) klb-f de torque, (0-32) y (0-44) klbs de peso en la broca, (10-900) y (10-800) para la tasa de perforación, (1-3.6) y (2-5.8) unidades de DDI para las secciones 12 ¼” y 8 ¾” respectivamente, identificando su distribución, valores atípicos, desviación, entre otros.

La exactitud generada por el modelo para datos de entrenamiento y prueba son: (>95%) y (70-95%) respectivamente, esto se debe principalmente a que, al validar el modelo con una mayor cantidad de datos, dicho modelo reconoce patrones; con lo cual la generación de nueva data es afín a la validación.

La proporción para la ROP predilecta fue 90/10 para entrenamiento y prueba respectivamente, al arrojar exactitudes de 81.59% y 70.34% para las secciones 12 ¼” y 8 ¾” respectivamente. Las proporciones para la RPM_FON elegidas fueron 70/30 y 80/20 con exactitudes de 95.41% y 84.55%, para las secciones 12 ¼” y 8 ¾” respectivamente. Las proporciones para el TORQUE elegidas fueron 80/20 y 90/10 con exactitudes de 72.63% y 93.99%, para las secciones 12 ¼” y 8 ¾” respectivamente. Las proporciones para el WOB elegidas fueron 80/20 y 90/10 con exactitudes de 72.18% y 71.55%, para las secciones 12 ¼” y 8 ¾” respectivamente. Las proporciones para el DDI elegidas fueron 90/10 y 80/20 con exactitudes de 92.91% y 84.61%, para las secciones 12 ¼” y 8 ¾” respectivamente.

El algoritmo de regresión supervisado Random Forest Regressor fue la mejor opción al presentar en todas las variables predichas exactitudes superiores al 70%, esto se debe a su comportamiento no lineal y el ajuste de hiper parámetros realizado el cual generó un aumento de 0.4% en promedio para cada una de las variables, lo cual es ideal para parámetros de perforación.

Se evidenció que el DDI tiene una correlación global; puesto que se presenta como de mayor importancia para cada variable de estudio en cada sección (Figura 27).

La relación entre la MSE y TORQUE es proporcional, como se muestra en (Ecuación 1) de manera numérica y gráficamente (Figura 34), lo cual físicamente es correcto; la fuerza torsional aumenta debido a la cantidad de energía requerida para remover un volumen de roca, contrario a lo que pasa entre MSE y ROP, la ROP disminuye mientras mayor cantidad de energía es requerida como se muestra en las (Figura 33) y (Figura 34) para las secciones 12 ¼” y 8 ¾” respectivamente.

Los resultados del indicador de calidad (WQS) evidencian la falta de control y seguimiento de parámetros de perforación. El 95% de los pozos muestra una clasificación “Calidad de pozo media”.

La creación de las líneas base de MSE dieron como resultado rangos de [15 – 35000] psi y [20000 - 60000] psi para las secciones 12 ¼” y 8 ½”, respectivamente.

BIBLIOGRAFÍA

- [1] H. Portilla y Y. Mejía, «OPTIMIZACIÓN DE PARÁMETROS DE PERFORACIÓN CON MSE E IMPACTO EN LA CONSTRUCCIÓN DE UN POZO EN EL CAMPO YARIGUÍ – CANTAGALLO,» *El reventón energético*, vol. 12, n° 2, 2014.
- [2] F. Gamperl, «Analysis of the impact of wellbore quality,» 2017. [En línea]. Available: <https://n9.cl/es>. [Último acceso: 17 Octubre 2020].
- [3] M. L. Waskom, «Seaborn: statistical data visualization,» *Journal of Open Source Software*, vol. 6, n° 60, 2011.
- [4] R. López Briega, «Libro Online de IAAR,» IAAR, 26 Octubre 2015. [En línea]. Available: <https://n9.cl/7rkra>. [Último acceso: 29 Abril 2021].
- [5] C. Szulma, «Academia,» 2015. [En línea]. Available: <https://n9.cl/tku0g>. [Último acceso: 15 Abril 2021].
- [6] J. Izquiero y S. Granados Ortíz, «Manual técnico,» 2011. [En línea]. Available: <https://n9.cl/95u9b>.
- [7] R. A. Jiménez, «FACTORES A TENER EN CUENTA AL MOMENTO DE REALIZAR UNA PERFORACIÓN PETROLERA PARA MINIMIZAR EL IMPACTO AMBIENTAL,» *Universidad Militar Nueva Granada*, 2017. [En línea]. Available: <https://n9.cl/uuh1n>. [Último acceso: 7 Junio 2021].
- [8] M. Salvador Figueras y P. Gargallo, «Análisis exploratorio de datos (AED),» 2003. [En línea]. Available: <https://n9.cl/6dcf3>. [Último acceso: 12 Mayo 2021].
- [9] A. Mendez Castro, «Course Hero,» 2013. [En línea]. Available: <https://n9.cl/uuh1n>. [Último acceso: 20 Mayo 2021].
- [10] N. Tabares Rodríguez y D. Tobar Castilla, «IMPLEMENTACIÓN DE UN MODELO PREDICTIVO DE MACHINE LEARNING PARA LA ESTIMACIÓN DE LOS PARÁMETROS ÓPTIMOS DE LA ROP Y LA MSE EN LA SECCIÓN 8½" Y 12 ¼" PARA LOS POZOS PERFORADOS CON MOTOR DE FONDO EN EL CAMPO YARIGUI - CANTAGALLO DURANTE EL 2019,» *Fundación Universidad de América*, 2021. [En línea]. Available: <https://n9.cl/uuh1n>. [Último acceso: 15 Junio 2021].

[11] C. J. Mason y C. David C-K, «The Perfect Wellbore!», *SPE*, 2005. [En línea]. Available: <https://n9.cl/uuh1n>. [Último acceso: 2 Junio 2021].

GLOSARIO

Algoritmo: serie de pasos repetibles para llevar a cabo cierto tipo de tarea con datos.

Argumento: son aquellos valores sobre los que se realizan las operaciones dentro de la función.

Atributo (Feature): variable de entrada que se usa para realizar predicciones.

Array: es una estructura de datos de secuencia que se ve muy parecida a una Lista, exceptuando que todos los miembros tienen que ser del mismo tipo.

Azimut: la dirección magnética de un levantamiento direccional o del pozo. Se expresa generalmente en grados con respecto al polo norte geográfico o magnético.

Big Data: una rama de las Tecnologías de la Información que estudia las dificultades inherentes a la manipulación de grandes conjuntos de datos.

Broca: una broca de perforación es un dispositivo conectado al extremo de la sarta de perforación que rompe, corta o aplasta las formaciones rocosas para perforar un pozo.

Ciencia de datos: es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos.

Conjunto de entrenamiento (Training Set): subconjunto del conjunto de datos que se usa para entrenar un modelo predictivo.

Conjunto de prueba (Test Set): subconjunto dentro del conjunto de datos que es utilizado para probar un modelo predictivo.

Dataframe: tipo de dato en Python utilizado para representar conjuntos de datos en es análogo a una tabla.

Dato: es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa.

Entrenamiento: en machine learning se conoce como entrenamiento al proceso de determinar los parámetros ideales que conforman un modelo.

Estadística: ciencia de aprender de los datos o como la ciencia de obtener conclusiones en la presencia de incertidumbre. Se divide en dos grandes ramas: Estadística descriptiva y Estadística inferencial

Etiqueta (Label): en el aprendizaje automático es considerada como la variable a predecir de un modelo.

Exactitud: fracción de predicciones que se realizaron correctamente en un modelo de clasificación. 89

Hiperparámetro: son los valores de las configuraciones utilizadas para el funcionamiento de un algoritmo durante el proceso de entrenamiento.

Indicador de calidad de hueco (WQS): lista de rangos óptimos de parámetros de perforación utilizada para medir la eficiencia de la operación.

Indicador de dificultad direccional(DDI): parámetro de perforación utilizado para evaluar la complejidad del perfil direccional

Línea base: valor de un parámetro utilizado como referencia para la medición del desempeño de este

Matplotlib: biblioteca para la generación de gráficos y visualizaciones en Python.

Matriz: conjunto de números o términos dispuestos en filas y columnas.

Modelo: en Machine Learning, se considera como modelo el objeto que va a representar la salida del algoritmo de aprendizaje. El modelo es lo que se utiliza para realizar las predicciones.

Numpy: biblioteca matemática de código abierto que proporciona operaciones entre matrices eficaces en Python.

Pandas: librería de Python que proporciona estructuras para el manejo de Dataframes.

Precisión: métrica que permite identificar la frecuencia con la que un modelo predijo correctamente.

Predicción: resultado de un modelo cuando se le proporciona información de entrada.

Python: lenguaje de programación creado en 1994 reconocido por su facilidad de uso y gran potencia, actualmente es uno de los lenguajes más utilizados en la inteligencia artificial y la ciencia de datos.

Regresión: consiste en encontrar la mejor relación que representa al conjunto de datos.

Scikit-Learn: una de las librerías más utilizadas para Machine Learning en Python.

Seaborn: librería para la visualización de datos en Python, basada en Matplotlib.

Sobreajuste (Overfitting): comportamiento de un modelo que coincide de tal manera con los datos de entrenamiento que no realiza predicciones correctas con datos nuevos. No aprende de los datos, se aprende los datos.

