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

NATHALIA TABARES RODRIGUEZ
DANIEL TOBAR CASTILLA

Proyecto integral de grado para optar al 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 Jurado

Ing. Ing.

_____

Jurado

Ing.

Bogotá D.C., enero 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. MARIA CLAUDIA APONTE GONZÁLEZ

Vicerrector Administrativo y Financiero
Dr. RICARDO ALFONSO PEÑARANDO

Secretaria General
Dra. ALEXANDRA MEJÍA GUZMAN

Decano Facultad de Ingeniería
Ing. JULIO CÉSAR FUENTES ARISMENDI

Director Programa de Ingeniería de Petróleos
Ing. JUAN CARLOS RODRÍGUEZ ESPARZA

DEDICATORIA

Dedico este trabajo a mis padres, quienes siempre han estado a mi lado brindándome su apoyo y consejos. A mi hermano Sebastián por su compañía en todo momento. A mi abuela, que desde el cielo me acompaña en cada paso. A mi amigo Daniel por ser mi compañero en la realización de este proyecto.

Nathalia Tabares Rodríguez

DEDICATORIA

Dedico este trabajo a mis papas y a mi abuela, quienes son lo más importante para mí y estuvieron a mi lado en todo este proceso. A mi hermano Sergio por el apoyo incondicional en todo momento. A Paola por sus consejos y recomendaciones. A mi apreciada amiga Nathalia por vivir este proceso y experiencia conmigo.

Daniel Tobar Castilla

AGRADECIMIENTOS

Agradecemos de manera muy especial a los ingenieros Ricardo Bustos, Wilmar Osorio y Ricardo García por su confianza, orientación, dedicación y apoyo incondicional en cada etapa del proyecto.

A La UNIVERSIDAD DE AMÉRICA por formarnos como profesionales y ciudadanos íntegros, por inculcarnos la excelencia en las diferentes áreas de conocimiento de nuestras carreras.

Al docente Sebastián Gómez por su apoyo incondicional y asesoría durante el desarrollo del proyecto.

A 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.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

LISTA DE ABREVIATURAS

ROP Tasa de penetración

RPM Revoluciones por minuto

WOB Peso sobre la broca

Q Caudal

TQ Torque

D Diámetro de la broca

ID Diámetro Interno del Pozo

OD Diámetro Externo del Pozo

PV Profundidad Vertical

AZ Azimut

MD Measure depth (Profundidad Medida)

Inc Inclinación

In Pulgadas

Ft pies

Psi Pound per square inch (Libra por pulgada cuadrada)

Klb-f Kilo libras fuerza

Fph Feet per hour (Pies por hora)

BHA Bottom Hole Assembly

GPM Galones Por Minuto

TVD True Vertical Depth (Profundidad Vertical Real)

NPT Non Productive Time (Tiempos No productivos)

MSE Energía Mecánica Específica

ML Machine Learning

RF Random Forest

EDA Análisis Exploratorio de Datos

Py Lenguaje de Programación Python

IA Inteligencia Artificial

RESUMEN

TITULO: Implementación de un modelo predictivo de machine learning para la estimación de los parámetros óptimos de la ROP y la MSE en la sección 8 ½'' y 12 ¼'' para los pozos perforados con motor de fondo en el Campo Yariguí – Cantagallo durante el 2019.

DESCRIPCIÓN: La implementación de un modelo predictivo de machine learning para la estimación de los parámetros óptimos de perforación surge por la necesidad de la industria de migrar hacia la ciencia de datos buscando optimizar procesos. A través de este proyecto de investigación se generó una base de datos correspondiente a los pozos perforados con motor de fondo durante el 2019 en el campo en mención, la cual fue sometida a un análisis exploratorio de datos (EDA). Seguido a esto, se realizó división de la misma para la estandarización y prueba del modelo predictivo. Una vez es realizada dicha división se implementó un algoritmo de aprendizaje automático supervisado como lo es Random Forest Regressor, teniendo como variables de entrada las revoluciones por minuto (RPM) de superficie y de fondo, el peso sobre la broca (WOB), el caudal (Q), el torque (TQ) y la información correspondiente a los topes de las formaciones geológicas perforadas, y se obtuvo como variables de salida la tasa de penetración (ROP) y la energía mecánica específica (MSE). El modelo se validó y verificó mediante la comparación de los valores predichos por el mismo contra los reales, y fue evaluado mediante el error absoluto porcentual medio, obteniendo como resultado precisiones del 72.57% y 71.51% para la ROP y 71.63% y 71.10% para la MSE en las secciones 8 ½'' y 12 ¼'' respectivamente. Por último, fueron establecidos los valores óptimos de los parámetros en estudio por formación geológica mediante la elaboración de mapas de parámetros basados en el concepto de mapas de calor.

PALABRAS CLAVE: Machine learning, random forest, tasa penetración (ROP), energía mecánica específica (MSE), caudal (Q), peso sobre broca (WOB), mapas calor.

INTRODUCCIÓN

Mundialmente la industria petrolera se encuentra en la búsqueda e investigación constante de nuevas tecnologías que permitan operaciones precisas, económicas y veloces, con el propósito principal de lograr la ejecución de procedimientos con bajo costo operacional y en el menor tiempo posible. La incursión de nuevas tecnologías en el mercado como lo son las técnicas de aprendizaje automático, mejor conocidas como machine learning, que según Javier Calvo se definen como "un conjunto de métodos capaces de detectar automáticamente patrones en los datos por medio de algoritmos, con el fin de usarlos en la predicción de datos futuros. Obteniéndose así sistemas de aprendizaje que mejoran manera autónoma por medio de la experiencia" [1], las cuales se convierten en una oportunidad invaluable y prometedora para el sector.

La inteligencia artificial cada vez tiene una mayor aplicación en todas las áreas de la industria. En el área de perforación, se han realizado diversos estudios e investigaciones dedicados a disminuir la incertidumbre de parámetros de perforación, implementando modelos de aprendizaje automático. Barbosa, Nascimiento y Mathiase 2019 describen diferentes modelos, estrategias y algoritmos de aprendizaje automático (machine learning) y cómo estos pueden servir para la predicción de las tasas de penetración durante la perforación de pozos, el estudio concluye y afirma la gran importancia que tienen las variables de entrada para una óptima y eficiente elaboración del modelo, y como las diferentes técnicas de aprendizaje automático superan de manera contunde a los procesos convencionales obteniendo variables de perforación con una mayor eficiencia [2]. Otro ejemplo es el estudio de C.Hedge y H. Daigle (2017) donde comparan tres modelos basados en la física con otros basados en datos estadísticos , para crear algoritmos de aprendizaje automáticos utilizando variables de superficie; el estudio concluyó que el mejor algoritmo estadístico es random forest, el cual predice de manera óptima el 84% de las veces la ROP necesaria para perforar nuevos pozos con una exactitud y optimización considerable, en comparación con otros modelos como el de regresión lineal con un 12% de exactitud y el de Bringham con un 46% respectivamente [3].

Actualmente, Ecopetrol S.A se encuentra comprometida en optimizar las actividades y procesos de la industria petrolera con el objetivo principal de reducir tiempos de operación y costos asociados, El presente estudio se realizó al campo Yariguí-Cantagallo, el cual cuenta

con más de cien (100) pozos perforados hasta el momento. Se encuentra ubicado en el Valle Medio del Magdalena cuya alta complejidad geológica dificulta las actividades de perforación horizontal y desviada [4]. Lo anterior ha generado que, a lo largo de la vida productiva del campo, muchos de los parámetros de perforación como lo son el peso sobre la broca (WOB), las revoluciones por minuto (RPM), el caudal (Q), y la tasa de penetración (ROP), tengan una errónea estimación. Al tener las causas del problema bien identificadas y la información disponible para su análisis, este trabajo de grado es el primer paso para utilizar toda la data acumulada para mitigar la incertidumbre de las variables anteriormente mencionadas, mediante modelos de predictibilidad.

Específicamente la alta incertidumbre en la estimación de la MSE en función de la ROP genera diferentes problemas asociados a la perforación como, la baja efectividad al momento de la perforación, el bajo desempeño de los equipos, el embotamiento y recalentamiento de la broca y las pegas de tubería. los cuales se pueden evidenciar en este proyecto de investigación en los pozos perforados con motor de fondo en el campo Yariguí-Cantagallo durante el 2019. La Compañía Operadora del Campo ha reportado igualmente incremento de los tiempos de operación y costos, razón por la cual la empresa debe invertir un mayor capital de lo esperado en la perforación de los pozos generando retrasos en la operación.

En la búsqueda a dar solución a la problemática planteada, este proyecto tiene como objetivo general implementar un modelo predictivo mediante la metodología machine learning para la estimación de los parámetros óptimos de la ROP y la MSE en la sección 8 ½'' y 12 ¼'' para los pozos perforados con motor de fondo en el Campo Yariguí-Cantagallo durante el año 2019. La realización de este proyecto se llevará a cabo, tomando como base información histórica de parámetros de perforación como el caudal, el peso sobre la broca, las revoluciones por minuto y las formaciones geológicas en la sección 8 ½'' para los pozos anteriormente mencionados y se establecerán los rangos óptimos de las variables involucradas para estimar los valores de la ROP y MSE durante la actividad de perforación y así optimizar los tiempos y costos de las operaciones asociadas.

Así entonces, la presente investigación se encuentra delimitada por los siguientes 4 objetivos específicos que se presentan a continuación:

Crear un dataset a partir del análisis de criterios estadísticos con la información de la sección 8 ½'' de los pozos perforados con motor de fondo en el Campo Yariguí-Cantagallo durante el 2019.

Estandarizar un modelo predictivo a partir de la implementación del algoritmo Random Forest de la librería Scikit Learn en Python, usando una muestra representativa del 70% del dataset de perforación determinada en el objetivo anterior.

Implementar el modelo predictivo usando el 30% de la data restante que no se usó para la estandarización del modelo.

Establecer los valores de ROP y MSE óptimos durante las operaciones de perforación, mediante la elaboraci
ón de mapas de calor y usando los resultados obtenidos de la implementación del modelo predictivo.

Mediante la implementación de un modelo predictivo de machine learning, tomando como base información histórica de parámetros de perforación: caudal, peso sobre la broca, revoluciones por minuto, formaciones geológicas y presión de fondo y de superficie en la sección 8 ½'' para los pozos perforados con motor de fondo en el campo Yarigui-Cantagallo durante 2019, se establecerán los rangos óptimos de las variables involucradas para estimar los valores de la ROP y MSE durante la actividad de perforación y así optimizar los tiempos y costos de las operaciones asociadas.

1. MARCO TEÓRICO

Esta sección hace referencia a los aspectos teóricos necesarios para el desarrollo y entendimiento de esta investigación. A continuación, se presentarán los conceptos básicos de: técnicas de aprendizaje automático (machine learning), consideraciones estadísticas para el manejo de la información y generalidades de las operaciones de perforación de pozos de hidrocarburos.

1.1 Aprendizaje automático

Las técnicas de aprendizaje automático (*machine learning*), se definen como "Un conjunto de métodos capaces de detectar automáticamente patrones en un conjunto de datos por medio de algoritmos, con el fin de usarlos en la predicción de datos futuros. El fin último del desarrollo de modelos predictivos es obtener sistemas de aprendizaje que mejoran de manera autónoma por medio de la experiencia". [1, 5, 6]

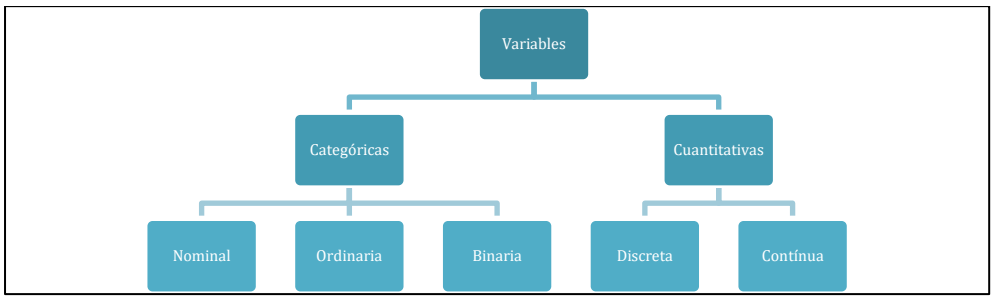
La implementación de un modelo de aprendizaje automático se inicia con la identificación de las variables (conjunto de datos) a trabajar. Posteriormente se determina el modelo y el método de aprendizaje basándose en el tipo de datos y en el objetivo final del mismo (clasificación o predicción) [7]. Estos últimos términos en esta investigación son asociados a los algoritmos predictivos de árboles de decisión (Bagging trees) y bosques aleatorios (Random Forest). Previo a la implementación del modelo seleccionado es necesario realizar una caracterización estadística de las variables con el fin de establecer su comportamiento, correlación, valores atípicos, entre otros [8]. La fiabilidad de los resultados de predicción puede determinarse mediante métricas de desempeño como lo son el error cuadrático medio, error absoluto medio, etc. [9]. A continuación, se hace una explicación detallada de los anteriores conceptos:

Las variables en un modelo de machine learning se definen como magnitudes que pueden almacenar datos utilizados en un algoritmo de programación [10]. Estas se clasifican en dos tipos:

- Variables cuantitativas: Son un tipo de variable que puede expresarse o medirse a través de números. Estas pueden ser: discretas, que utilizan valores enteros y no finitos, por ejemplo, número de libros vendidos en un mes; o continuas, las cuales se caracterizan por tener valores decimales, como el peso de una persona [10].

- Variables categóricas: Son aquellas características o cualidades que no pueden ser determinadas con números, sino que son clasificadas con palabras. Dentro de ellas se destacan las nominales, las cuales no siguen ningún orden específico, por ejemplo, el color de cabello; las ordinales, caracterizadas por seguir una jerarquía, como los rangos militares, y las binarias, que permiten obtener solo dos características, tales como aprobado o reprobado. [10]

Figura 1.
Tipos de variables



Nota: Tipos de variables: cualitativas las cuales pueden ser nominales u ordinarias y cuantitativas, que se pueden subdividir en discretas, continuas y binarias. Tomado de: Enciclopedia Económica. Variable estadística (2018). https://enciclopediaeconomica.com/variable-estadistica/

1.1.1 Modelos de aprendizaje

Los modelos de aprendizaje están basados en los tipos de variables suministradas y el objetivo final de su implementación, estos permiten el uso de algoritmos para la interpretación de los datos y su comportamiento [5]. Existen tres principales modelos que se pueden ejecutar según la información suministrada y la finalidad del mismo, como se muestra en la siguiente imagen

Figura 2.

Modelos de aprendizaje

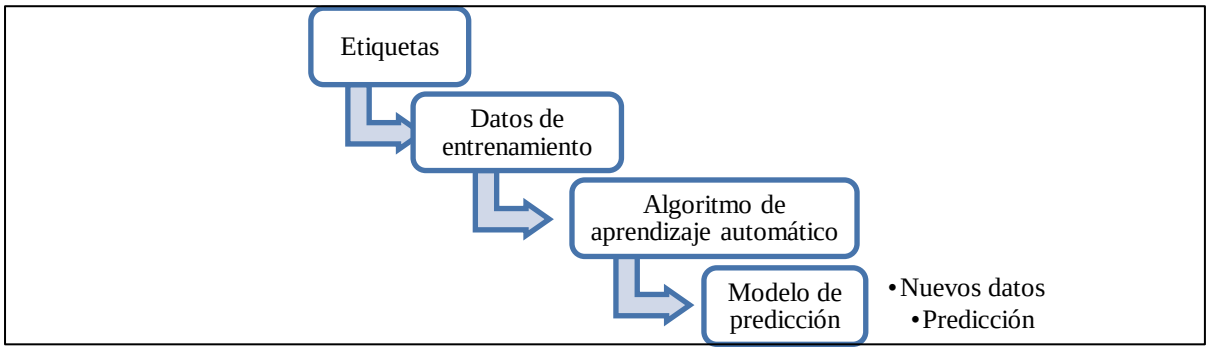
Aprendizaje supervisado	Aprendizaje no supervisado	Aprendizaje reforzado
• Datos etiquetados • Feedback directo • Predicción de resultados/futuro	• Sin etiquetas • Sin feedback • Encontrar estructuras ocultas en los datos	• Proceso de decisión • Interacción con el entorno • Aprender series de acciones

Nota: Modelos de aprendizaje: supervisado, no supervisado y reforzado, sus principales características y objetivos. Se adaptó del libro Python Machine Learning por Raschka S.

El primer modelo es el aprendizaje supervisado, este término se refiere a un conjunto de muestras donde las señales de salida deseadas (etiquetas) ya se conocen. Su objetivo principal es aprender de un modelo a partir de datos de entrenamiento etiquetados, permitiendo realizar predicciones sobre datos futuros. [5].

Este tipo de modelo es muy útil cuando la variable a predecir hace parte del conjunto de datos implementado en el modelo [7]. Un ejemplo clásico de este tipo de aprendizaje es la predicción del tipo de flor de acuerdo a las características de sus pétalos [11].

Figura 3.
Aprendizaje supervisado

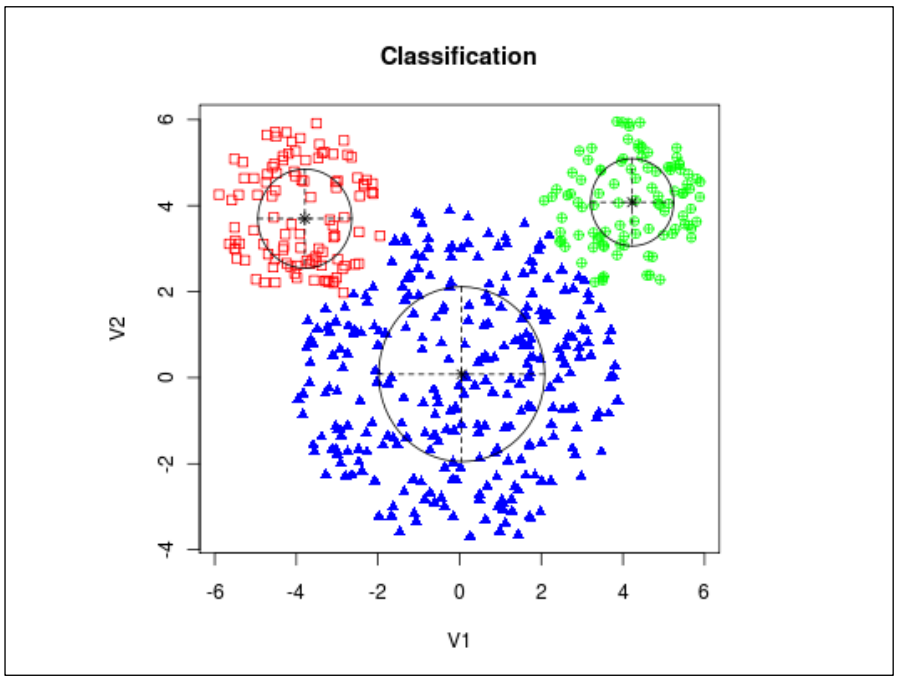


Nota: Descripción de un modelo de aprendizaje supervisado. Las etiquetas corresponden a las variables a predecir, los datos de entrenamiento a la información suministrada para que el modelo aprenda de forma autónoma, el algoritmo de aprendizaje a la metodología empleada para la predicción y el modelo final hace referencia al código listo para ser implementado. Se adaptó del libro Python Machine Learning por Raschka S.

A diferencia de los modelos de aprendizaje supervisados, los no supervisados, son aquellos con datos sin etiquetas o con estructura desconocidas. Mediante su aplicación se busca explorar la estructura de los datos para extraer información significativa sin la necesidad de una variable resultado, buscando así identificar estructuras internas (clasificaciones) dentro de los datos [5]. . Un ejemplo clásico de los modelos de aprendizaje no supervisados es la clasificación de los clientes tomando como base sus intereses, con el fin de desarrollar programas de marketing exclusivos. Este tipo de modelo es muy útil cuando se busca identificar subgrupos con características significativas dentro de un conjunto de datos. A diferencia de los no supervisados este modelo genera clasificaciones en un grupo de información que comparten un

cierto grado de semejanza, siendo útiles para estructurar la información y derivar relaciones significativas de los datos [5].

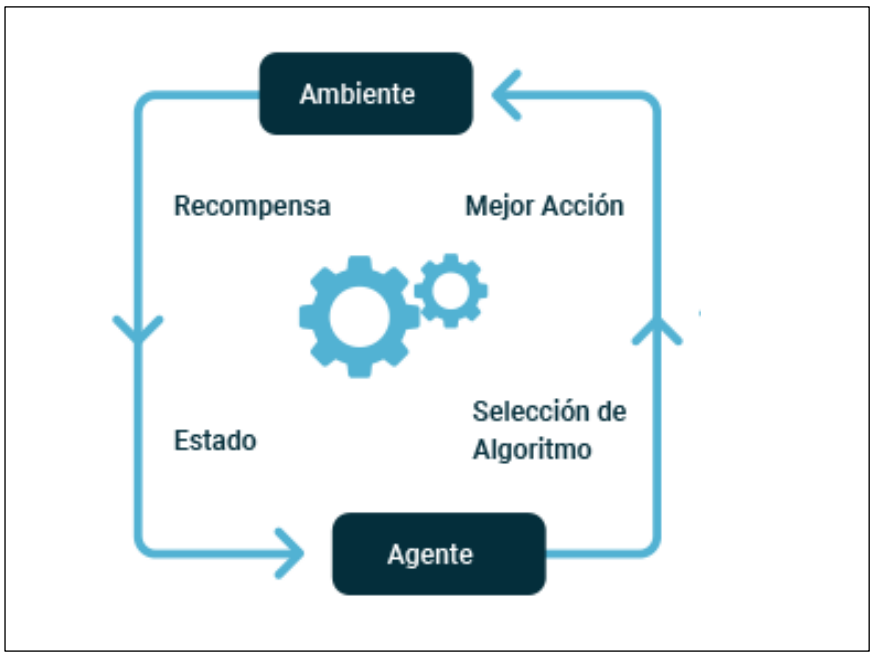
Figura 4.
Aprendizaje no supervisado



***Nota*:** Resultado de un modelo de aprendizaje no supervisado. Los colores rojo, verde y azul hacen referencia a la clasificación generada por este tipo de modelos. Tomado de: Sancho, F. (2020). Aprendizaje supervisado y no supervisado.

Por último se encuentra el aprendizaje reforzado, el cual tiene como objetivo principal desarrollar un sistema (agente) que mejore su rendimiento basado en las interacciones con el entorno donde se desarrolla. El fin último de este modelo es que dicho agente sea capaz de realizar la acción más adecuada frente a una situación planteada. Una de las aplicaciones más comunes de este tipo de aprendizaje es en el entrenamiento de los robots para que aprendan a realizar una tarea en específico. [7]

Figura 5.
Aprendizaje reforzado



Nota: Aprendizaje reforzado y sus componentes: Agente (sistema) encargado de optimizar el rendimiento del modelo y acciones permiten mejorar las interacciones con entorno de desarrollo. Tomado de: Medium. Tipos de aprendizaje automático. Aprendizaje por refuerzo. https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2

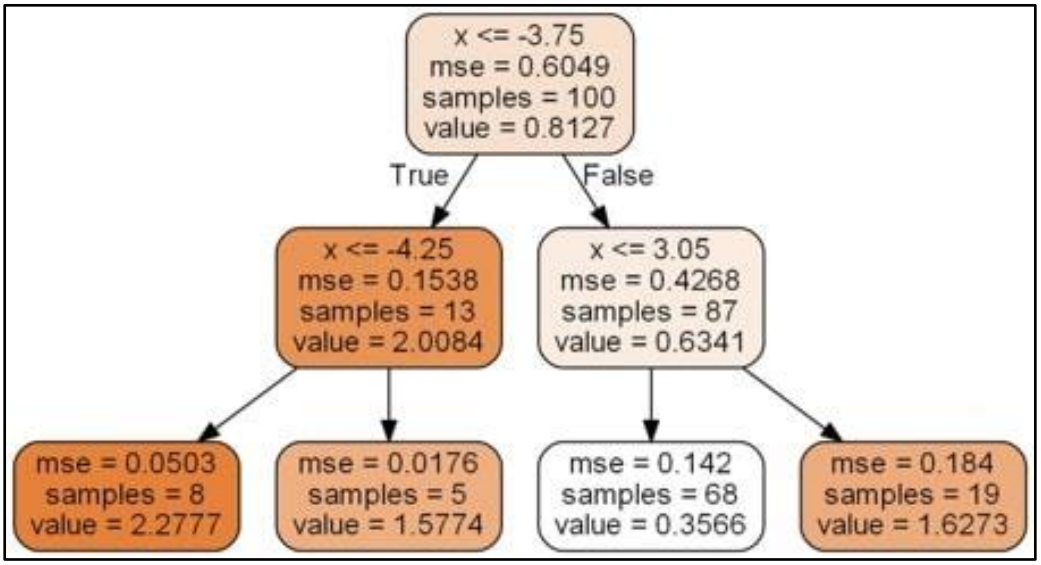
Es válido aclarar que no es posible comparar la eficiencia de un modelo de aprendizaje frente a otro, cada uno tiene su área de aplicabilidad de acuerdo al objetivo final de su implementación, y su predictibilidad será medida en función de la cantidad y calidad de la información suministrada al modelo, así como del ajuste de los parámetros internos de la técnica seleccionada [7].

1.1.2. Métodos de aprendizaje

Una vez determinado el modelo de aprendizaje se debe establecer el método de aprendizaje a implementar. Este último concepto se define como el mecanismo mediante el cual será realizada la predicción o clasificación deseada. Los métodos de aprendizaje supervisado más destacados se conocen como árboles de decisión (*bagging trees*). Este método consiste en una descomposición de los datos mediante la toma de decisiones, tomando como base para su desarrollo la subdivisión con base en criterios de los parámetros suministrados al modelo.

Tienen un primer nodo llamado raíz (*root*) y luego se descomponen el resto de atributos de entrada en dos o más ramas dependiendo de la situación planteada [12]. Por ejemplo el análisis de riesgo en la concesión de créditos bancarios tomando como base salarios, gastos, género y condición social [13]. Este método es muy usado cuando la relación entre las variables no tiene un comportamiento lineal [14] .

Figura 6.
Árbol de decisión



***Nota*:** Ejemplo de la subdivisión de nodos de acuerdo al valor de una variable *x* en un árbol de decisión en Python. Tomado de: Bagnato, J. (2018). Árboles de decisión en Python: clasificación y predicción.

Existen métodos de aprendizaje supervisado más robustos como es el bosque aleatorio (*random forest*), mediante el cual se ensamblan y promedian varios árboles de decisión, obteniendo por cada uno de ellos un puntaje o voto a la salida del mismo, el más votado dará como resultado la opción más opcionada del modelo. La principal ventaja de este método es la disminución del efecto de overfitting o sobreajuste de los datos, evitando que el modelo solo aprenda casos particulares y sea incapaz de reconocer nueva información. Uno de los parámetros más importantes a tener en cuenta es el número de árboles *k* que se eligen para el modelo predictivo, entre más alto sea el número de árboles mejor tiende a ser el rendimiento del bosque aleatorio. [15]

Una vez definidos los aspectos básicos sobre las técnicas de aprendizaje automático, es necesario conocer y entender uno de los conceptos de necesarios previos a la implementación

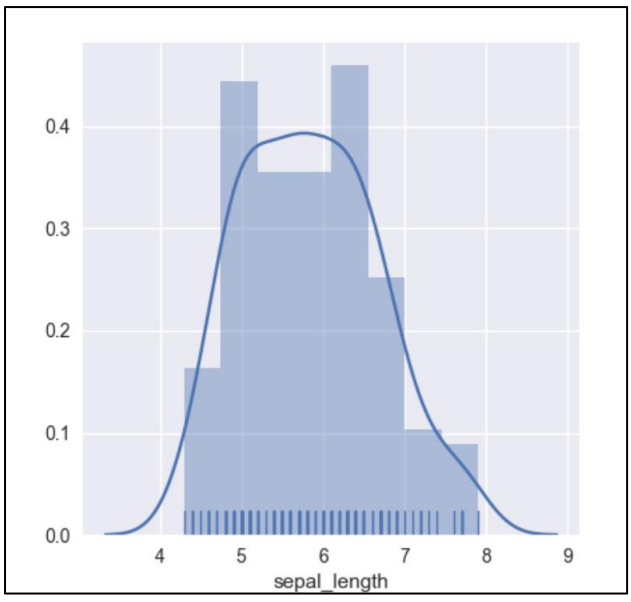
un algoritmo predictivo, el cual tiene como objetivo principal evaluar la calidad y de la información que será cargada en el modelo, Este procedimiento permite eliminar los valores atípicos que no coinciden con el comportamiento de la información, permitiendo la creación de una base de datos congruente [16]; lo anterior se realiza mediante un Análisis exploratorio de datos (*EDA*), el cual se define a continuación.

1.1.3. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos es una técnica que permite verificar la consistencia de la información empleada en un modelo de aprendizaje. Consiste en un conjunto de técnicas y conceptos estadísticos que facilitan el entendimiento y proporcionan calidad en la información suministrada para poder desarrollar un modelo predictivo [16].

Este tipo de análisis no posee un conjunto de procedimientos específicos para su desarrollo, sin embargo, dentro de los principales conceptos estadísticos empleados se encuentran los histogramas, los cuáles son una representación gráfica que permite observar las cantidades acumuladas de los valores obtenidos para la variable en estudio [6].