Subajuste (Underfitting): comportamiento de un modelo que no permite que el mismo sea capaz de reconocer tendencias o comportamientos en los datos

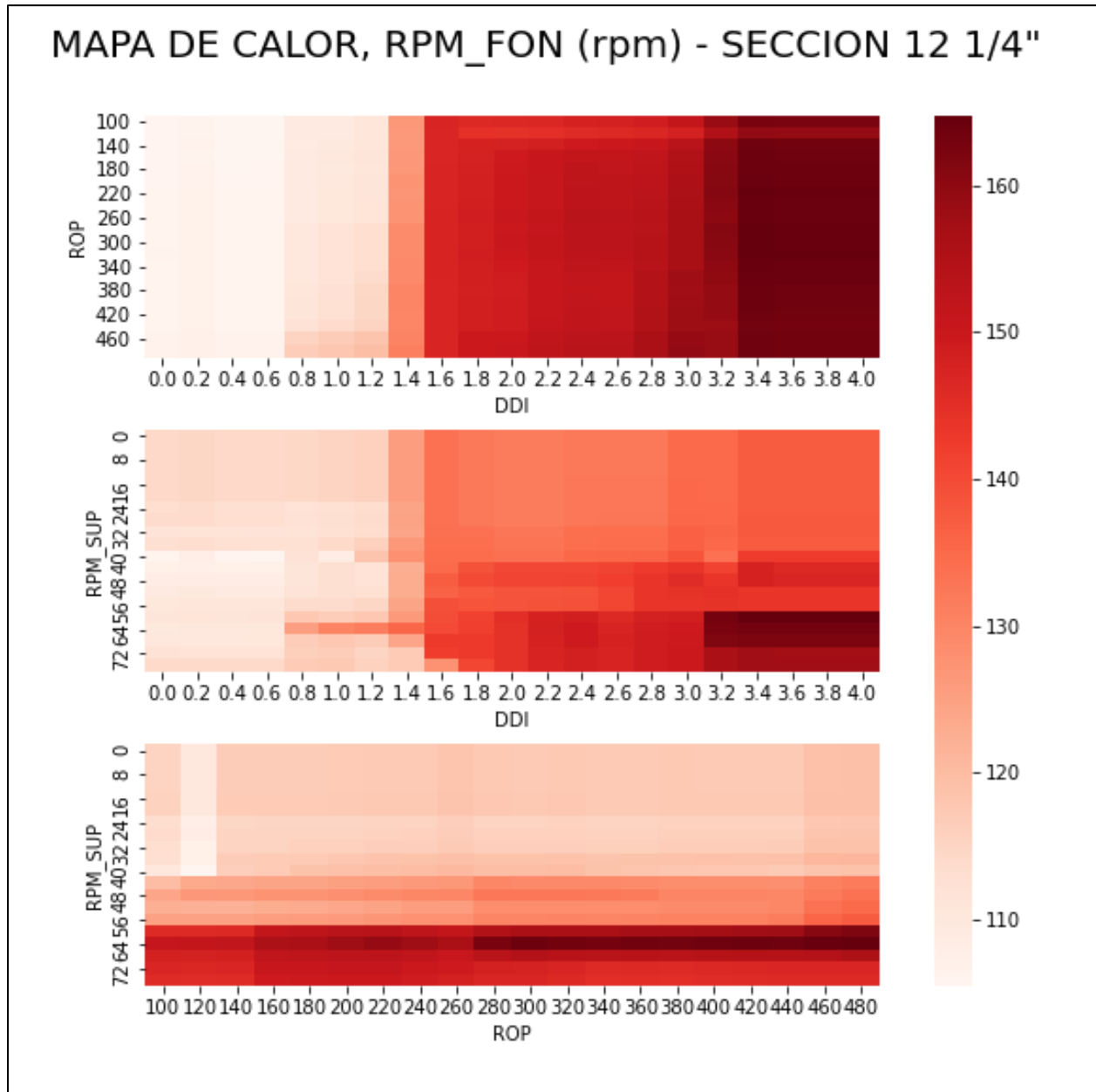
ANEXOS

ANEXO 1.

MAPAS DE CALOR PARA PARÁMETROS DE PERFORACIÓN, SECCIONES 12 ¼" Y 8 ½"

Figura 44.

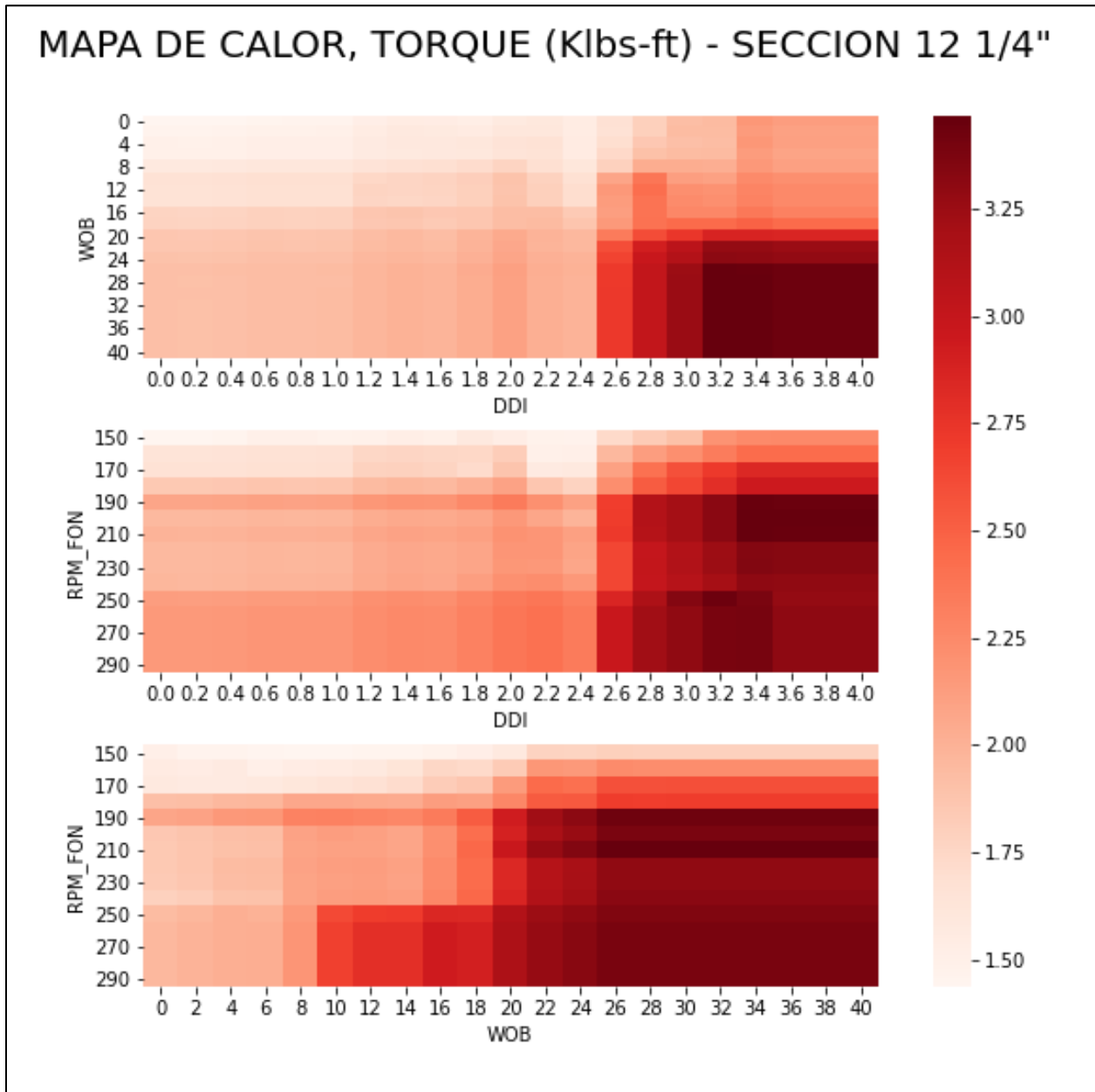
Mapa de calor para RPM_FON, sección 12 ¼"



Nota. Mapa de calor para RPM_FON en la sección 12 ¼" con un rango operativo de 110 a 160 rpm

Figura 45.

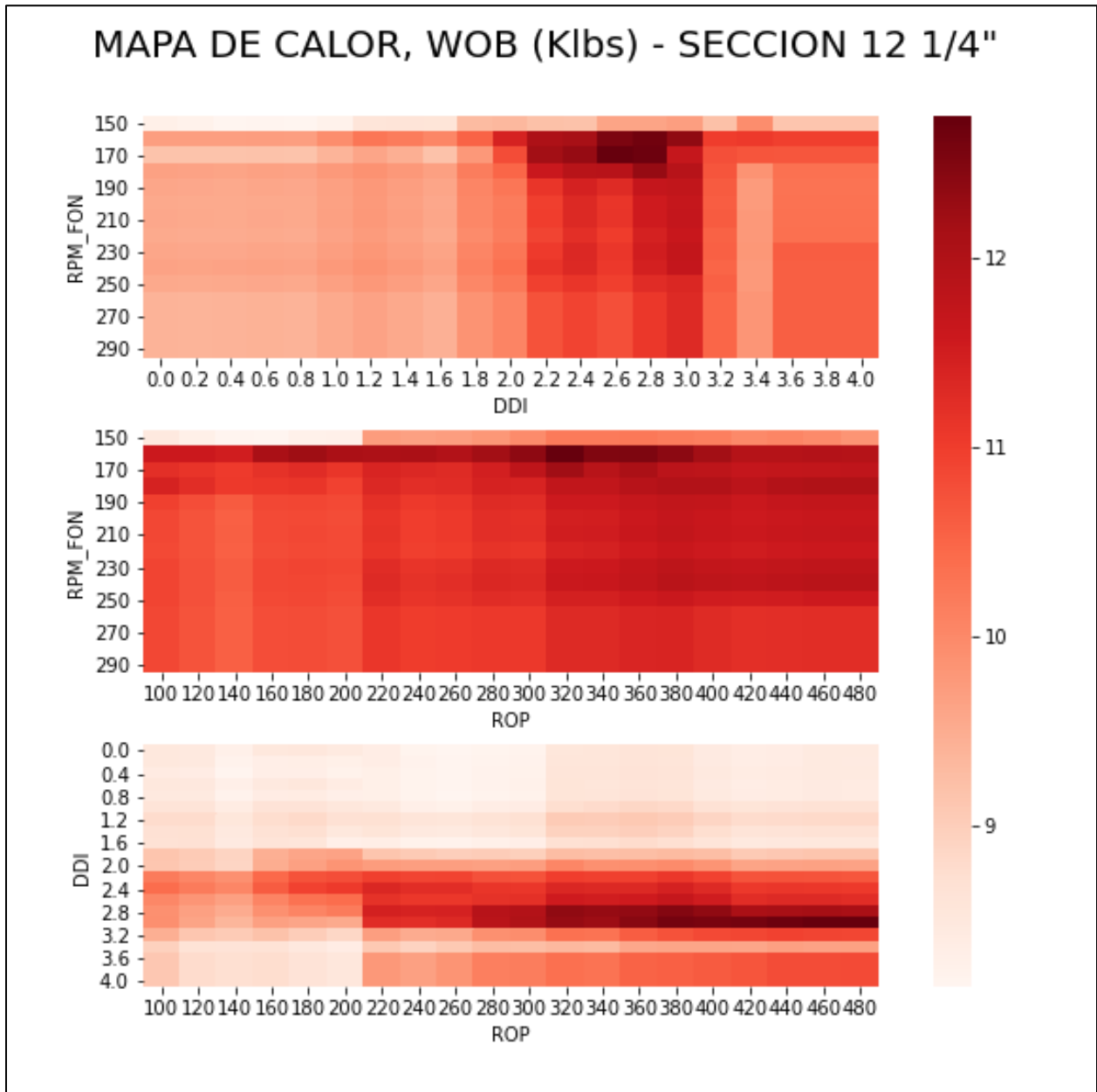
Mapa de calor para TORQUE, sección 12 1/4"



Nota. Mapa de calor para TORQUE en la sección 12 1/4" con un rango operativo de 1.5 a 3.25 Klbs-ft

Figura 46.

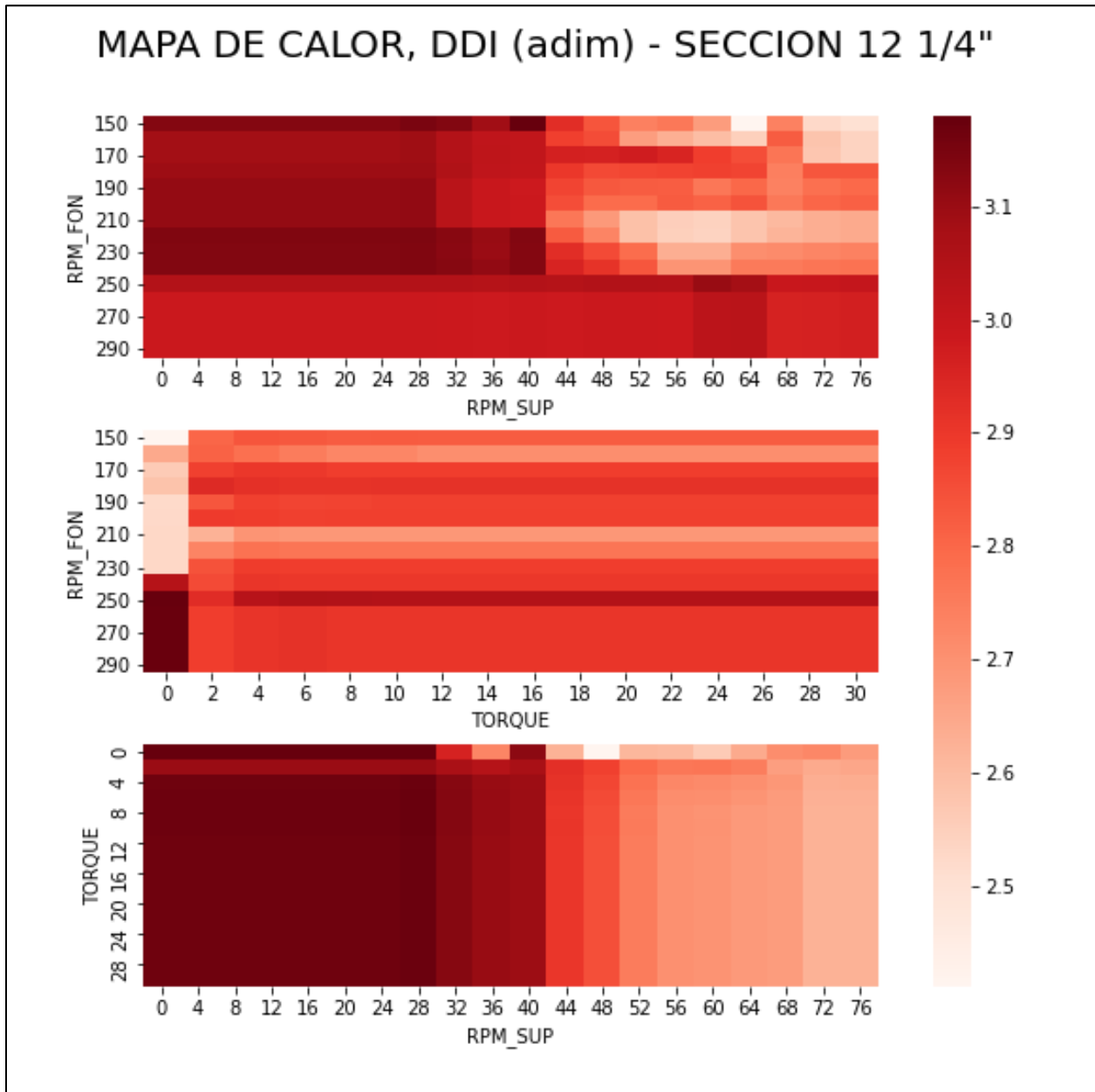
Mapa de calor para WOB, sección 12 1/4"



Nota. Mapa de calor para WOB en la sección 12 1/4" con un rango operativo de 9 a 12 Klbs

Figura 47.

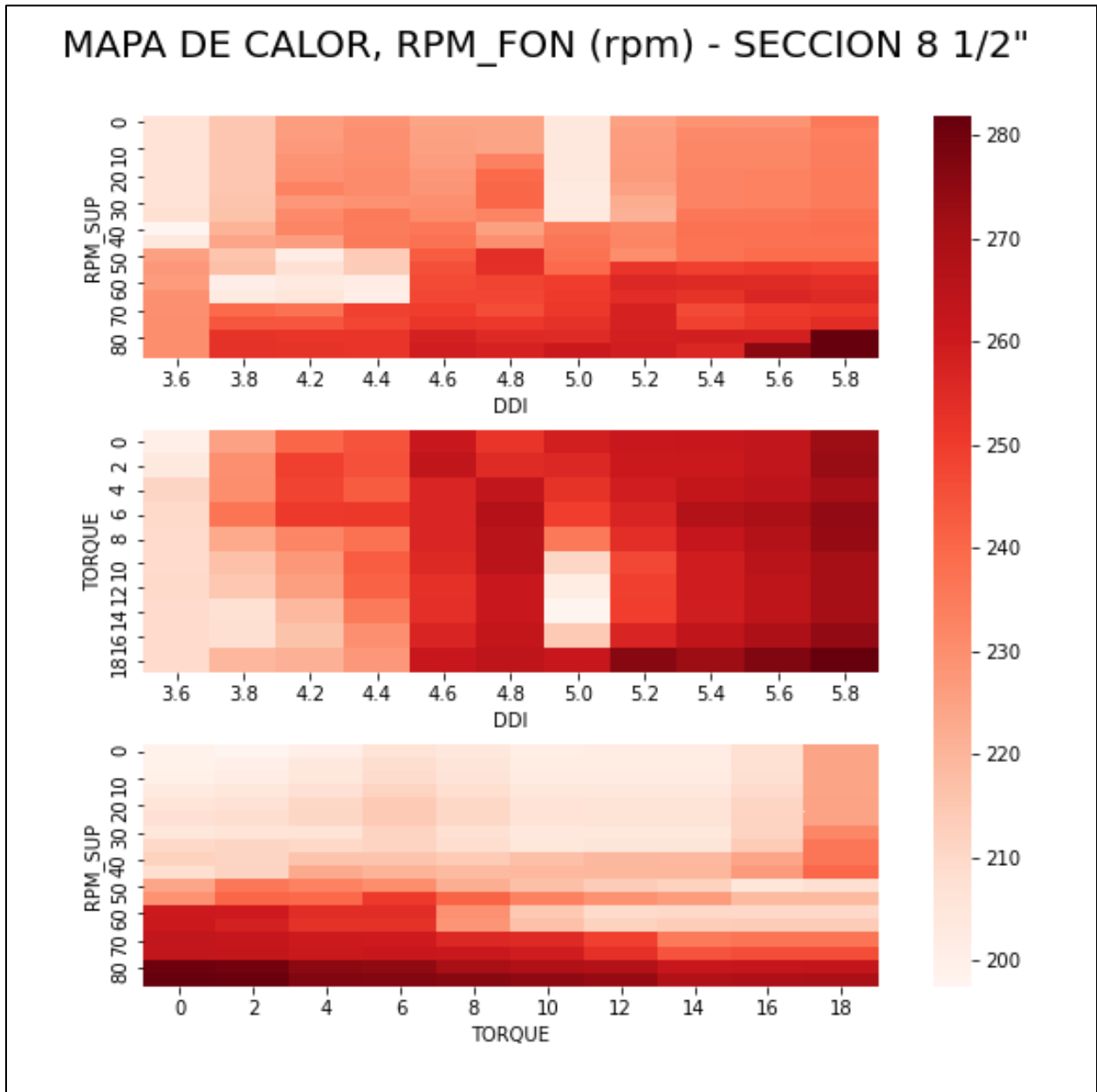
Mapa de calor para DDI, sección 12 1/4"



Nota. Mapa de calor para DDI en la sección 12 1/4" con un rango operativo de 2.5 a 3.1 adimensional.

Figura 48.