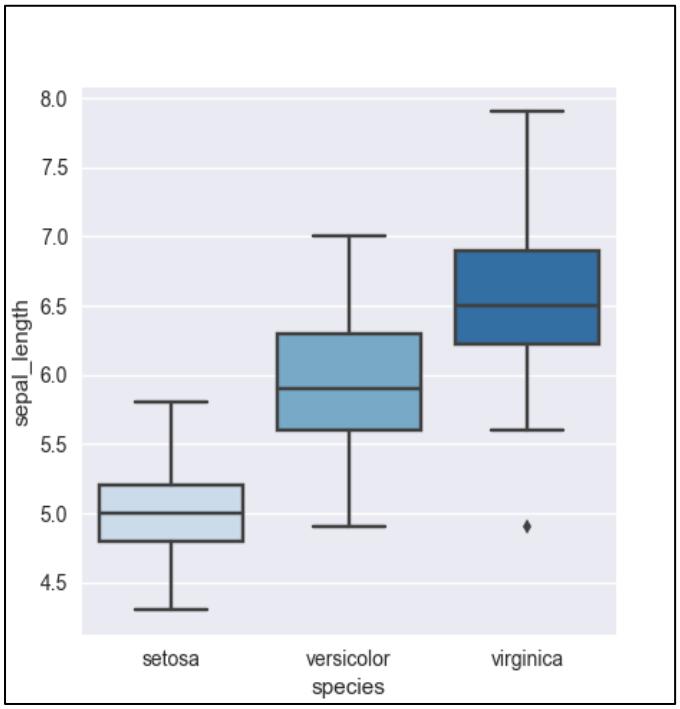
Figura 7.
Histograma



***Nota*:** Ejemplo de un Histograma con distribución normal. Tomado de: The Python Graph Gallery. Hystogram.

Otra manera eficiente de verificar la consistencia de los datos son los diagramas de cajas, los cuáles se definen cómo un tipo de representación visual para la distribución de valores de una muestra. Este diagrama dará una idea sobre la dispersión de una variable, permitiendo a su vez por medio de una línea que sale de los extremos de la caja detectar valores inusuales y atípicos. [6].

Figura 8.
Diagramas de caja



Nota: Diagramas de caja de varias variables, distintos tipos de distribución y valores atípicos. Tomado de: The Python Graph Gallery. Boxplot

1.2. Perforación

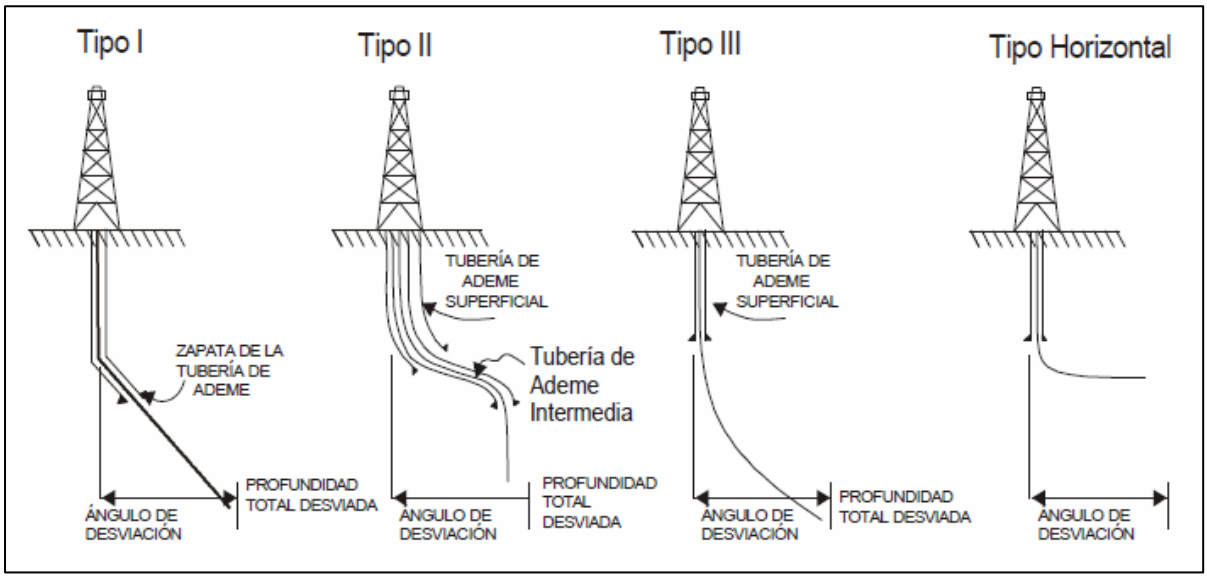
Los conceptos anteriormente mencionados tienen como objetivo principal en este trabajo ser implementados a los parámetros de perforación (Caudal, peso sobre la broca, revoluciones por minuto y torque) en el campo de estudio para correcta estimación de la tasa de penetración y la energía mecánica específica, como se mencionó anteriormente en los objetivos.

De esta manera, es importante tener fundamentos teóricos de la perforación de pozos petroleros, la cual es conocida como un proceso mediante la cual se realiza un agujero empleando una broca para establecer comunicación entre la superficie y una formación

productora con el fin extraer hidrocarburos desde el subsuelo de una forma segura y económicamente rentable [17].

Dentro de las técnicas de perforación se encuentra la perforación direccional, en la cual se genera una desviación intencional del pozo con respecto a la vertical con el fin de hacer más eficiente la operación. Esta a su vez puede subdividirse en 4 tipos, los cuales se pueden observar en la siguiente Figura. [18]

Figura 9.
Tipos de pozos petroleros de acuerdo a su geometría



Nota: Tipos de pozos petroleros de acuerdo a su geometría y ángulo de desviación. Los pozos Tipo I, se perforan de modo que la desviación inicial se realice a poca profundidad; los pozos Tipo II son de configuración en "S"; en los pozos Tipo III la desviación es a mayor profundidad y el ángulo promedio de inclinación se mantiene hasta llegar al objetivo, y los pozos Horizontales, los cuales una vez es construido el ángulo se mantienen en geo-navegación de la arena productora. Tomado de: Madrid, M. (2016) Portal del petróleo. Perforación Direccional, Tipos de Perforación, Propósitos y Motor de Fondo. https://www.portaldelpetroleo.com/2016/03/perforacion-direccional-tipos-de_6.html

1.2.1. Parámetros de perforación

Tabla 1.
Parámetros de perforación

Parámetro	Sigla	Unidades	Descripción
Tasa de penetración	ROP	Pies/hr	Velocidad a la que se profundiza durante la perforación
Revoluciones por minutos	RPM	Rpm	Tasa a la que la broca rota durante la operación
Torque	TQ	Klbs-ft	Fuerza creada por la sarta de perforación debida a la rotación
Peso sobre la broca	WOB	Klbs	Peso ejercido sobre la broca
Presión en la tubería	SPP	Psi	Presión de circulación del fluido de perforación circulando en la tubería
Caudal	GPM	Gpm	Tasa de flujo del fluido de perforación en el sistema
Peso del lodo	MW	Lgs/gl	Densidad del fluido de perforación en el sistema

Nota: Parámetros de perforación medidos en superficie, su definición y sus unidades. Tomado de: Portilla E., Suarez F., Corzo R.. Metodología para la optimización de parámetros de perforación a partir de propiedades geomecánicas. Revista El Reventón energético.

1.2.2. Energía mecánica específica (MSE)

Los parámetros mencionados previamente en conjunto permiten generar una métrica de desempeño de la operación conocida como Energía mecánica específica (MSE), la cual se define como la energía requerida para remover una unidad de volumen de roca [19]. Esta puede ser calculada mediante la siguiente ecuación:

Ecuación 1.

Energía mecánica específica

$$MSE = \frac{480 * T * RPM}{D^2 * ROP} + \frac{4 * WOB}{\pi * D^2}$$

En donde,

MSE = Energía mecánica especifica

T = Torque

RPM = Revoluciones por minuto

D = Diámetro de la broca

ROP = Tasa de penetración

WOB = Peso sobre la broca

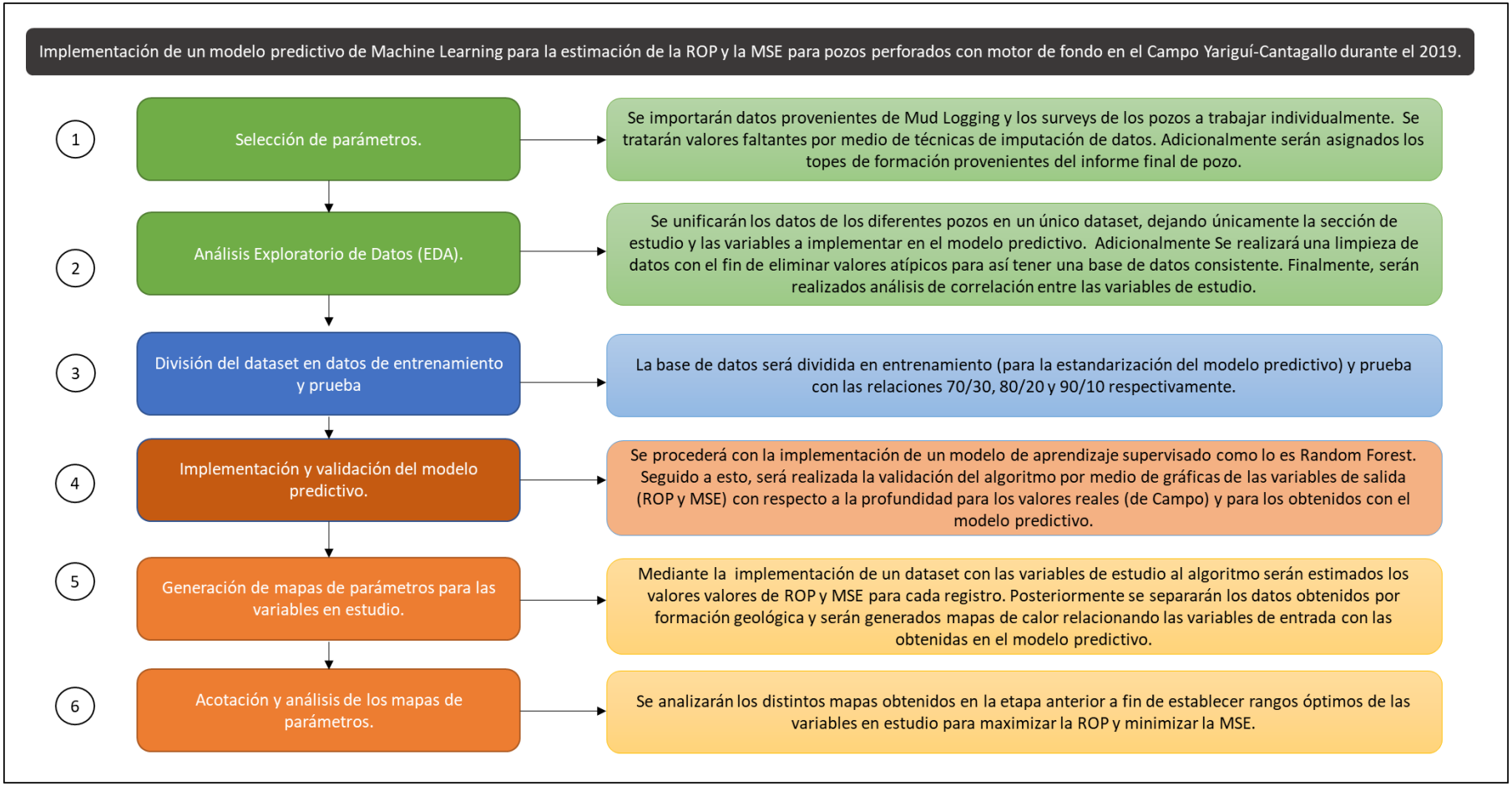
Este concepto fue introducido por Theale en 1964, y es equivalente a la energía impuesta sobre la broca de perforación con respecto a la profundidad alcanzada en el tiempo [20]. El uso de esta métrica permite optimizar el desempeño de la operación al poder analizar de forma conjunta los principales parámetros asociados a la misma, siendo la tasa de penetración el parámetro con mayor influencia sobre esta con una relación inversa. Sin embargo, parámetros como el torque, las revoluciones por minuto y el peso sobre la broca conservan una relación directa con la métrica en mención, y al ser manejados en conjunto también permiten optimizar el desempeño de la operación [21].

2. METODOLOGÍA Y DATOS

En esta sección se expone la metodología propuesta para desarrollar la presente investigación. Esta inicia con la recopilación de la información correspondiente a los parámetros de perforación para los pozos del campo en estudio. A partir de allí se exhibe el proceso para la creación de la base de datos, que consta el tratamiento de valores faltantes, el cálculo de la profundidad vertical verdadera (TVD) y la asignación de topes de formación. Seguidamente, se realiza un análisis exploratorio de datos EDA, el cual busca obtener una descripción estadística de las variables, permitiendo identificar y eliminar valores atípicos de la información con el fin de generar una base de datos consistente [22]. Llegados a este punto se procede a implementar el modelo de aprendizaje automático supervisado *Random Forest*, inicialmente con las proporciones 70/30, 80/20 y 90/10 de la base de datos para su entrenamiento y validación respectivamente. Finalmente se generan mapas de calor con los parámetros de estudio por formación geológica, con el fin de establecer los rangos óptimos de las variables para maximizar la ROP y minimizar la MSE.

A continuación, se procederá a explicar paso a paso la metodología para ejecutar y lograr los objetivos del proyecto (Figura 10):

Figura 10.
Diagrama de la metodología propuesta



Nota: Metodología para la implementación de un algoritmo predictivo de machine learning para la estimación de la ROP y la MSE en el Campo Yariguí-Cantagallo durante el 2019. Los colores verde, azul, rojo y naranja corresponden a los cuatro objetivos específicos del presente proyecto.

2.1. Selección de parámetros

Conocer la geometría del pozo, tener la información correspondiente a los parámetros de perforación del campo en estudio, las secciones y formaciones geológicas perforadas es clave en la generación de la base de datos a implementar en el modelo predictivo para la estimación de la ROP y la MSE. Teniendo en cuanta la Ecuación (1) en el marco teórico se deben trabajar con las siguientes variables: las secciones perforadas, el caudal (Q), las revoluciones por minuto de superficie y de fondo (RPM), el peso sobre la broca (WOB), el torque (TQ), la tasa de penetración (ROP) y la energía mecánica específica (MSE). Adicionalmente se requiere información referente la profundidad medida (MD), la profundidad vertical verdadera (TVD) y el ángulo de desviación (INCL) para la asignación de las formaciones geológicas perforadas. Toda la información proveniente de los reportes finales de: Surveys para los datos asociados a profundidad y ángulo de inclinación y Mud Logging para las variables correspondientes a los parámetros de perforación. Dicha información fue proporcionada por Ecopetrol S.A.

2.1.1. Importar librerías y bases de datos

La visualización e implementación del modelo predictivo se realiza en Python importando diferentes librerías de trabajo. Estas se encuentran instaladas en el gestor de entorno Anaconda Navigator, el cual funciona como interfaz gráfica para la ejecución de algoritmos de programación [23]. Las librerías usadas se listan y describen a continuación:

Tabla 2.
Librerías de Python utilizadas en el proyecto

Libreria	Descripcion	Aplicación
Pandas	Herramienta de manipulación de DataFrames y análisis de datos. Un DataFrame permite almacenar y manipular datos tabulados en filas (información) y columnas (variables).	Se usa para importar las variables de los pozos y trabajar la base de datos en el Objetivo 1.
Mathplotlib	Generación de gráficos.	Empleada para la generación de las gráficas del Análisis Exploratorio de Datos en el Objetivo 1.
Missigno	Visualización de valores faltantes.	Visualizar valores faltantes en la base de datos en el Objetivo 1.
Seaborn	Generación de gráficos	Se usa para la generación de las matrices de correlación y los mapas de calor (Objetivos 1 y 4).
Numpy	Manejo y adecuación de la información para implementar modelos predictivos.	Se utiliza para la conversión del DataFrame a matriz, previo a la implementación del modelo predictivo (Objetivo 2).
ScikitLearn	Algoritmos de aprendizaje automático.	Utilizada para la división de la base de datos en entrenamiento y prueba, así como para la implementación y ajuste del modelo predictivo (Objetivos 2 y 3).

***Nota*:** Se describen las librerías empleadas en la implementación del modelo predictivo así como su uso principal. Tomado de: Python Machine Learning.

2.1.2. Verificación del dataframe y tratamiento de valores faltantes

Teniendo en cuenta lo mencionado anteriormente se importaron las bases de datos de cada uno de los pozos a trabajar, con ayuda de la librería Pandas, las cuales serán manejadas como DataFrames.

Una vez visualizados (cada uno de) los DataFrame se procede a determinar el tipo de dato como flotante (float) y objeto (object), dependiendo de si son datos cuantitativos continuos o categóricos nominales respectivamente. Así pues las variables sección (SECTION), profundidad medida (MD), inclinación (INCL), revoluciones por minuto de superficie y de fondo (RPM_SURF, RPM), toque (TQM), peso sobre la broca (WOBK), caudal (GPM), tasa de penetración (ROP) y energía mecánica específica (MSE), son del tipo cuantitativo

continuo (float), y las variables nombre y tipo de pozo (WELL_NAME, TYPE) corresponden al tipo categórico nominal (object), como puede apreciarse en la Tabla 3. Finalmente se tienen 10 variables float y 2 object.

Tabla 3.
Caracterización inicial de las variables

Tipo de variable	Variable	Non-null count
Object	WELL_NAME	8954 Non-null
	WELL_TYPE	8954 Non-null
Float	SECTION	8954 Non-null
	MD	8954 Non-null
	INCL	120 Non-null
	RPM_SURF	8954 Non-null
	RPM	8954 Non-null
	TQM	8954 Non-null
	WOBK	8954 Non-null
	GPM	8954 Non-null
	ROP	8954 Non-null
	MSE	8954 Non-null

Nota: Descripción de las variables para un pozo. Las variables tipo *object* (verde) hacen referencia a variables cualitativas contínuas, y las tipo *float64* (azul) a variables categóricas nominales. El número de entradas corresponde a la longitud de cada pozo. Se puede observar que la variable inclinación posee valores faltantes (cuadro rojo), al tener menor cantidad de registros (*non-null)*, con respecto a los demás parámetros.

Con el fin de obtener una base de datos generalizada para las secciones en estudio, se hace necesario el cálculo de la profundidad vertical verdadera (TVD), dicha magnitud es estimada mediante la siguiente ecuación:

Ecuación 2.
Cálculo de la profundidad vertical verdadera (TVD)

$$TVD_i = [(MD_i - MD_{i-1}) * COS(INCL_i)] + TVD_{i-1}$$

Donde:

TVD_i Profundidad vertical actual

MD_i Profundidad medida anterior

$INCL_i$ Inclinación actual

TVD_{i-1} Profundidad vertical anterior

Como se puede identificar en la Figura 15 la variable inclinación posee menos campos que el resto de variables (leer la columna "No null count"), mientras las demás variables tienen 8954 entradas esta solo posee 120 entradas, esto debido a que la medición de los parámetros de perforación es realizada cada pie, mientras que el ángulo de inclinación es registrado cada 100 pies aproximadamente. Con el objetivo de completar la información faltante, los valores de INCL de la ecuación anterior fueron estimados mediante la interpolación lineal, siendo esta la técnica de imputación de datos más adecuada para la variable en mención, al permitir una transición gradual entre un valor y otro [25].

2.1.3. Asignación de topes de formación

Una vez son tratados los valores faltantes y es realizado el cálculo de la profundidad vertical verdadera, se procede a asignar variables categóricas (tipo "*object*") correspondientes a las formaciones geológicas perforadas, las cuales fueron estimadas a través de registros, principalmente Gamma Ray, obtenidos de los informes finales de cada uno de los pozos, proporcionados por las compañías de servicios. Estas formaciones serán asignadas con

respecto a la TVD y serán usadas principalmente para la generación de mapas de parámetros en el cuarto objetivo.

Tabla 4.
Profundidades mínimas y máximas de las Formaciones Geológicas perforadas

Formación	TVD Mínima (ft)	TVD Máxima (ft)
Real	1450	2900
Colorado	2740	4180
La Cira Shale	2370	4320
Mugrosa	3340	6220
La Paz (C)	5650	7070
La Paz (CG)	6410	8340

Nota: Rangos mínimos y máximos de las formaciones geológicas perforadas con respecto a la profundidad vertical verdadera.

Partiendo de esta información fueron generados dos data sets teniendo en cuenta la geometría de los pozos (Horizontal o tipo "S"), las secciones y variables de estudio (TVD, Formación geológica, RPM_SURF, RPM, TQ, WOBK, GPM, ROP y MSE) . Las bases de datos fueron generadas a través de la librería Pandas como se muestra a continuación.

Tabla 5.
Pozos con motor perforados en el campo Yariguí-Cantagallo durante el 2019

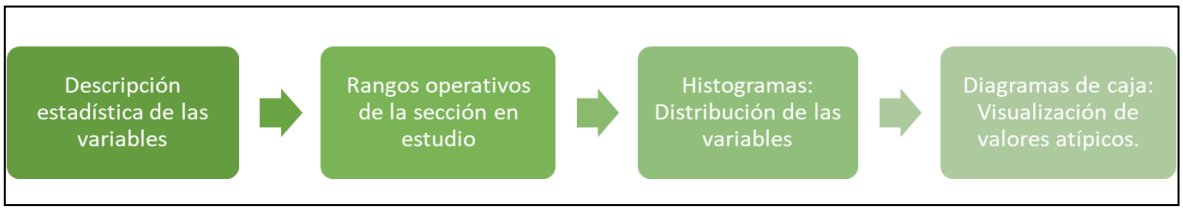
Cantidad de pozos	Geometría	Sección de estudio	Formaciones perforadas
6	Tipo "S"	8 ½''	Real, Colorado, La Cira Shale, Mugrosa, La Paz
5	Horizontal	12 ¼ ''	Real, Colorado, La Cira Shale, Mugrosa

Nota: La figura discretiza los pozos por geometría, sección de estudio y formaciones geológicas para las bases de datos a implementar en el modelo predictivo.

2.2. Análisis Exploratorio de datos.

Una vez establecidas las bases datos a implementar en el modelo predictivo, se realiza una descripción de cada una de las variables involucradas en el presente proyecto, con el fin de describir su comportamiento, eliminar valores atípicos (*outliers*) y establecer correlaciones entre las mismas, siendo empleados los criterios mostrados en la Figura 11.

Figura 11.
Etapas del Análisis Exploratorio de Datos



Nota: Se describen las diferentes etapas que conforman el Análisis Exploratorio de Datos de la presente investigación, con el objetivo principal de generar una base de datos consistente para implementar en el modelo predictivo.

Estas métricas son generadas a través de las librerías Matplotliby Pandas. Por medio de los conceptos estadísticos previamente mencionados, se busca tener valores representativos de cada una de las variables durante la operación.

2.2.1. Descripción estadística de las variables

Con las bases de datos establecidas anteriormente es realizada una caracterización estadística de las variables en estudio para cada una de las bases de datos a implementar en el modelo predictivo, las cuales pueden apreciarse en los Anexos 1 y 2.

Asimismo, durante la planeación de un pozo son establecidos los rangos de parámetros teniendo en cuenta su geometría, experiencia operacional y ensamblaje de fondo [26]. Dichos rangos para las secciones y pozos en estudio se encuentran en la Tabla 6, y serán tomados como referencia para las bases de datos manejadas en la presente investigación.

Tabla 6.
Datos de entrada. Rangos operaciones para las secciones en estudio.

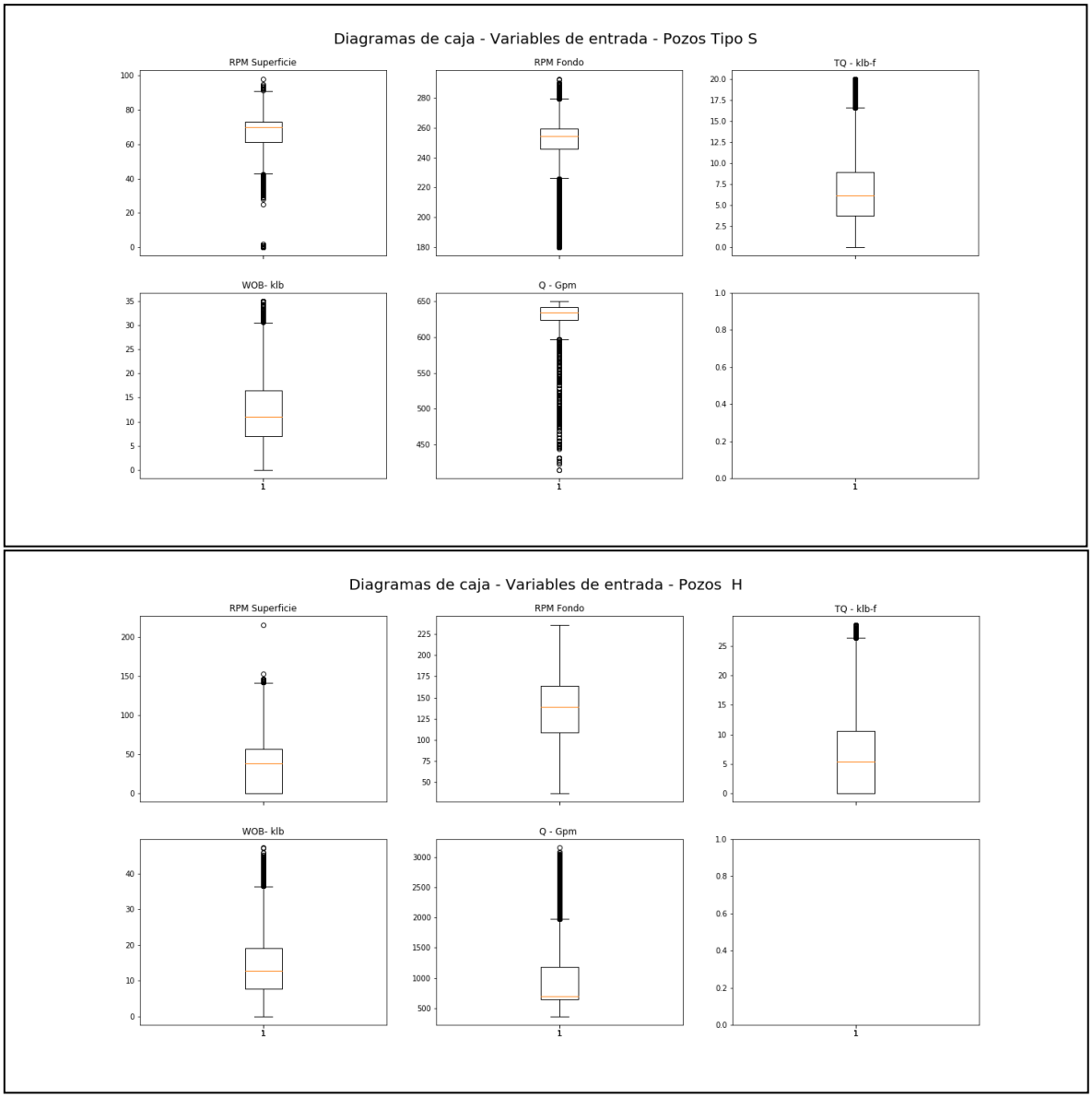
Parámetro	Unidades	Tipo S – Sección 8 ½''	Horizontales – Sección 12 ¼''
RPM Superficie	Rev/min	0-100	0-100
RPM Fondo	Rev/min	100-300	60-250
Torque	Klb-f	0-28	0-32
Peso sobre la broca	Klb	1-35	1-42
Caudal	Gal/min	400-700	700-1500

Nota: La figura describe los rangos operacionales establecidos de cada una de las secciones para las variables de entrada del modelo predictivo.

De igual manera fueron generados diagramas de caja para las variables de entrada del modelo, las cuales permiten describir el comportamiento de los datos identificando la distribución de los mismos en cuatro grupos segmentados por: el mínimo, primer cuartil (Q1), mediana, tercer cuartil (Q3) y máximo (Figura 12). Adicionalmente, fueron generados histogramas con el fin de establecer la distribución de los valores de cada una de las variables (Figura 13).

Figura 12.