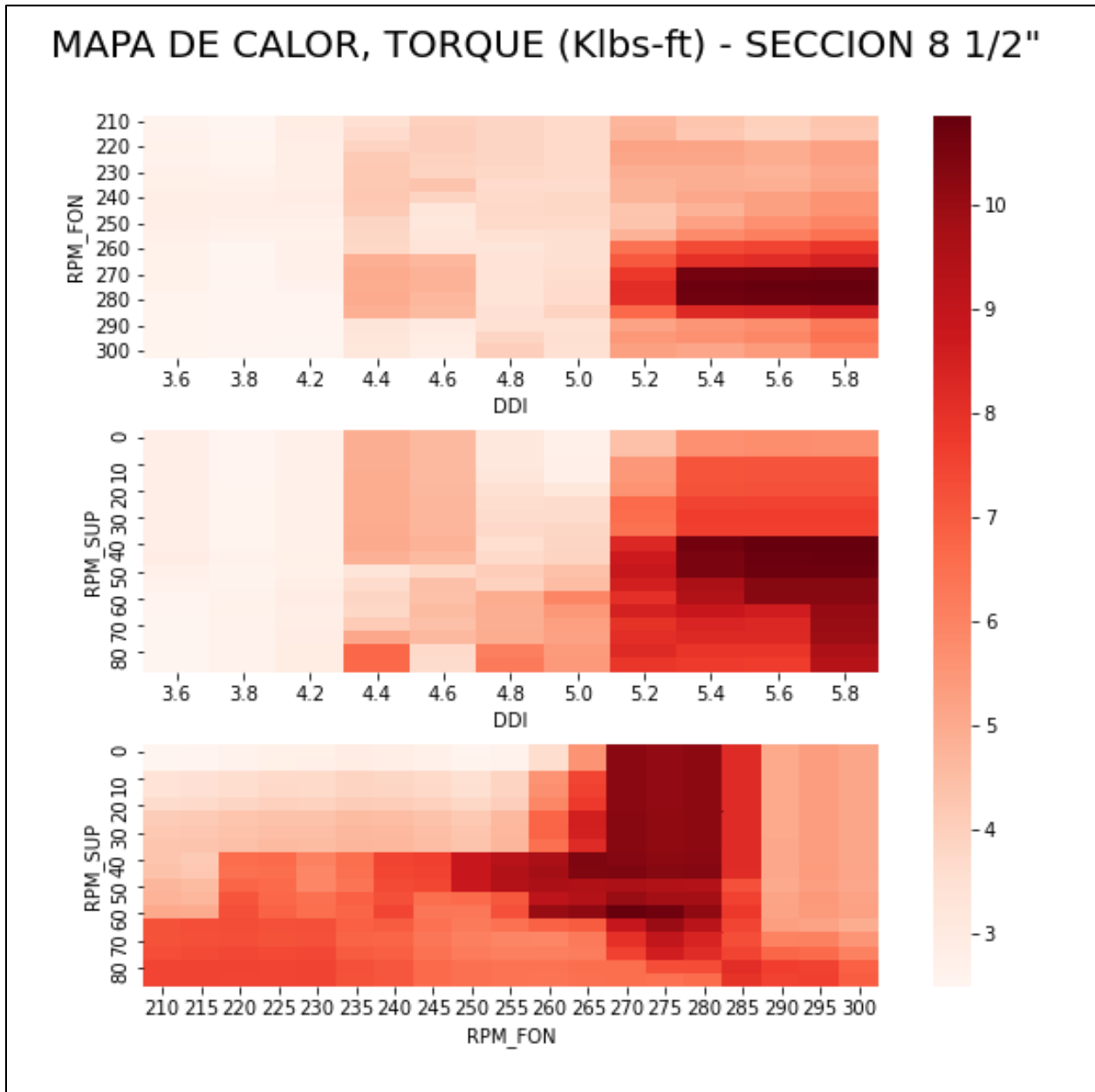
Mapa de calor para RPM_FON, sección 8 1/2"



Nota. Mapa de calor para RPM_FON en la sección 8 1/2" con un rango operativo de 200 a 280 rpm

Figura 49.

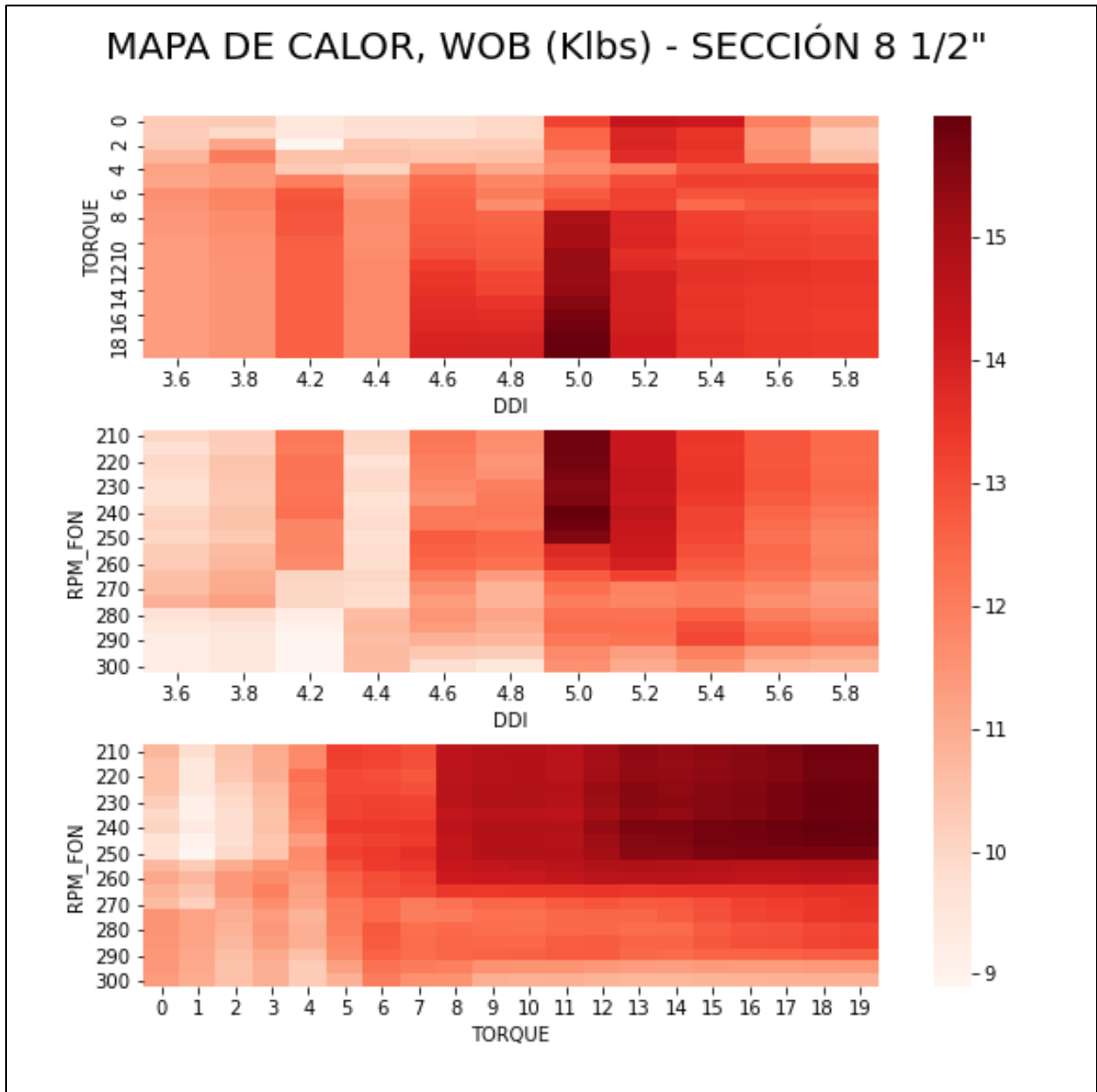
Mapa de calor para TORQUE, sección 8 1/2"



Nota. Mapa de calor para TORQUE en la sección 8 1/2" con un rango operativo de 3 a 10 Klbs-ft

Figura 50.

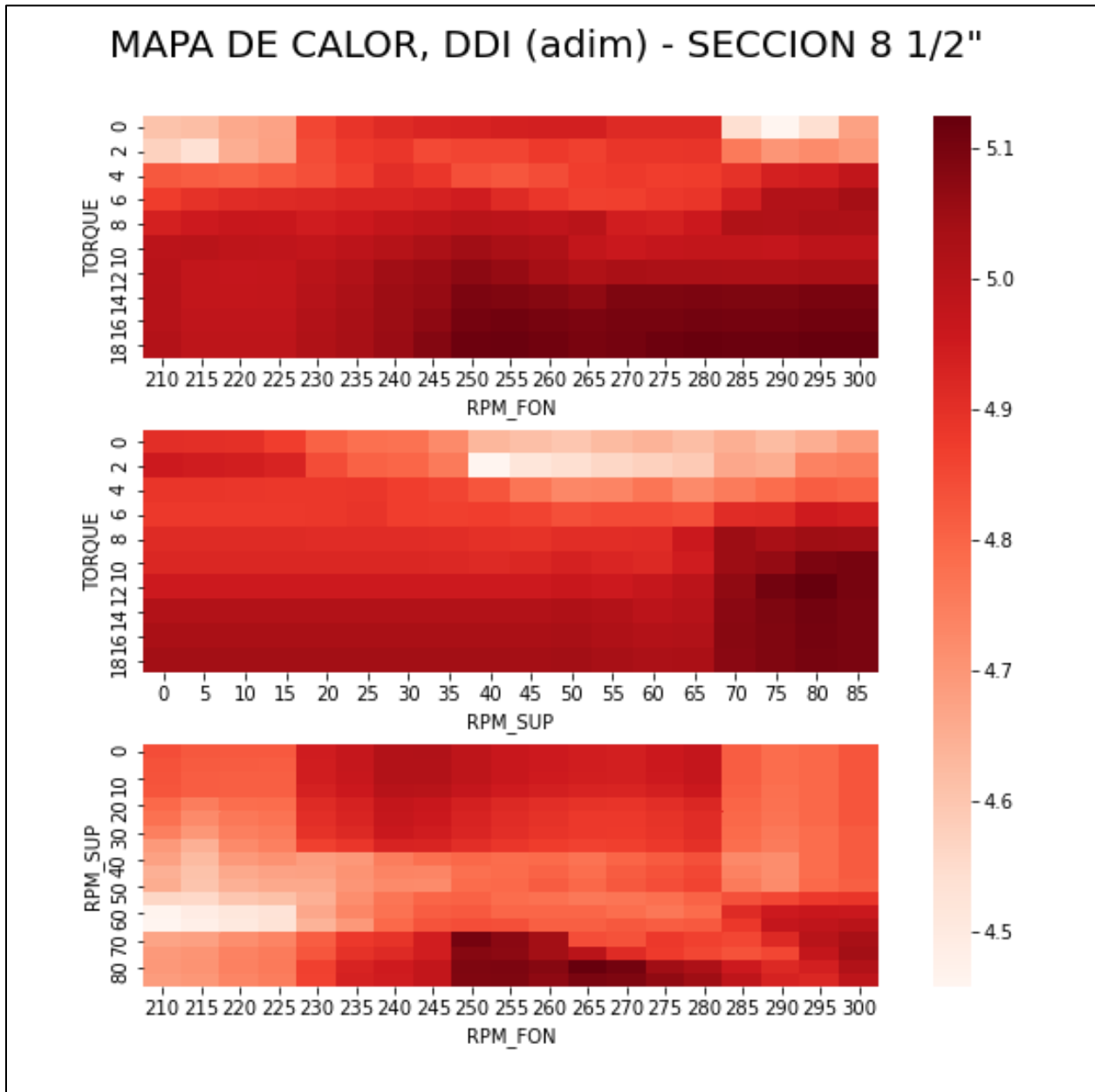
Mapa de calor para WOB, sección 8 1/2"



Nota. Mapa de calor para WOB en la sección 8 1/2" con un rango operativo de 9 a 15 Klbs

Figura 51.