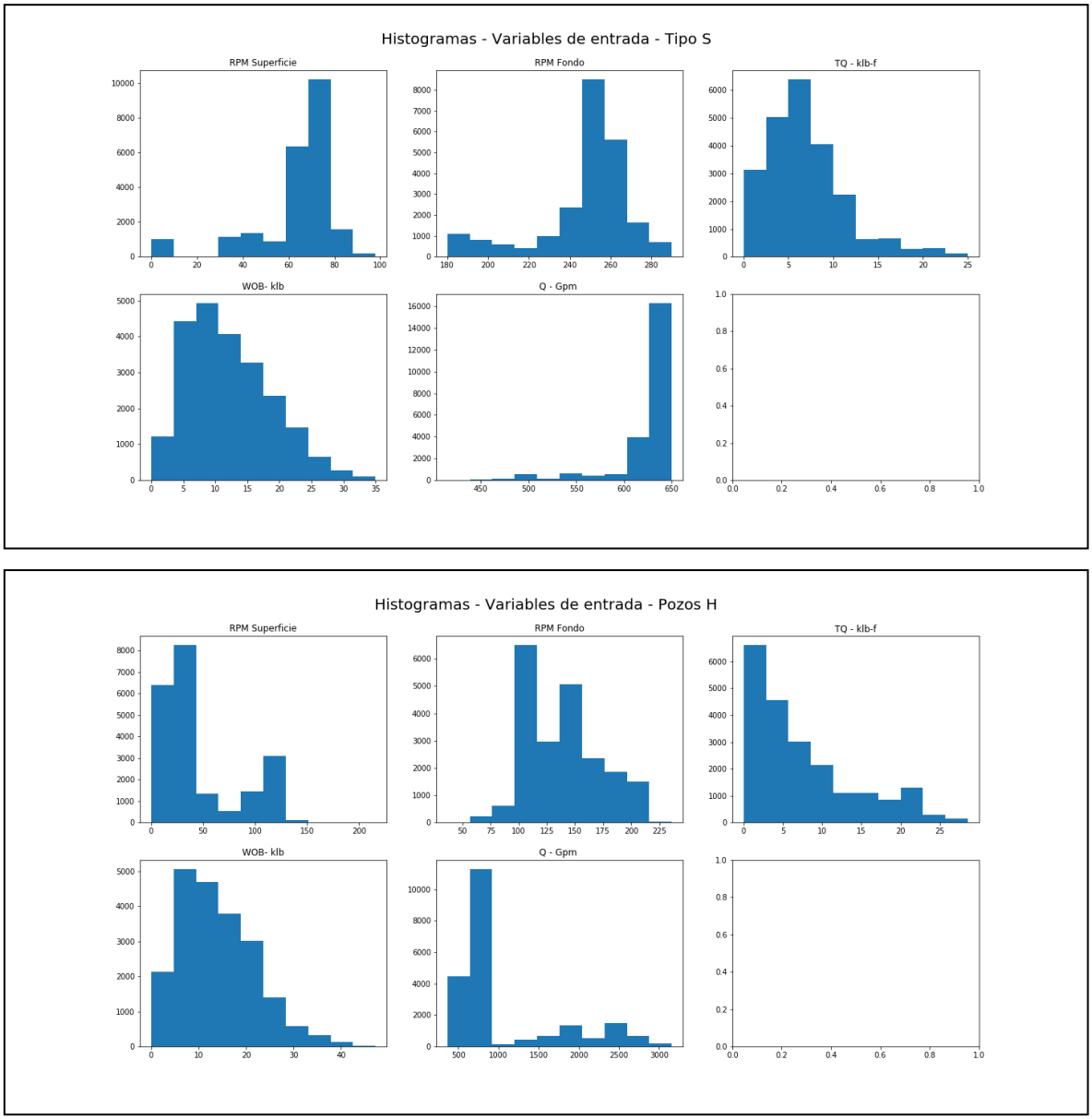
Diagramas de caja para las variables y secciones en estudio



Nota: La figura muestra los diagramas de caja para las dos bases de datos trabajadas. La línea naranja corresponde a la mediana de los datos, los puntos negros por fuera de los bigotes son considerados como valores atípicos.

Figura 13.

Histogramas para las variables y secciones en estudio



Nota: La figura muestra los histogramas para las dos bases de datos trabajadas. Las variables torque y peso sobre la broca evidencian sesgo hacia a izquierda, mientras que el caudal y las revoluciones por minuto de superficie y de fondo se encuentran sesgadas a la derecha.

Estas Figuras permiten describir el comportamiento de las variables de entrada del modelo predictivo, en el caso de las revoluciones por minuto, los rangos en superficie para las secciones en estudio pueden oscilar entre 0 (al momento de deslizar) y 100 (lo máximo permitido por el motor). Las variaciones de este parámetro suelen ser consecuencia del uso de un caudal superior o inferior al proyectado [26, 4].

Con respecto al caudal, para las secciones 8 ½'' y 12 ¼'' el caudal se mantiene estable en valores cercanos a los 650 y 750 GPM respectivamente, a excepción del *drillout* y las primeras paradas de cada sección, mientras el lodo alcanza la temperatura y condiciones adecuadas para la perforación [26, 4].

El peso sobre la broca se encuentra dentro de los rangos operacionales establecidos para cada una de las secciones en estudio, concentrando la mayor cantidad de valores en el rango 5-15 klbs para la sección 8 ½'' y 10-20 klbs en la sección 12 ¼''. En caso de presentar valores mayores a los establecidos, la broca tiene una mayor probabilidad de desgaste; por el contrario, si su valor es menor, la Tasa de Penetración (ROP) tiende a disminuir, afectando la eficiencia de la operación [27].

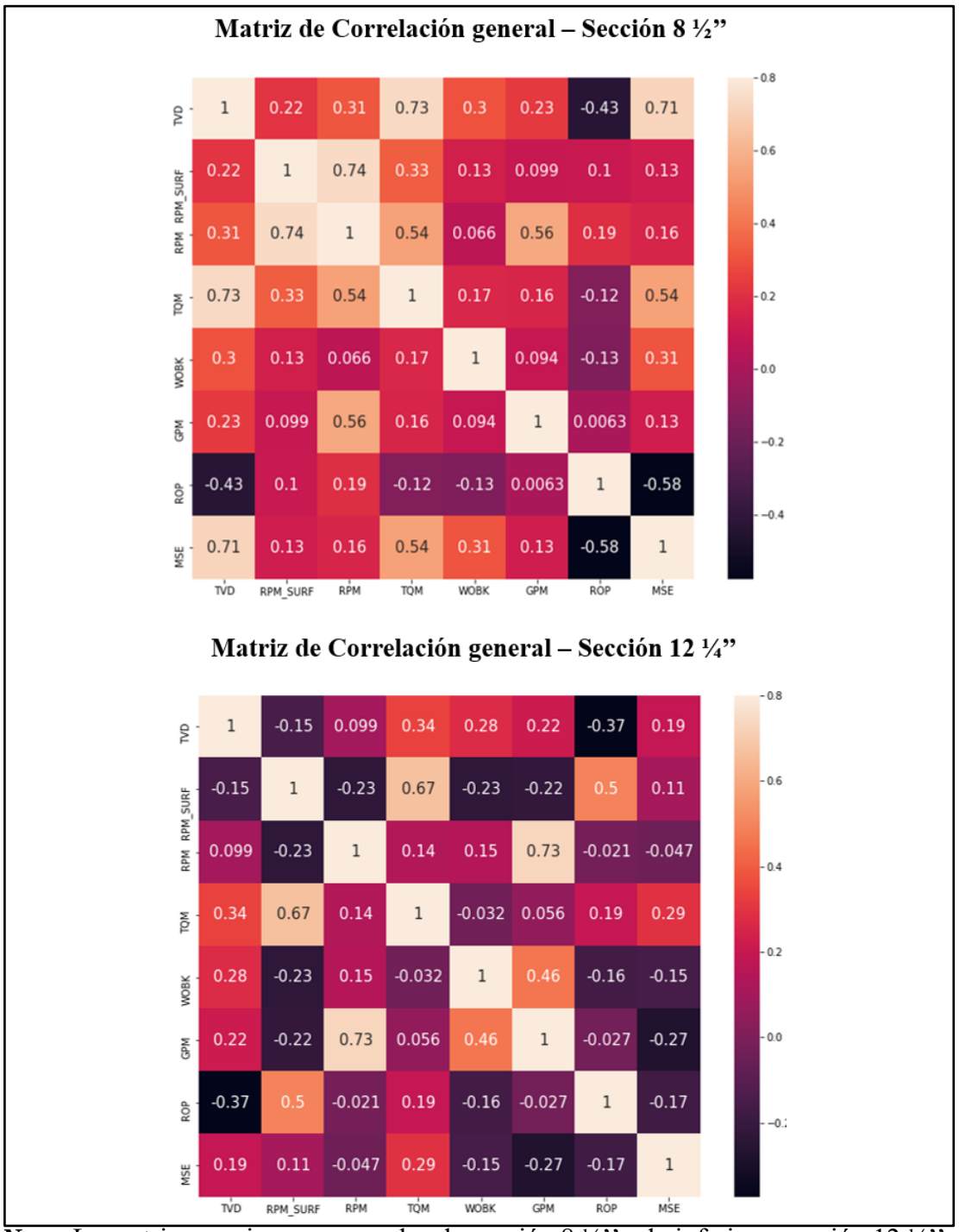
Las mediciones de torque registradas en las bases de datos son consistentes de acuerdo con los valores establecidos en los rangos operacionales para cada una de las secciones. Los valores de esta variable se ven afectados directamente por el peso sobre la broca y las revoluciones por minuto, así como en el mantenimiento de la inclinación para la construcción ángulos durante la operación [28].

2.2.2. Análisis de correlación entre las variables

Una vez realizada la limpieza y descripción estadística de cada una de los parámetros de perforación, se procede a establecer la correlación entre los mismos y su incidencia en las variables a predecir para las secciones objeto de estudio de forma general como puede observarse en la Figura 14 y por formación geológica en los Anexos 1 y 2.

Figura 14.

Matrices de correlación para las secciones en estudio



***Nota*:** La matriz superior corresponde a la sección 8 ½’’ y la inferior a sección 12 ¼’’. Estas matrices permiten evidenciar el grado correlación directa (colores rojizo y naranja) e inversa (colores morado y negro) entre variables asociadas a la perforación para las dos secciones en estudio.

A partir de la Figura se puede observar la incidencia de cada una de los parámetros de entrada sobre las variables a predecir. El caudal, el peso sobre la broca, el torque, las revoluciones por minuto de superficie y de fondo y la profundidad vertical tienen una relación directa sobre a la Energía Mecánica Específica e inversa con respecto a la Tasa de Penetración. Esta información será correlacionada con los mapas de calor generados en el cuarto objetivo del proyecto, a fin de establecer los rangos óptimos y las relaciones de las variables en estudio.

2.3.División del Dataset en datos de entrenamiento y prueba

En este punto se cuenta con una base de datos consistente para la implementación del modelo predictivo, sin embargo, para la implementación de este tipo de algoritmos se requiere una adaptación previa de la base de datos, la cual se describe en la Figura 15.

Figura 15.
Procedimiento para la división de la base de datos en entrenamiento y prueba

Nota: Se evidencian los tres pasos fundamentales a desarrollar en una base de datos para la implementación de un modelo predictivo.

La etapa codificación (Etapa A) consiste en asignar una variable cuantitativa a las variables categóricas, ya que el algoritmo de predicción solo admite variables de tipo numérico. En este caso las Formaciones Geológicas son variables categóricas. Para ello, se empleará algoritmo *get_dummies* de la librería Pandas, el cual genera una columna por cada una de las variables categóricas presentes (formaciones geológicas perforadas), asignando automáticamente un valor de 1 si es la formación correspondiente al registro y 0 en caso contrario [29, 30]

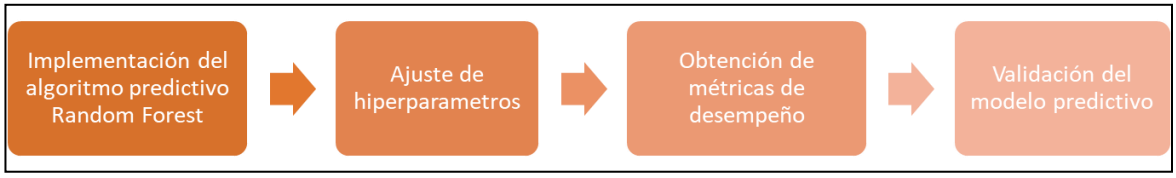
Seguido a esto, a través de la librería Numpy se realizó la conversión del DataFrame en una matriz (*Array)* para la implementación del modelo predictivo (Etapa B), así como la clasificación del mismo en atributos (*Features)* para las variables de entrada del modelo predictivo (Caudal, revoluciones por minuto, peso sobre la broca, formación geológica y torque) y etiquetas (*labels*) para la variable a predecir (Tasa de penetración y energía mecánica específica).

Para finalizar es realizada la división del data-set en datos de entrenamiento, los cuales servirán para el aprendizaje a través de la experiencia del algoritmo, y datos de prueba, para verificar la exactitud en la predicción del mismo, las proporciones a emplear para la división de la base de datos es 70/30, 80/20 y 90/10 respectivamente, con el fin de establecer la división óptima para la predicción. Esta segmentación de la base de datos es realizada con la Librería Scikit-Learn, a través del algoritmo *Train_Test_Split*, el cual, al indicarle el tamaño de las muestras para entrenamiento y prueba, genera la división aleatoria del Dataset para la calibración y validación del modelo predictivo.

2.4. Implementación y validación del modelo predictivo

Teniendo en cuenta las divisiones realizadas en el numeral anterior, se procede a implementar el modelo de aprendizaje supervisado conocido como *Random Forest Regressor* de la librería ScikitLearn, siguiendo los pasos mostrados en la Figura 16.

Figura 16.
Implementación del algoritmo predictivo Random Forest Regressor



Nota: Se describen las diferentes etapas que conforman la implementación del modelo predictivo de la presente investigación.

Inicialmente se mantendrán los hiperparámetros por defecto para obtener un desempeño inicial. Se entiende por hiperparámetro un criterio del modelo que se establece antes del inicio del proceso de aprendizaje.

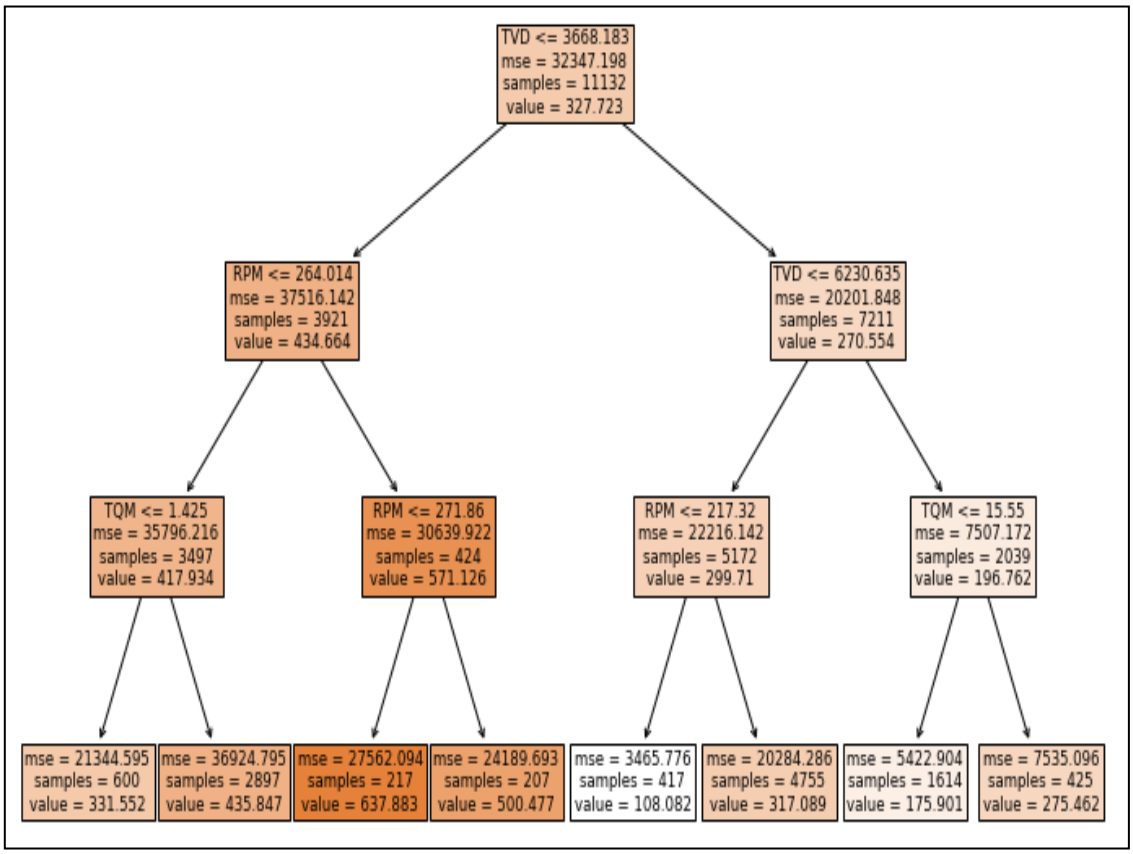
Para este caso en específico, los principales hiperámetros a trabajar son [31]:

- *n_estimators:* especifica el número de árboles en modelo. El valor predeterminado para este parámetro es 100.
- *max_depth:* la profundidad máxima de cada árbol. El valor predeterminado para max_depth es None, lo que significa que cada árbol se expandirá hasta que cada hoja tenga una única clase (variable).
- *min_samples_split:* es el número mínimo de muestras necesarias para dividir un nodo. El valor predeterminado para este parámetro es 2, esto quiere decir que un nodo interno debe tener al menos dos muestras antes de poder dividirlo.
- *min_samples_leaf:* número mínimo de muestras que debe haber en un nodo final. El valor predeterminado para este parámetro es 1, lo que significa que cada hoja debe tener al menos 1 muestra para esta clase.

El algoritmo selecciona aleatoriamente registros de la base de datos para conformar una muestra, esto es conocido como *Bootstrapping*. Seguido a esto tomará el parámetro con mayor incidencia sobre la misma como nodo inicial y posteriormente se irá subdividiendo bajo el mismo criterio en los demás parámetros hasta alcanzar la profundidad establecida (Figura 17). Esto es realizado para la cantidad de árboles de decisión determinada en el ajuste de los hiperparametros, siendo el resultado final un promedio del resultado obtenido a través de todas las muestras (Figura 18) [32].

Figura 17.

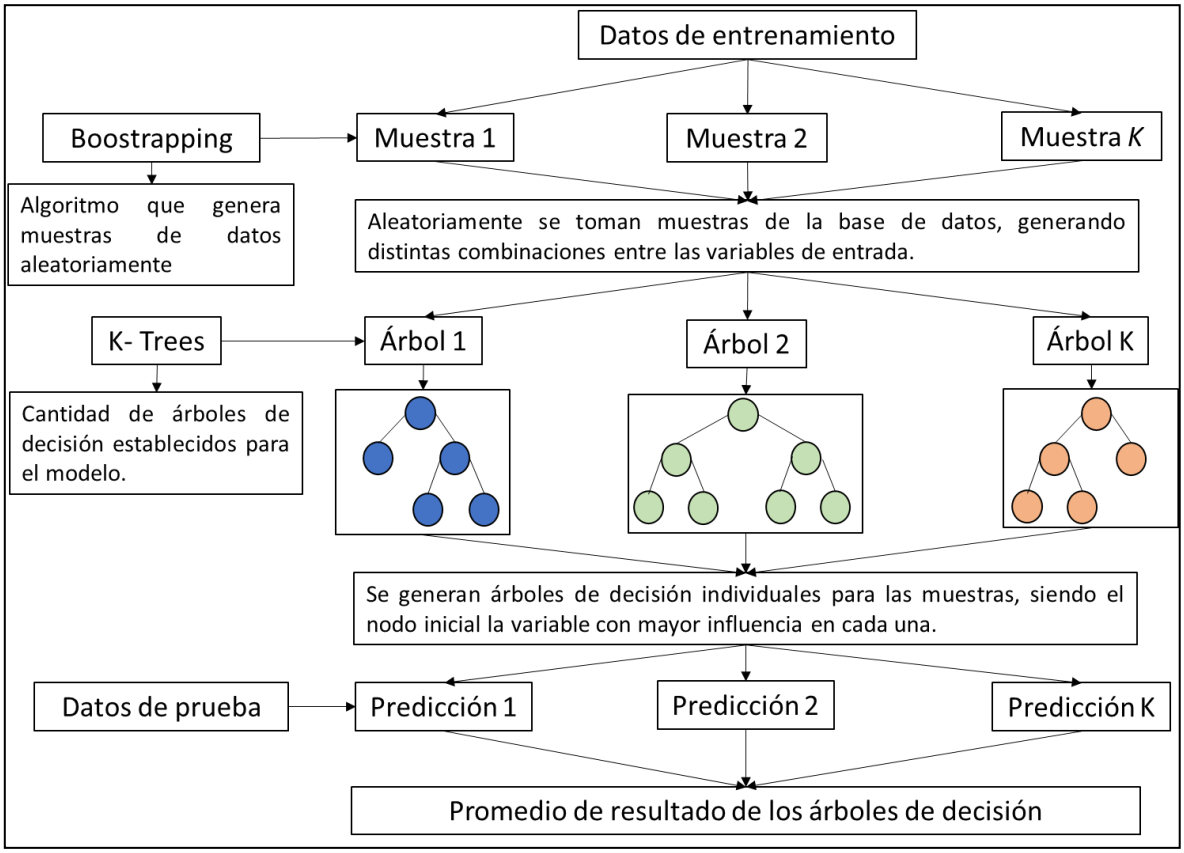
Árbol de decisión del modelo



Nota: La figura muestra uno de los 500 árboles de decisión del modelo predictivo, como nodo principal se encuentra la profundidad vertical verdadera, seguido por las revoluciones por minuto y torque.

Figura 18.

Esquema de funcionamiento del algoritmo de aprendizaje Random Forest Regressor



Nota: La figura muestra como la información de entrenamiento del modelo es seccionada con muestras de variables para posteriormente generar árboles de decisión de cada una de ellas. El resultado final es por medio de un promedio de los árboles de decisión generados internamente.

Una vez es realizada la implementación del modelo predictivo, se realiza la validación del mismo con el porcentaje restante del set de datos, con el fin de obtener las variables de salida ROP y MSE. Con el fin de establecer un métrica de desempeño del modelo realizado, se empleará el Error Porcental Absoluto Medio, *(MAPE - Mean Absolute Percentage error),* la cual es una relación entre las variables totales acertadas con respecto al número de precciones realizadas, tal como se muestra en la siguiente ecuación [9]:

Ecuación 3.
Error Absoluto Porcentual Medio

$$MAPE = \sum_{i=1}^{N} \frac{x_i - \hat{x}_i}{x_i} * 100\%$$

Donde

x_i Observaciones reales

$\hat{x}_i$ Predicciones realizadas

De esta manera, se evaluará el Error Porcentual Absoluto Medio para los tres escenarios contemplados en la fase de división de la base de datos sin ajuste de hiperparametros para las secciones 8 ½'' y 12 ¼'' respectivamente. El modelo será considerado aceptable con valores superiores al 70% de esta métrica de desempeño.

2.4.1. Optimización de los hiperparámetros del modelo

Con el propósito de obtener un modelo de aprendizaje optimizado, se ajustarán los hiperparámetros mencionados anteriormente para aumentar la precisión del mismo (disminuir el MAPE), este proceso es conocido como *Tunning* [31] e igualmente es ejecutado por medio de la librería ScikitLearn.

Para este proceso se establecerán rangos de cada uno de los hiperparámetros realizando una combinatoria entre los mismos, los cuales serán probados en el modelo por medio de un proceso iterativo para finalmente obtener los valores óptimos de cada hiperparámetro para el algoritmo predictivo [33].

Adicionalmente será evaluada la importancia de cada una de las variables del modelo predictivo, con el fin de establecer los parámetros con mayor incidencia en la Tasa de Penetración y la Energía Mecánica específica durante la ejecución del algoritmo.

2.4.2. Validación del modelo predictivo

Seguido a esto será realizada la validación del proyecto, por medio de gráficas correspondientes a las variables de salida reales (obtenidas de la base de datos) y las predichas

por el modelo con respecto a la profundidad, esto será realizado con el porcentaje de la base de datos destinado para el *test* y con los hiperparámetros previamente optimizados en el algoritmo predictivo.

2.5. Generación de mapas de parámetros para las variables en estudio

Con el modelo previamente optimizado y validado se procede a generar mapas de parámetros generales y por formación geológica tomando como base el concepto de mapas de calor, el cual es una representación gráfica de los valores contenidos en una matriz mediante el uso de colores [22]. Los ejes horizontales y verticales tomarán las variables de entrada del modelo, mientras que el color utilizado dentro del mapa cada representará las variables de salida ROP o MSE.

Para la creación de estos mapas se generó un dataset con las variables de entrada del modelo para cada sección de estudio, teniendo en cuenta los rangos operacionales asociados a cada una (Inciso 2.2.1., Tablas 10 y 11)., generando todas las combinaciones posibles entre estas. Estas bases de datos generadas se introducen en los respectivos modelos a fin de predecir la ROP y la MSE para cada registro (Tabla 7).

Tabla 7.

Rangos de parámetros para la generación de mapas de calor

Parámetros	**Sección 8 ½''**			**Sección 12 ¼''''**		
	Min	**Max**	**Paso**	**Min**	**Max**	**Paso**
RPM_SURF	0	80	84	0	80	4
RPM	100	250	10	150	300	10
WOB	0	40	2	0	42	2
Q	300	750	30	700	1200	30
TQM	0	28	2	0	32	2
Formaciones	Real, La cira shale, Colorado, Mugrosa, La paz C, La paz CG			Real, La cira shale, Colorado, Mugrosa		

***Nota*:** La figura muestra los rangos de parámetros a introducir en el modelo predictivo con el fin de generar los mapas de calor para las variables ROP y MSE por formación geológica.

2.6.Acotación y análisis de los mapas de parámetros

Los mapas de parámetros realizados permitirán acotar los rangos óptimos de los parámetros de perforación por formación geológica para optimizar las variables de salida ROP y MSE. Éstas serán correlacionadas con la información obtenida en las matrices del inciso 2.2.1., a fin de verificar y establecer los parámetros con mayor incidencia y sus respectivos rangos para optimizar las variables objetivo en cada una de las formaciones en estudio.

3. RESULTADOS Y ANÁLISIS

El presente Capitulo muestra los resultados del proceso de entrenamiento del modelo, implementación y estimación de parámetros óptimos de operación. Lo anterior en el marco de la metodología y los datos presentados en la sección anterior.

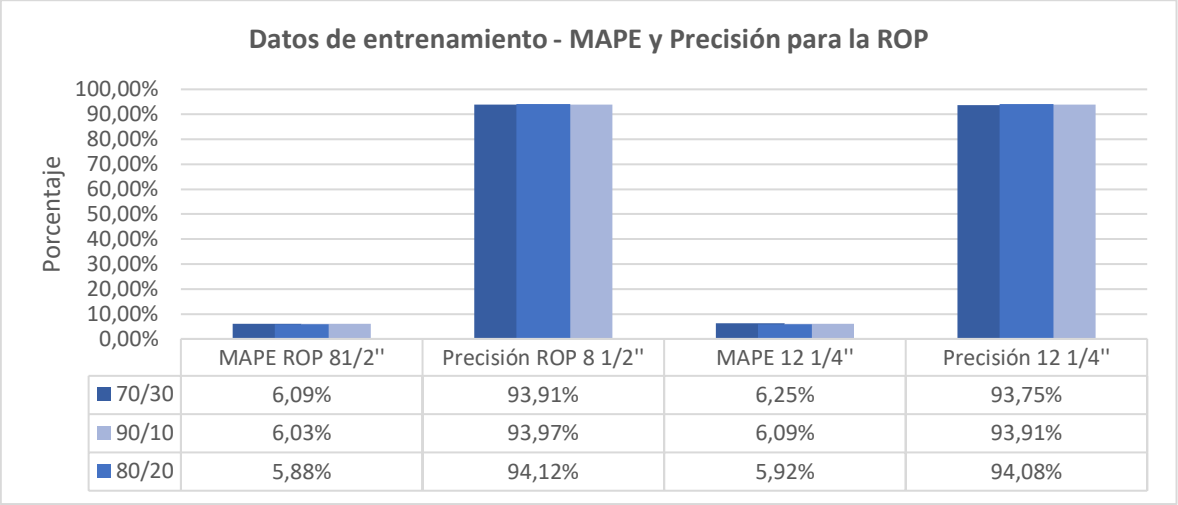
Una vez se termina de programar y ejecutar el código, se procederá a la interpretación de los resultados arrojados por el modelo predictivo. Inicialmente se verificó la precisión con los hiperparámetros que vienen por defecto en el algoritmo de programación *Random Forest Regressor*, para establecer, cuál es la mejor división de los datos de entrenamiento y prueba para el modelo. Con la división óptima de la base de datos es realizado el ajuste de hiperparametros del algoritmo con el fin de optimizar los resultados obtenidos. Una vez realizado el paso anterior se estableció la importancia de las variables en el desarrollo del algoritmo y se generaron los mapas de calor de los parámetros de entrada por formación geológica para la acotación y estimación de la ROP y MSE con su análisis respectivo.

A manera de ilustración en los Anexos 3, 4, 5 y 6 se encuentran árboles de decisión de cada uno de los modelos predictivos.