Mapa de calor para DDI, sección 8 1/2"



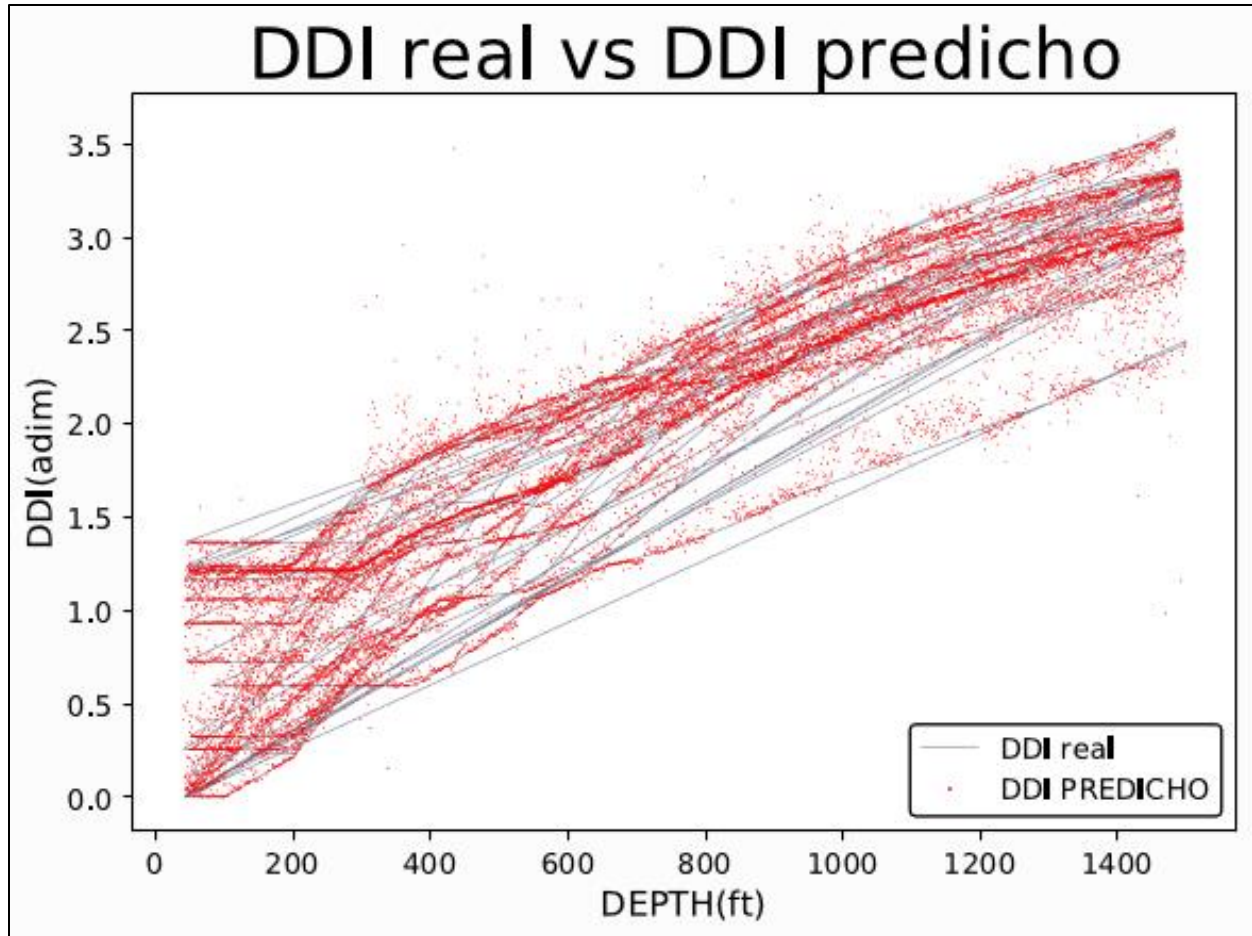
Nota. Mapa de calor para DDI en la sección 8 1/2" con un rango operativo de 4.5 a 5.1 adimensional

ANEXO 2.

VALIDACIONES DE MODELOS PREDICTIVOS EN LAS SECCIONES 12 ¼” Y 8 ½”

Figura 52.

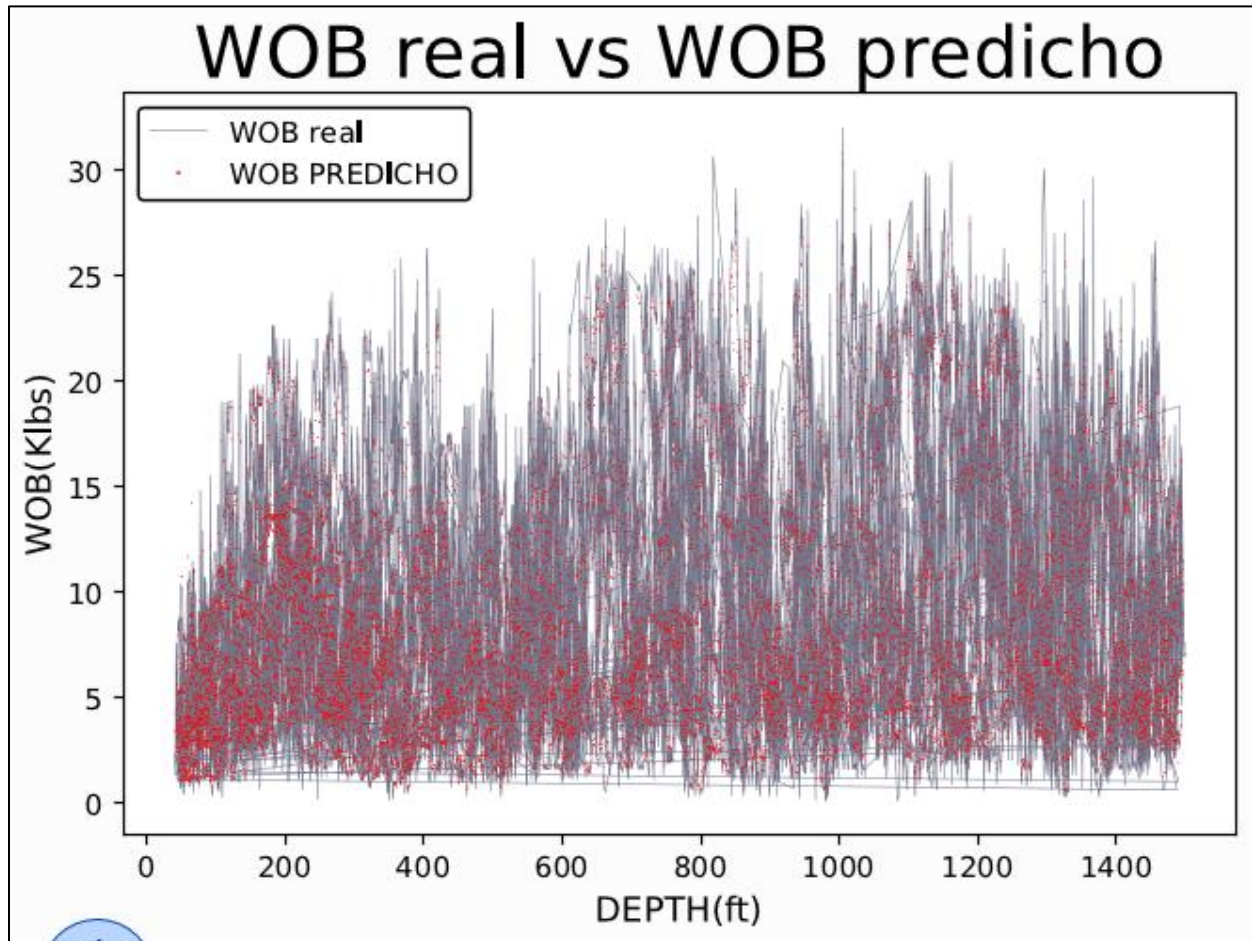
Validación para el modelo predictivo de DDI, sección 12 ¼”



Nota. Validación para DDI en la sección 12 ¼” con una exactitud del 92.91%

Figura 53.

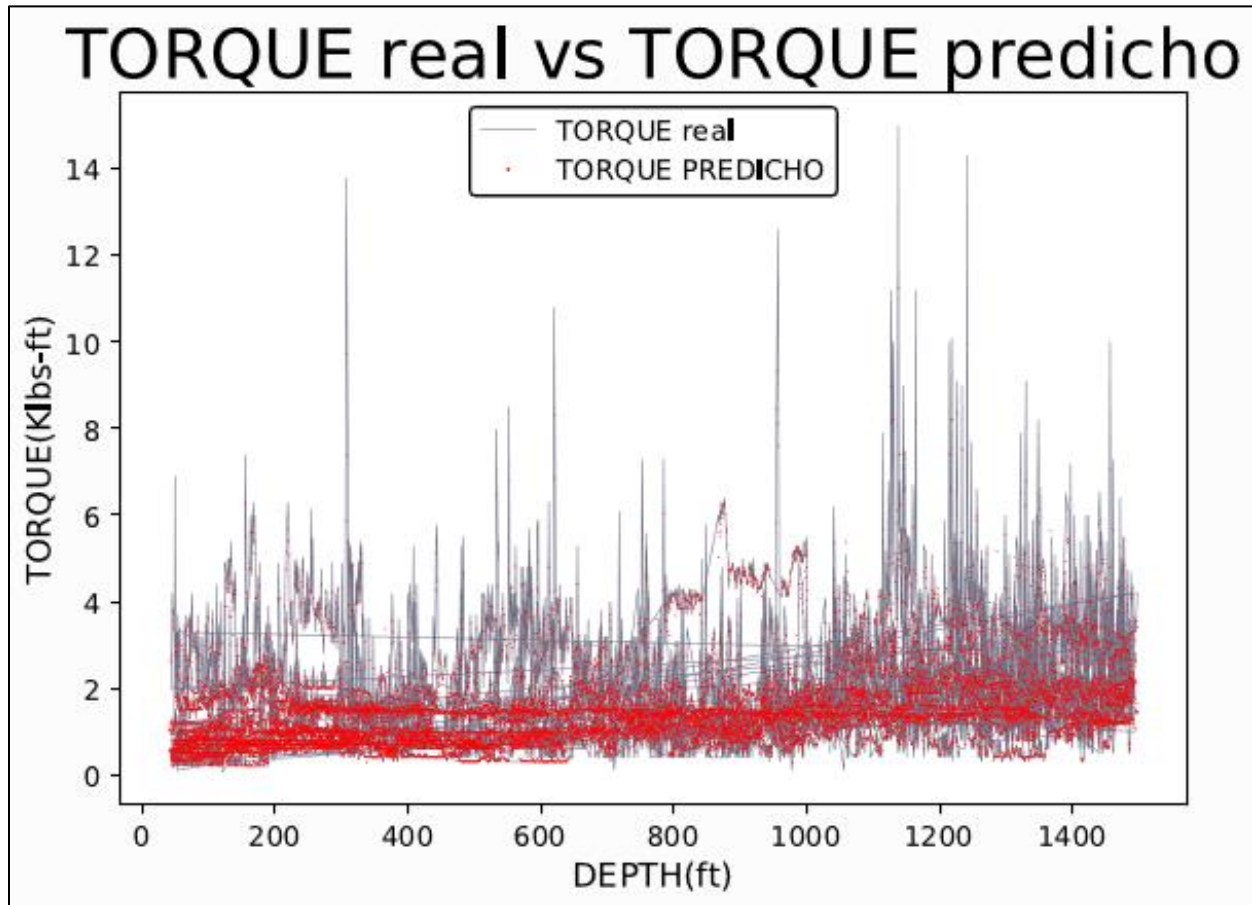
Validación para el modelo predictivo de WOB, sección 12 ¼”