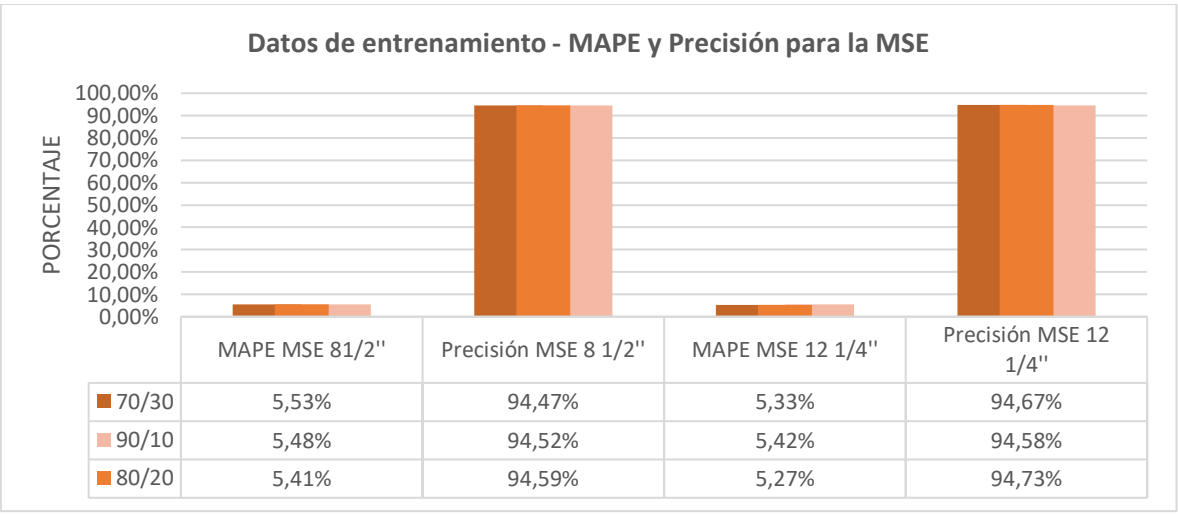
3.1. Entrenamiento y ajuste del modelo

El proceso de entrenamiento de un modelo de aprendizaje automático consiste en destinar un porcentaje de la base de datos a partir del cual el algoritmo detecta patrones en la información para poder realizar predicciones sobre las variables deseadas, en este caso para la ROP y la MSE, con el fin de que posteriormente el modelo pueda ser utilizado en futuras estimaciones. Para las secciones y variables en estudio, se obtuvieron las siguientes métricas de desempeño en sobre los datos de entrenamiento:

Figura 19.
Resultados del entrenamiento del modelo



	MAPE ROP 81/2''	Precisión ROP 8 1/2''	MAPE 12 1/4''	Precisión 12 1/4''
70/30	6,09%	93,91%	6,25%	93,75%
90/10	6,03%	93,97%	6,09%	93,91%
80/20	5,88%	94,12%	5,92%	94,08%



	MAPE MSE 81/2''	Precisión MSE 8 1/2''	MAPE MSE 12 1/4''	Precisión MSE 12 1/4''
70/30	5,53%	94,47%	5,33%	94,67%
90/10	5,48%	94,52%	5,42%	94,58%
80/20	5,41%	94,59%	5,27%	94,73%

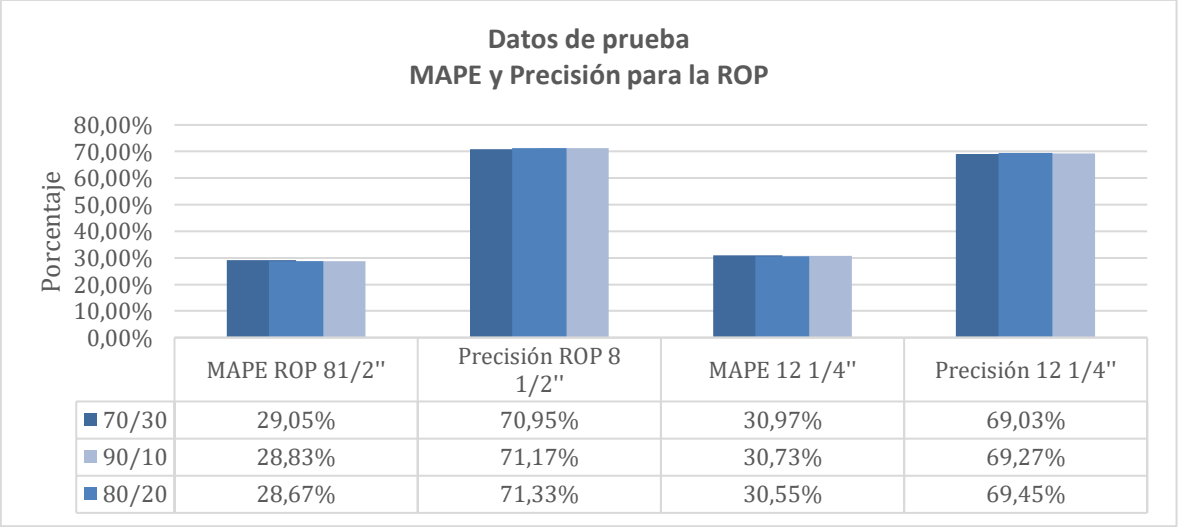
***Nota*:** La Figura muestra el MAPE y la precisión para el entrenamiento del modelo en los escenarios de división de datos de entrenamiento y prueba 70/30, 80/20 y 90/10 para la ROP y la MSE en las secciones en estudio.

Se puede evidenciar un desempeño por encima del 90% para las secciones y variables en estudio, observándose mayores precisiones en la división de datos 80/20, con errores del 5.88% y 5.92% en la estimación de la ROP y de 5.41% y 5.27% para la MSE en las secciones 8 ½'' y 12 ¼'' respectivamente. Con respecto a las divisiones de 70/30 y 90/10 se obtuvieron diferencias menores al 1% con respecto al error mínimo observado. Estos valores se deben principalmente a que en este proceso se entrena y prueba el modelo con los mismos datos, buscando que el algoritmo identifique patrones en la información para que sea capaz de brindar altos desempeños con nueva información.

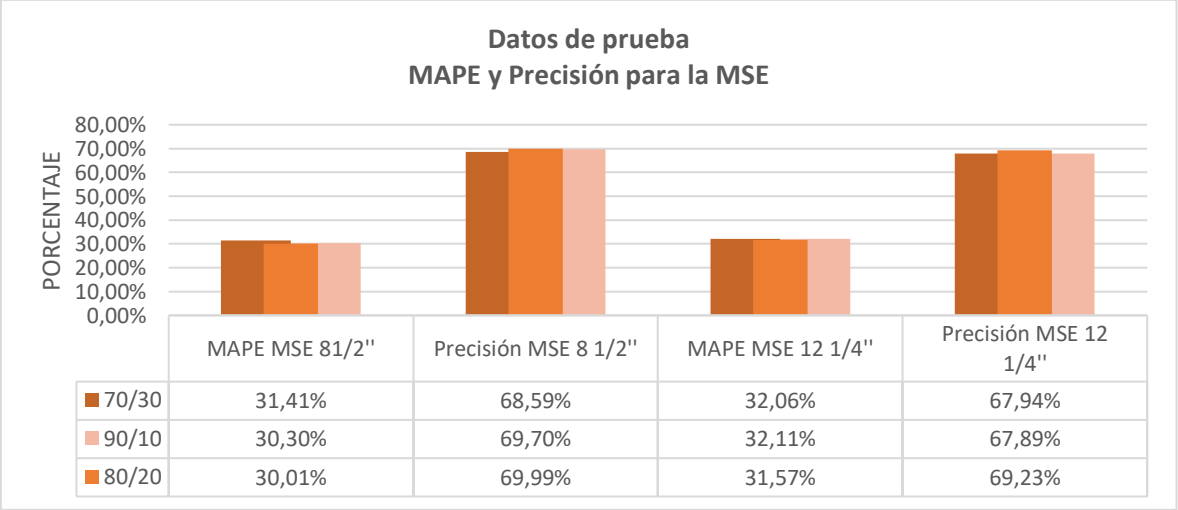
De esta manera, altas precisiones con los datos de entrenamiento indican un buen aprendizaje del algoritmo predictivo. [34]

Para estimar la precisión del modelo predictivo con los datos de prueba se ejecutaron diferentes tipos de pruebas. La primera corriendo el algoritmo *Random Forest Regressor* con los hiperparámetros que vienen por defecto en el código (100 arboles de decisión para cada variable en estudio y sección), en donde se dividió el data set en porcentajes de entrenamiento y prueba, obteniendo resultados de MAPE y precisión de la ROP y MSE para cada sección estudiada como se muestra en la Figura 20.

Figura 20.
Resultados de la prueba del modelo



	MAPE ROP 81/2''	Precisión ROP 8 1/2''	MAPE 12 1/4''	Precisión 12 1/4''
70/30	29,05%	70,95%	30,97%	69,03%
90/10	28,83%	71,17%	30,73%	69,27%
80/20	28,67%	71,33%	30,55%	69,45%



	MAPE MSE 81/2''	Precisión MSE 8 1/2''	MAPE MSE 12 1/4''	Precisión MSE 12 1/4''
70/30	31,41%	68,59%	32,06%	67,94%
90/10	30,30%	69,70%	32,11%	67,89%
80/20	30,01%	69,99%	31,57%	69,23%

Nota: La Figura muestra el MAPE y la precisión para la prueba del modelo en los escenarios de división de datos de entrenamiento y prueba 70/30, 80/20 y 90/10 para la ROP y la MSE en las secciones en estudio.

Con los hiperparámetros que vienen por defecto y la división óptima de datos en entrenamiento y prueba (80/20), el error porcentual absoluto medio (MAPE) estimado por el algoritmo fue de 28,67% y 30,55% para la ROP y 30% y 31,57% para la MSE y la precisión del modelo fue de 71,33% y 69,45% para la ROP y 69,99% y 69,23% para la MSE, para las secciones 8 ½ y 12 ¼ respectivamente. El modelo predictivo será siendo considerado aceptable con valores de precisión superiores al 70%..

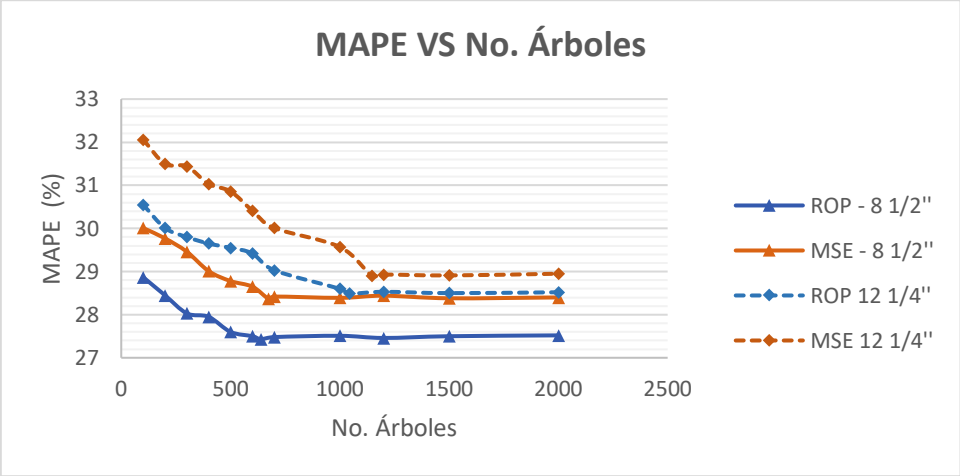
Para las demás pruebas, se observó que el mayor MAPE correspondió para la distribución de 70-30 alcanzó valores entre 29,05% y 30,97% para la ROP y 31,41% y 30,97% para la MSE de las secciones en estudio, lo cuál es inversamente proporcional a la precisión del modelo, obteniendo la precisión 70,95% y 69,03% para la ROP y 68,59% y 67,94% para la MSE, en las secciones 8 ½ y 12 ¼ .

Como puede observarse en las figuras anteriores, existe una mayor precisión en la predicción de la ROP para ambos modelos. Esto se debe principalmente a que este parámetro presenta menor variabilidad en comparación a la MSE, lo cual afecta directamente al modelo predictivo empleado. Adicionalmente, la mejor métrica de desempeño evidenciada en la implementación del algoritmo, corresponde a la división de datos de entrenamiento y prueba 80/20, respectivamente. Este porcentaje será utilizado para el ajuste de los hiper-parámetros del algoritmo predictivo.

El comportamiento del MAPE relaciona las variables totales acertadas con respecto al número de predicciones realizadas, tal como se explico en la sección 2.4 del capitulo de metodología y datos del presente trabajo, entre más alto sea el valor del error medio porcentual absoluto, más baja era la precisión del modelo trabajado.

Una vez establecido el mejor porcentaje de distribución entrenamiento prueba (80-20) se procede a ajustar los hiperparámetros del modelo. Como se explicó en la metodología, este paso consiste principalmente en estimar el número óptimo de árboles de decisión a ejecutar por el algoritmo. Para ello calculamos el MAPE con diferentes números de árboles de decisión para la distribución 80-20. Los resultados de las pruebas son mostrados en la Figura 21.

Figura 21.
Error Absoluto Porcentual Medio con respecto al número de árboles de decisión.



Nota: La figura muestra cómo el modelo disminuye su error al aumentar el número de árboles de decisión entre 600 – 1000, sin embargo, al intentar valores superiores a los mencionados, no se observa una mayor optimización en el proceso iterativo.

Se puede concluir que después de 600 árboles de decisión el MAPE para las sección 8 ½'' tiene una tendencia casi plana, esto debe principalmente a que para los pozos tipo S se tiene mayor información (tanto en número de pozos como en profundidad), razón por la cual, se logra estandarizar el modelo con menos árboles de decisión. Para la sección 12 ¼ la curva se tarda un poco más en normalizar, aproximadamente en 1000 arboles de decisión , es importante aclarar que entre mas información tenga el modelo más efectiva será esta normalización [24].

Con estos resultados se ajusta el número de árboles (n_estimators) para cada uno de los modelos por sección de estudio, pasando de 100 arboles de decisión a 673 y 1044 para la ROP y de 639 a 1147 para la MSE en las secciones 8 ½ y 12 ¼ respectivamente

Una vez son ajustados los hiperparámetros para los modelos predictivos, se obtiene un modelo de aprendizaje con mejores métricas de desempeño para la estimación de los parámetros en estudio (Tabla 8).

Tabla 8.
Métricas de desempeño finales de los modelos de aprendizaje

Dataset	ROP			MSE		
	MAPE	Precisión	Mejora	MAPE	Precisión	Mejora
Pozos Tipo S -Sección 8 ½''	27,43%	72,57%	1,24%	28.37%	71.63%	1.64%
Pozos H - Sección 12 ¼''	28,49%	71,51%	2,06%	28,90%	71,1%	1,87%

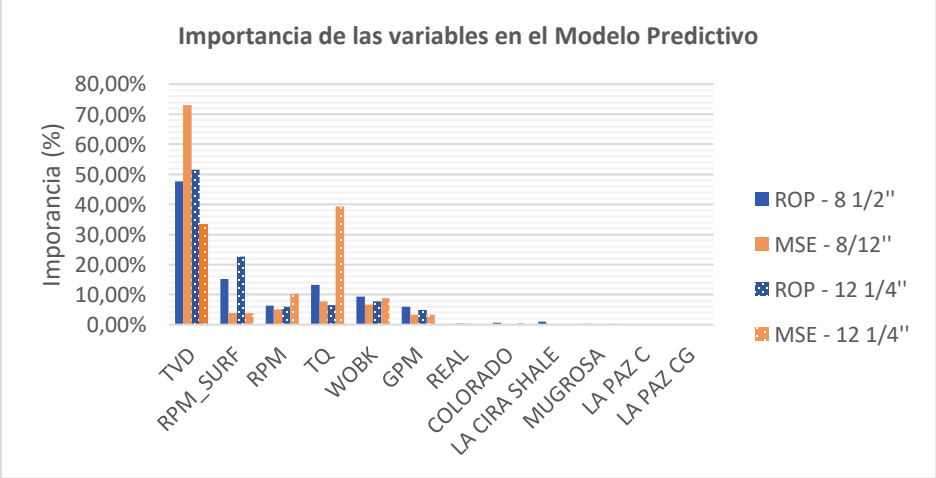
Nota: Optimización del MAPE para los modelos predictivos con el ajuste de hiperparametros. La mejora oscila entre un 1,24% y un 2,06%.