Nota. Validación para WOB en la sección 12 ¼” con una exactitud del 72.18%

Figura 54.

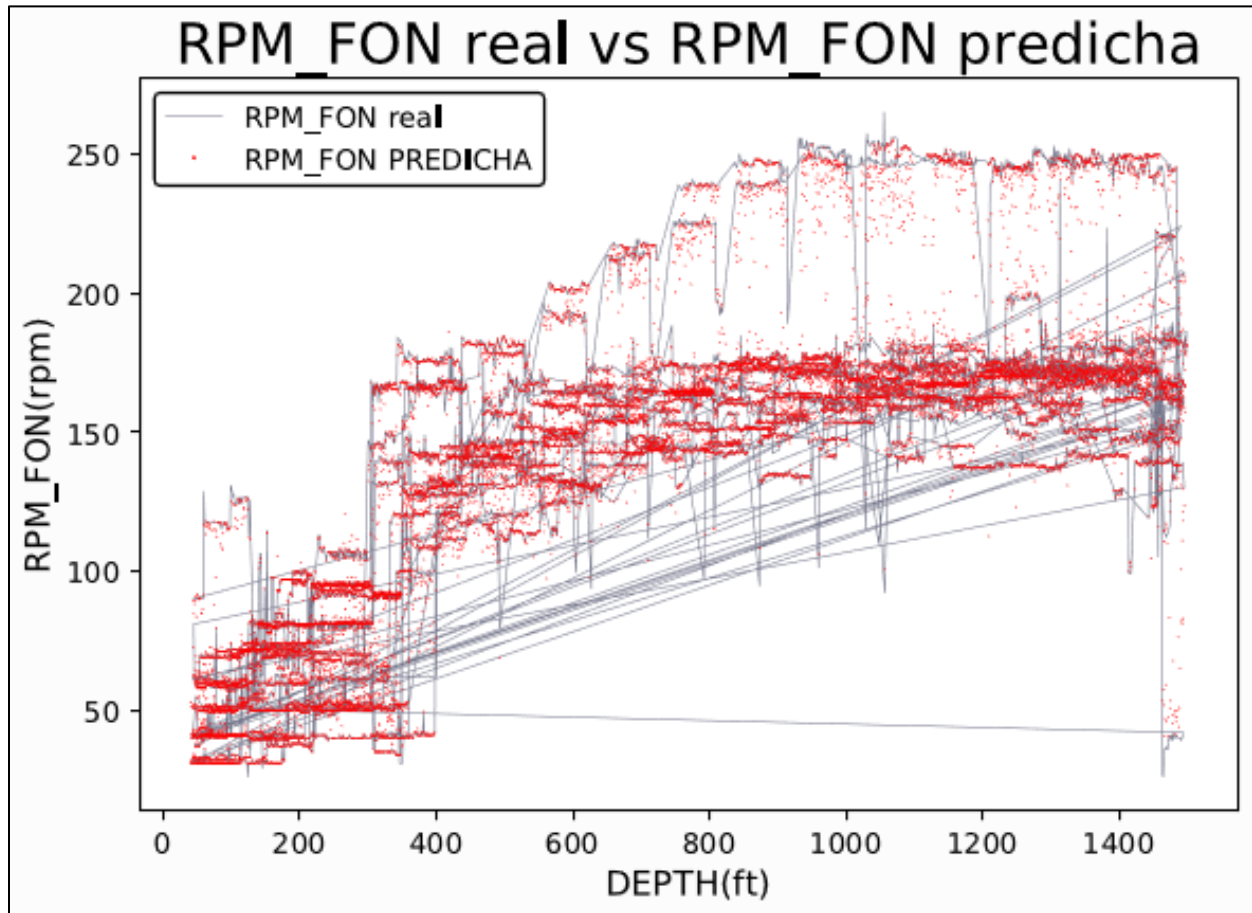
Validación para el modelo predictivo de TORQUE, sección 12 ¼”



Nota. Validación para TORQUE en la sección 12 ¼” con una exactitud del 74.04%

Figura 55.

Validación para el modelo predictivo de RPM_FON, sección 12 ¼”



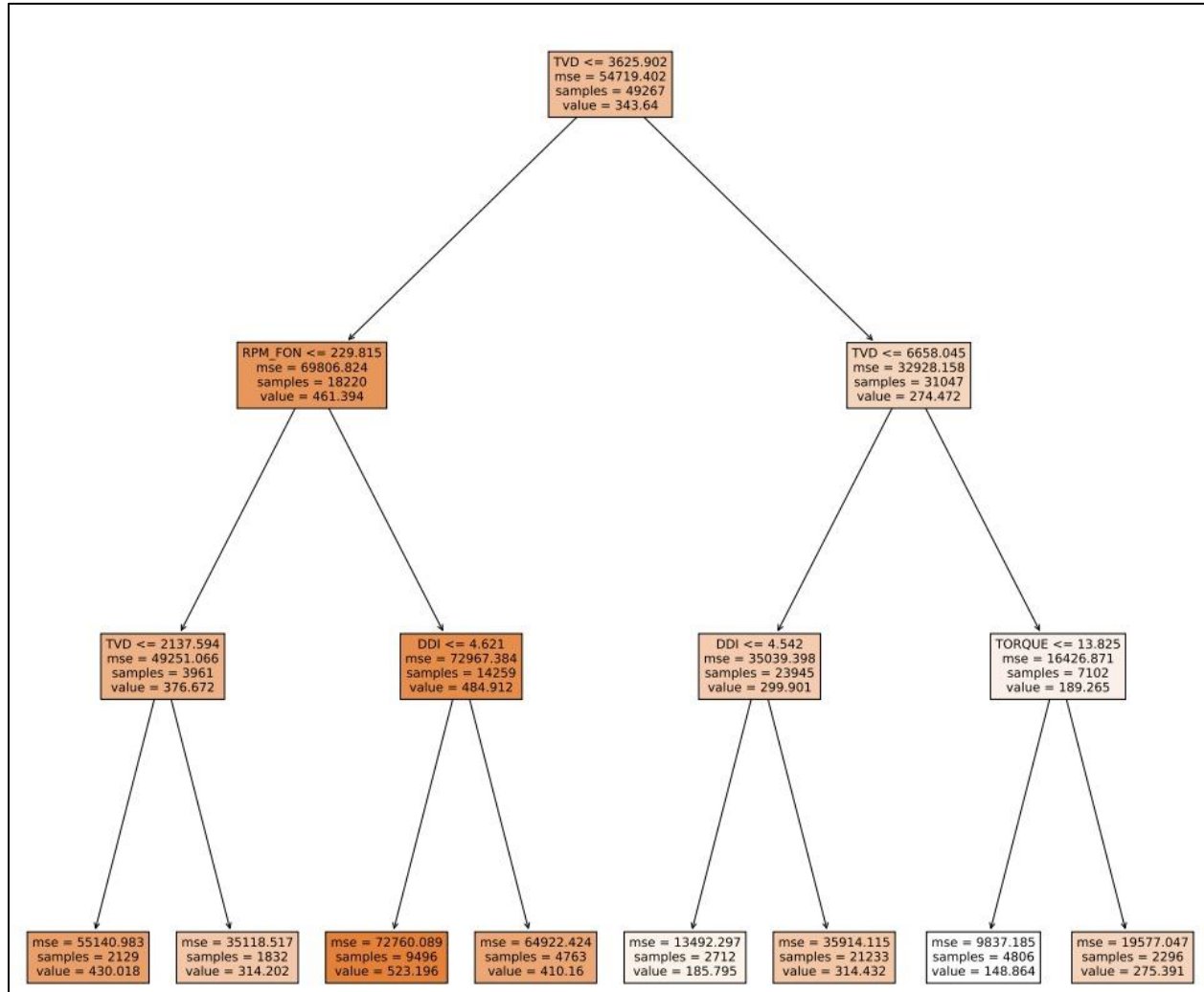
Nota. Validación para RPM_FON en la sección 12 ¼” con una exactitud del 95.41%

ANEXO 3.

ARBOLES DE DECISIÓN PARA LA SECCIÓN 12 1/4"

Figura 56.

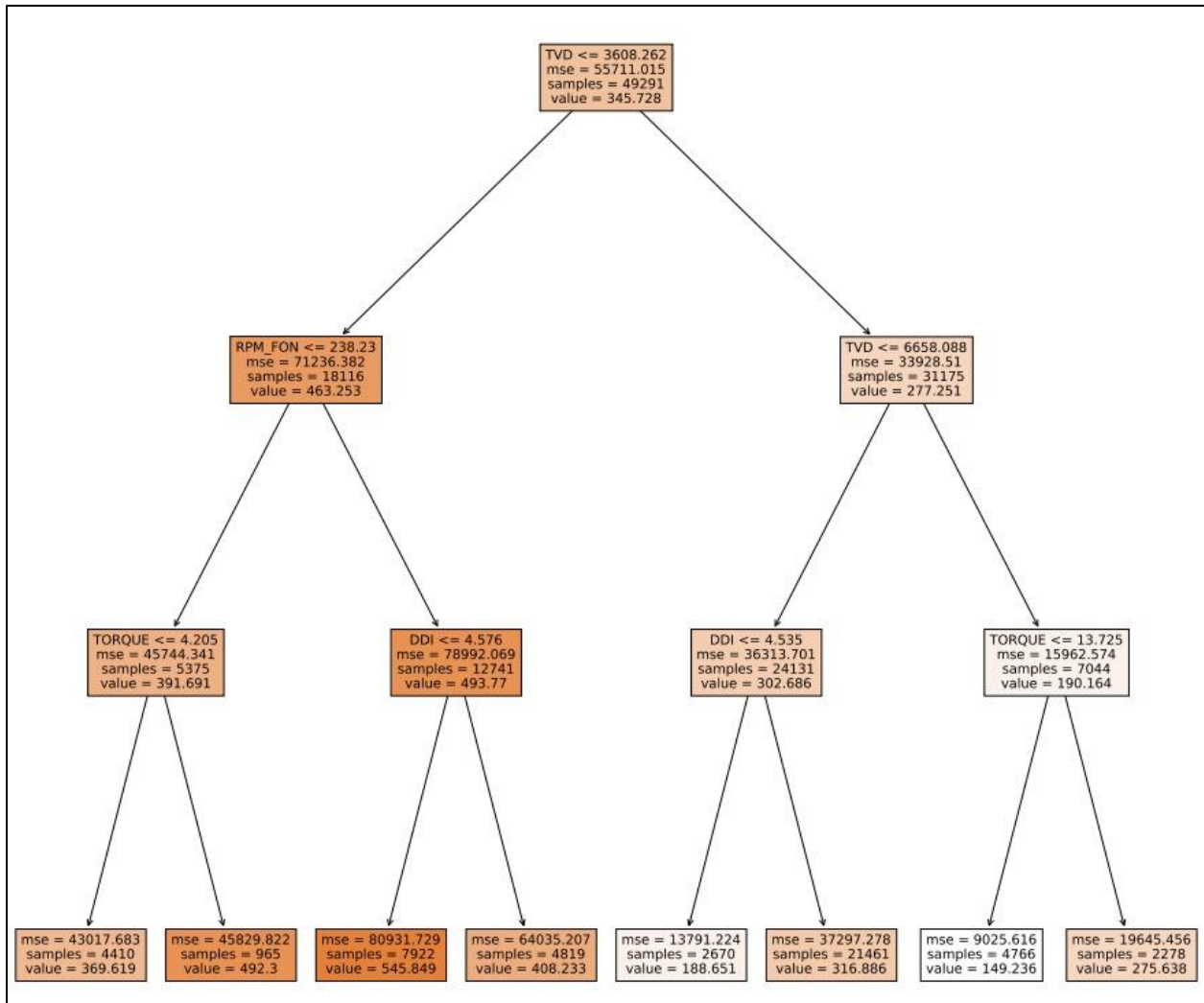
Árbol de decisión ROP, sección 12 1/4"



Nota. Árbol de decisión para la ROP hasta el 3 nivel.

Figura 57.

Árbol de decisión TORQUE, sección 12 ¼”.



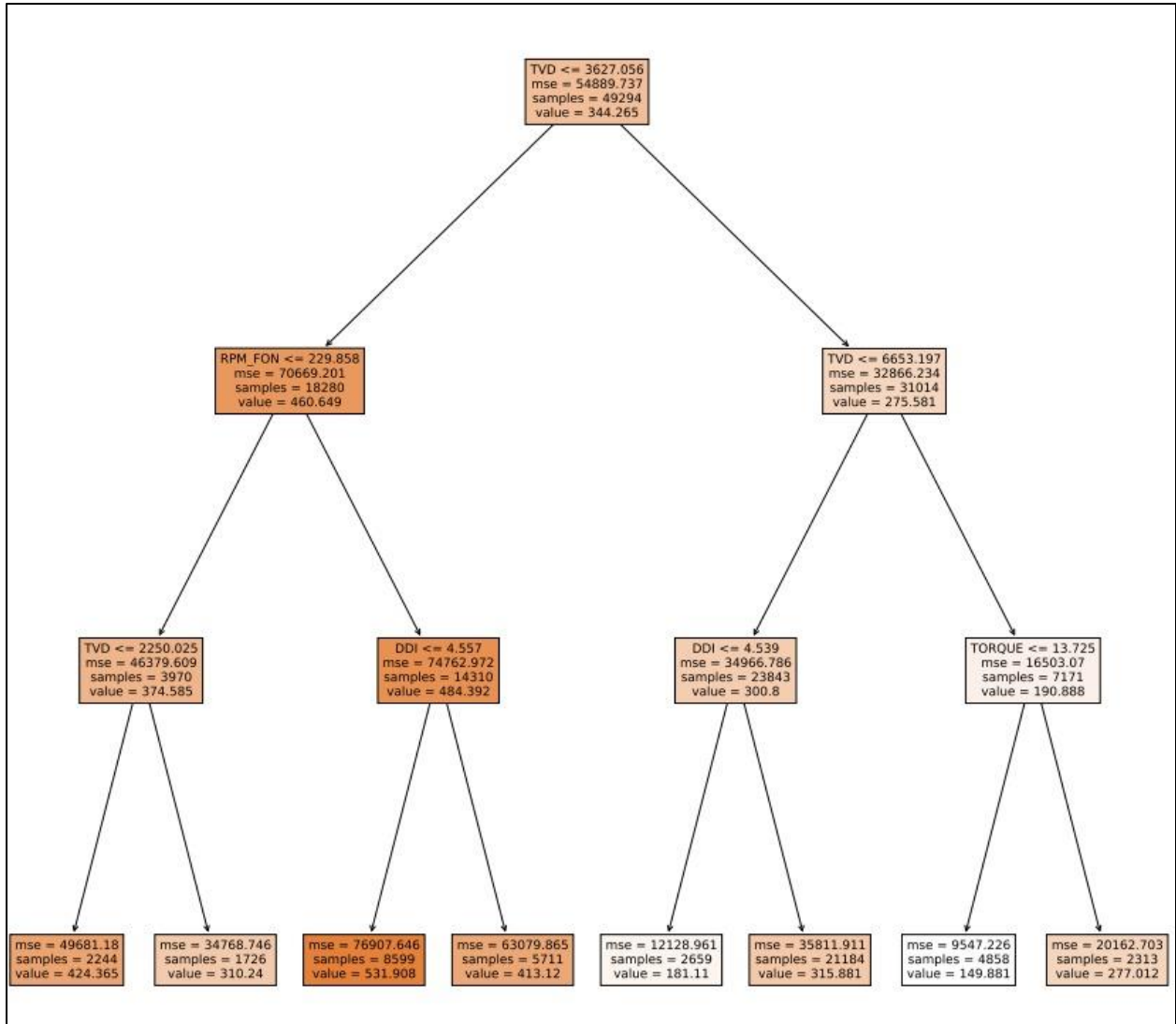
Nota. Árbol de decisión para el TORQUE hasta el 3 nivel.

ANEXO 15.

ÁRBOL DE DECISIÓN WOB, SECCIÓN 12 1/4"

Figura 58.

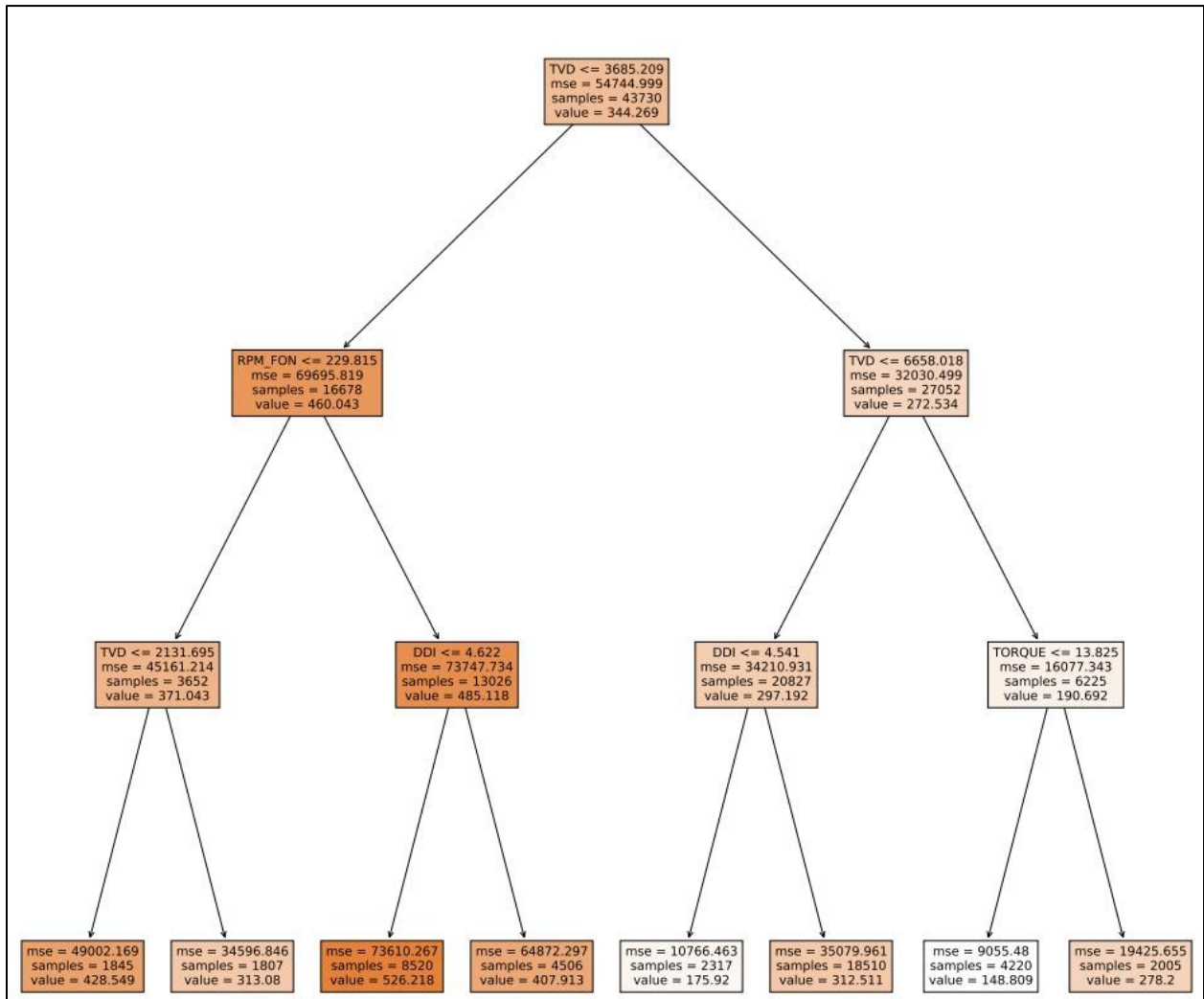
Árbol de decisión WOB, sección 12 1/4".



Nota. Árbol de decisión para el WOB hasta el 3 nivel.

Figura 59.

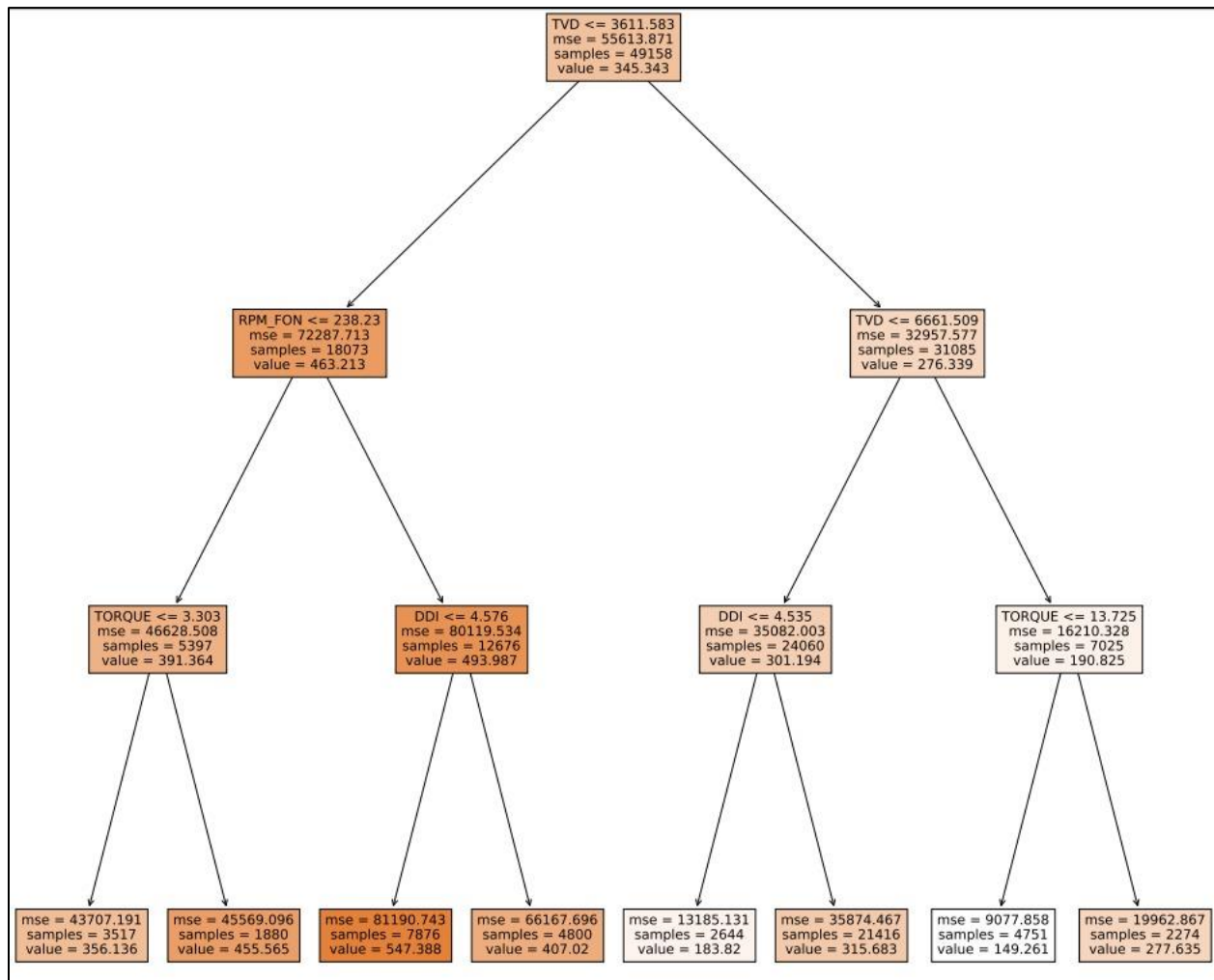
Árbol de decisión para la RPM_FON , sección 12 1/4".



Nota. Árbol de decisión para la RPM_FON hasta el 3 nivel.

Figura 60.

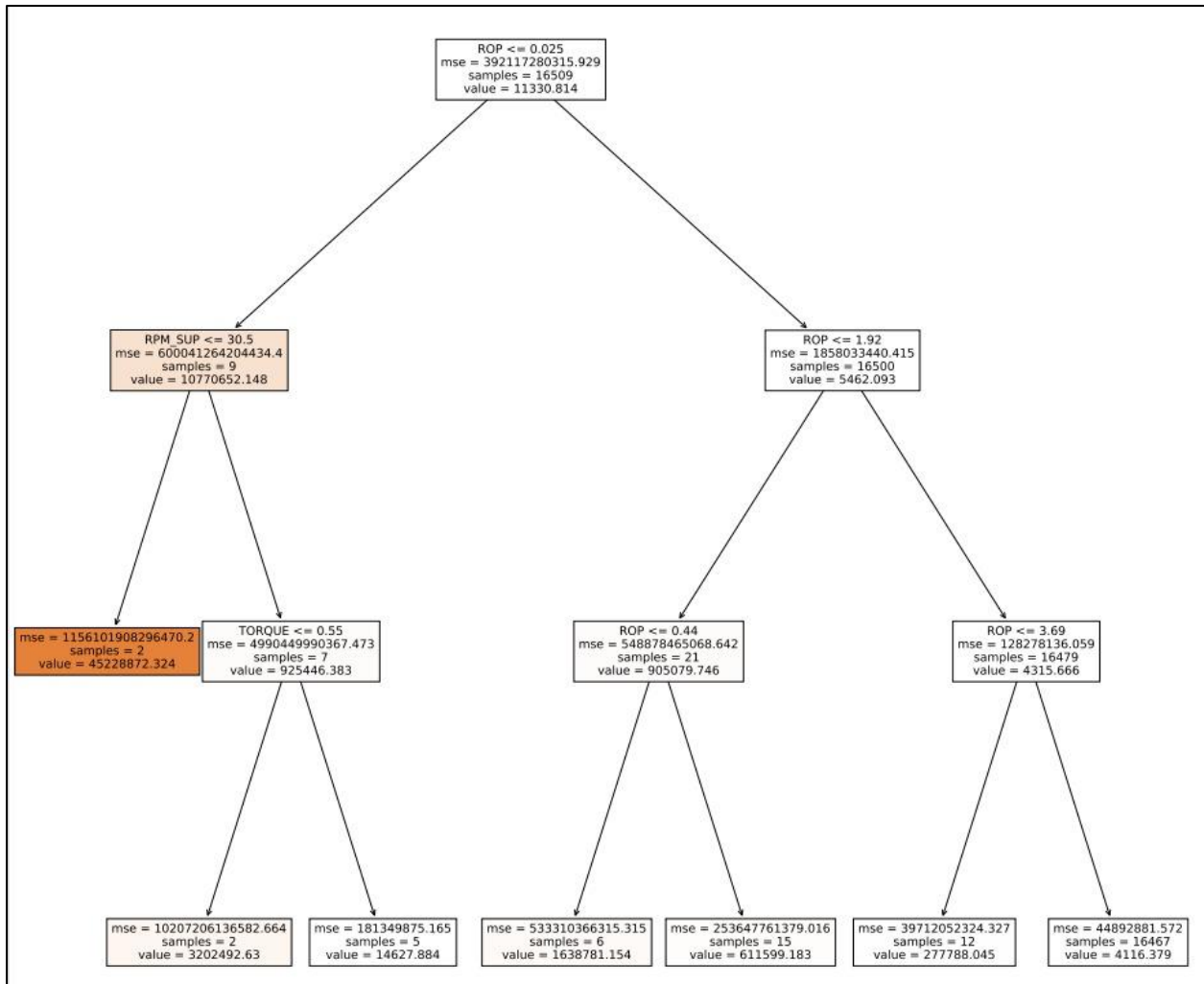
Árbol de decisión DDI, sección 12 1/4".



Nota. Árbol de decisión para el DDI hasta el 3 nivel.

Figura 61.

Árbol de decisión ROP, sección 12 1/4"



Nota. Árbol de decisión para la ROP hasta el 3 nivel.

ANEXO 4.

RECOMENDACIONES

Analizar más variables involucradas en actividades de perforación para conformar un dataset más robusto y por lo tanto modelos predictivos más precisos al momento de obtener nuevos datos para las variables de entrada. Variables como espiralamiento, vibración de la sarta y propiedades reológicas del lodo son muy buenas candidatas a complementar un próximo dataset.

Discriminar las secciones de estudio por formación geológica, con el fin de tener claro qué valores son los apropiados para perforar dichas formaciones y no ocasionar rompimiento, hinchamiento y otras consecuencias de una mala perforación.

Realizar un análisis exploratorio de datos (AED) incluyendo más variables estadísticas, como pueden ser: curtosis, simetría, coeficiente de variación y error estándar encaminado a tener un mejor estudio y claridad del comportamiento estadístico de los parámetros de perforación analizados.

Crear modelos predictivos con base en otros algoritmos de regresión tales como vecinos más cercanos, máquinas de vectores de soporte y Naive Bayes, entre otros; y evaluar su desempeño.

Continuo a la creación de modelos predictivos, realizar una corrida para verificar cuál es el error más eficiente (mse o MAPE) para su respectivo modelo.

Generar listas de parámetros más largas al momento de construir los mapas de calor, esto se hace para tener más predicciones de la variable objetivo y obtener rangos óptimos más precisos.

Contar con información más detallada sobre costos para realizar un análisis financiero más completo y consistente por pozo en cada año de actividad de perforación.