El ajuste de los hiper-parámetros en el modelo es realizado también para evitar el efecto de overfitting o sobre-ajuste de los datos, en el cual el algoritmo aprenderá solamente en casos particulares. Esto quiere decir que solamente reconocerá los datos de entrenamiento y no los nuevos datos de entrada, afectando la efectividad de la predicción [35]. Al realizar el ajuste de los arboles de decisión (n_estimators) se obtendrá una mayor de cantidad de combinaciones posibles y disminuirá notablemente la probabilidad de caer en el efecto de overfitting, mejorando la veracidad y el rendimiento del modelo predictivo [22].

3.1.1. Importancia de las variables

Seguido a esto, se estableció la incidencia de las variables de entrada para cada algoritmo predictivo. Dichos resultados pueden observarse en la Figura 22.

Figura 22.
Importancia de las variables en los modelos predictivos



Nota: La figura muestra importancias de las variables de entrada para cada uno de los modelos predictivos.

Estas gráficas permiten evidenciar que la variable profundidad vertical verdadera, tiene la mayor incidencia en la predicción de la ROP y la MSE, seguido de las revoluciones por minuto y el torque. Es válido aclara que estas importancias son las establecidas por el algoritmo predictivo, para la generación de los nodos en los diversos árboles de decisión.

Comparando los resultados obtenidos con las matrices de correlación establecidas en el inciso 2.4 de Metodología y Datos (Figura 14), donde para la sección 8 ½'' las variables con mayor relación con la ROP y la MSE son la Profundidad vertical verdadera, con correlaciones de -0.43 y 0.71, seguidos de las Revoluciones por minuto y el Torque con valores de 0.19 y 0.54 respectivamente. Para la sección 12 ¼'' la mayor relación sobre las variables en estudio se evidencia en las Revoluciones por minuto de superficie y el torque con un grado de correlación de 0.49 y 0.57, posteriormente se encuentran la profundidad vertical verdadera y las revoluciones por minuto de fondo, con valores de -0.38 y 0.42.

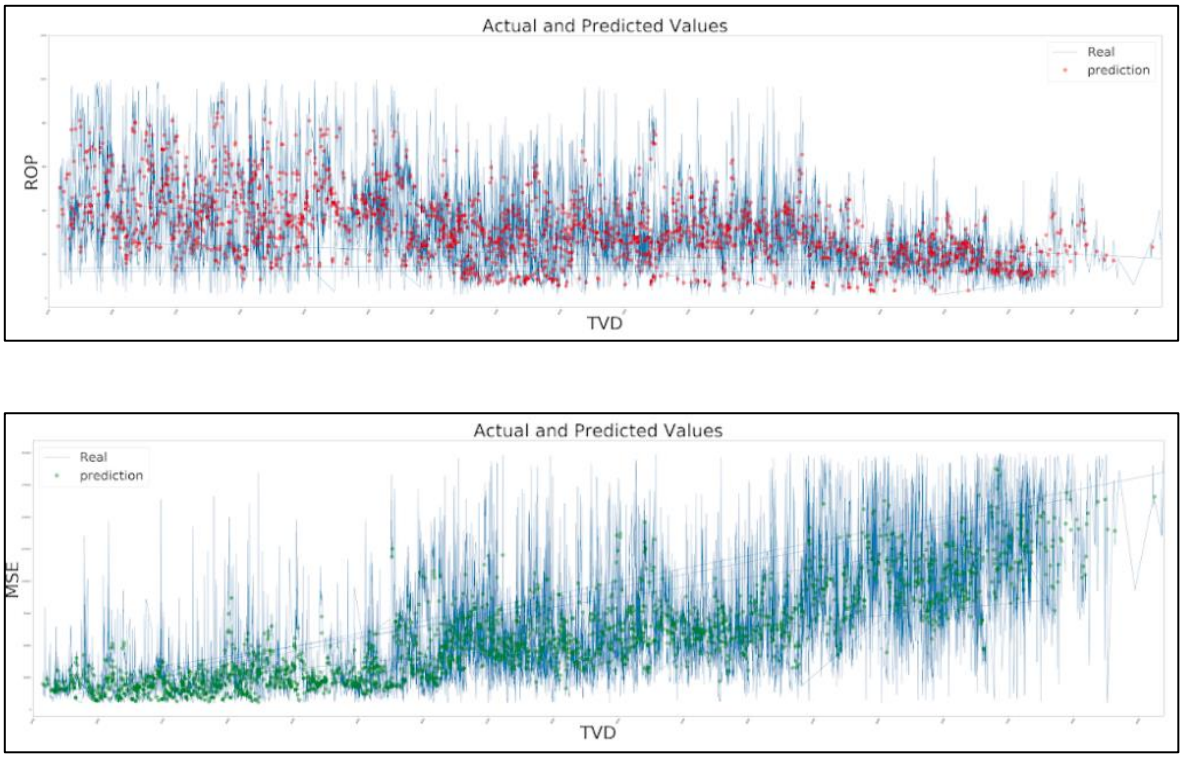
Como se menciono anteriormente la profundad vertical verdadera es la variable que tiene una mayor incidencia en el modelo, ya que a mayor profundidad, existirá más complejidad al momento de perforar, esto debido principalmente a la intercalación de unidades geológicas, generando un aumento de energía mecánica especifica y variaciones en la tasa de penetración [3]. Seguido a la TVD las variables con mayor incidencia en el modelo son las revoluciones por minuto y el torque, las cuales tienen una relación directamente proporcional con la energía mecánica especifica, como se evidencia en la Ecuación 1.5.2 del marco teórico , esto quiere decir que un aumento abrupto de las RPM generará mayores torques, incrementando la energía mecánica específica y afectando directamente la eficiencia de la operación [19]. Era de esperarse que las variables con una mayor incidencia en el modelo fueran las anteriormente descritas, ya que son parámetros operacionales, por lo que tienen una relación más estrecha con las variables en estudio, caso contrario a las formaciones geológicas, que como se puede observar en la gráfica tienen muy poca importancia en relación a los parámetros de perforación, debido a que fueron tenidas en cuenta como variables con el único fin de generar los mapas de calor [30] .

3.2. Validación del modelo predictivo

Con el 20% de los datos destinada para la prueba del algoritmo se realizó la validación del modelo predictivo, se elaboraron las gráficas correspondientes por sección, teniendo en cuenta la profundidad

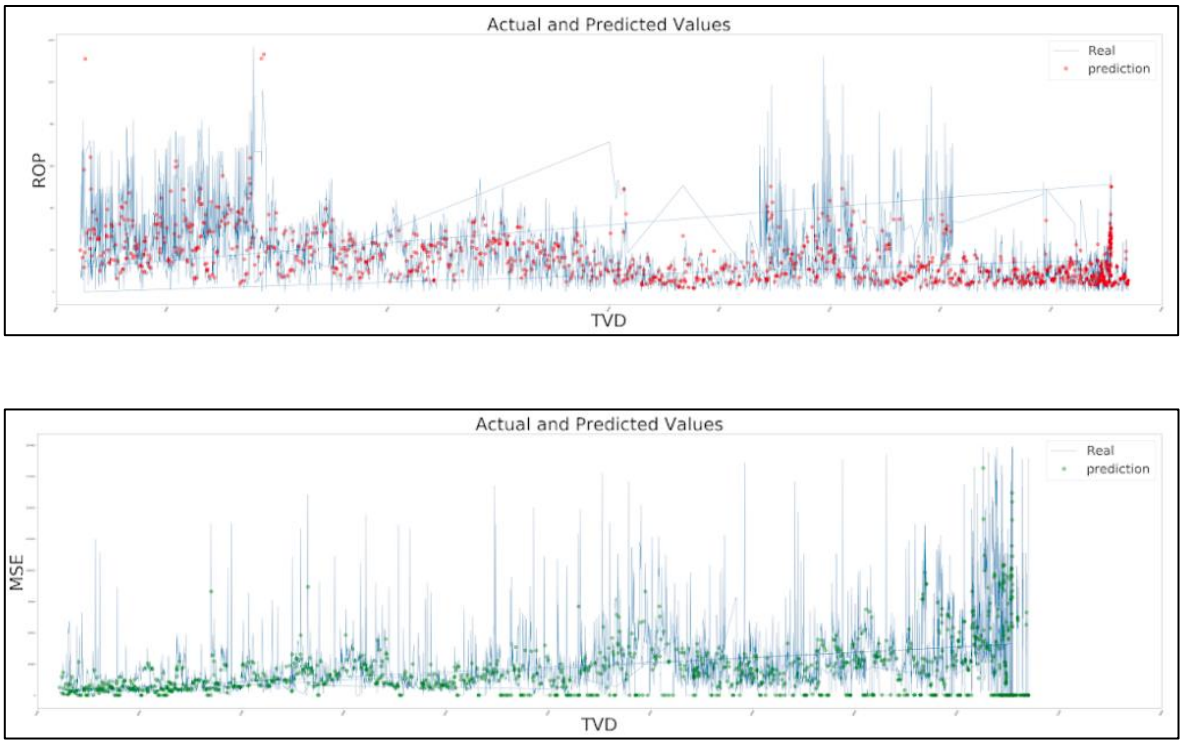
del modelo (TVD) y las variables que se buscan predecir (ROP y MSE), como se muestra a continuación:

Figura 23.
Validación del modelo predictivo para la ROP y la MSE de la sección 8 ½ “



Nota: Las gráficas expuestas muestran el comportamiento de los datos reales (medidos en campo) representados con la línea azul, y las predicciones realizadas por el algoritmo con los puntos rojos para la ROP y verdes para la MSE con respecto a la profundidad para la sección en estudio. Se puede observar la relación inversa entre ambas variables.

Figura 24.
Validación del modelo predictivo para la ROP y la MSE de la sección 12 ¼''



Nota: Las gráficas expuestas muestran el comportamiento de los datos reales (medidos en campo) representados con la línea azul, y las predicciones realizadas por el algoritmo con los puntos rojos para la ROP y verdes para la MSE con respecto a la profundidad para la sección en estudio. Se puede observar la relación inversa entre ambas variables.

3.3 Estimación de los parámetros óptimos para las secciones en estudio

Como se mencionó en la sección 2.5 de la metodología, se generaron mapas de calor para cada una de las secciones en estudio. Se procederá a presentar y analizar cada uno de los mapas obtenidos por el modelo, para así poder determinar y establecer los valores óptimos para la ROP y la MSE, durante las actividades de perforación.

Tanto para las secciones 8 ½" y 12 ¼" se generaron los siguientes mapas de calor:

- Torque vs. Peso sobre la broca
- Revoluciones por minuto en superficie vs peso sobre la broca
- Revoluciones por minuto en superficie vs torque
- Revoluciones por minuto vs galones por minuto

Tabla 9.

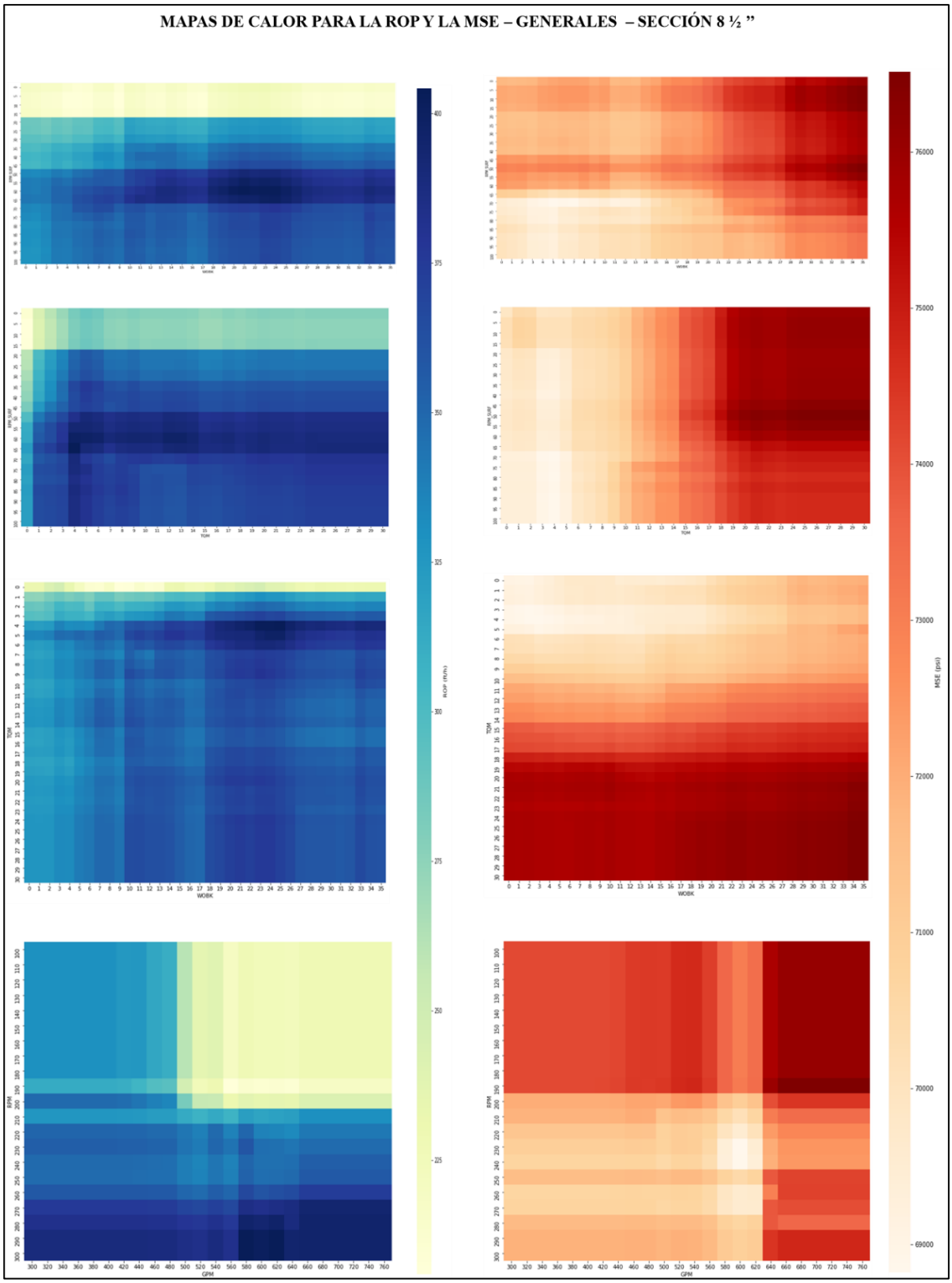
Abreviaciones y unidades de las variables utilizadas en los mapas de calor.

Variable		Unidades
TQM	Torque	klb
RPM_SURF	Revoluciones por minuto de superficie	rev/min
RPM	Revoluciones por minuto de fondo	rev/min
GPM	Galonaje	gal/min
WOB	Peso sobre la broca	klb
ROP	Rata de penetración	ft/h
MSE	Energía mecanica especifica	psi

Nota: Variables y unidades para la generación de mapas de calor.

Los mapas de calor fueron elaborados de manera general para cada una de las secciones en estudio, los cuales se presentan a continuación, así como para cada una de las formaciones geológicas perforadas, los cuales se encuentran en los Anexos 7-16 de este documento.

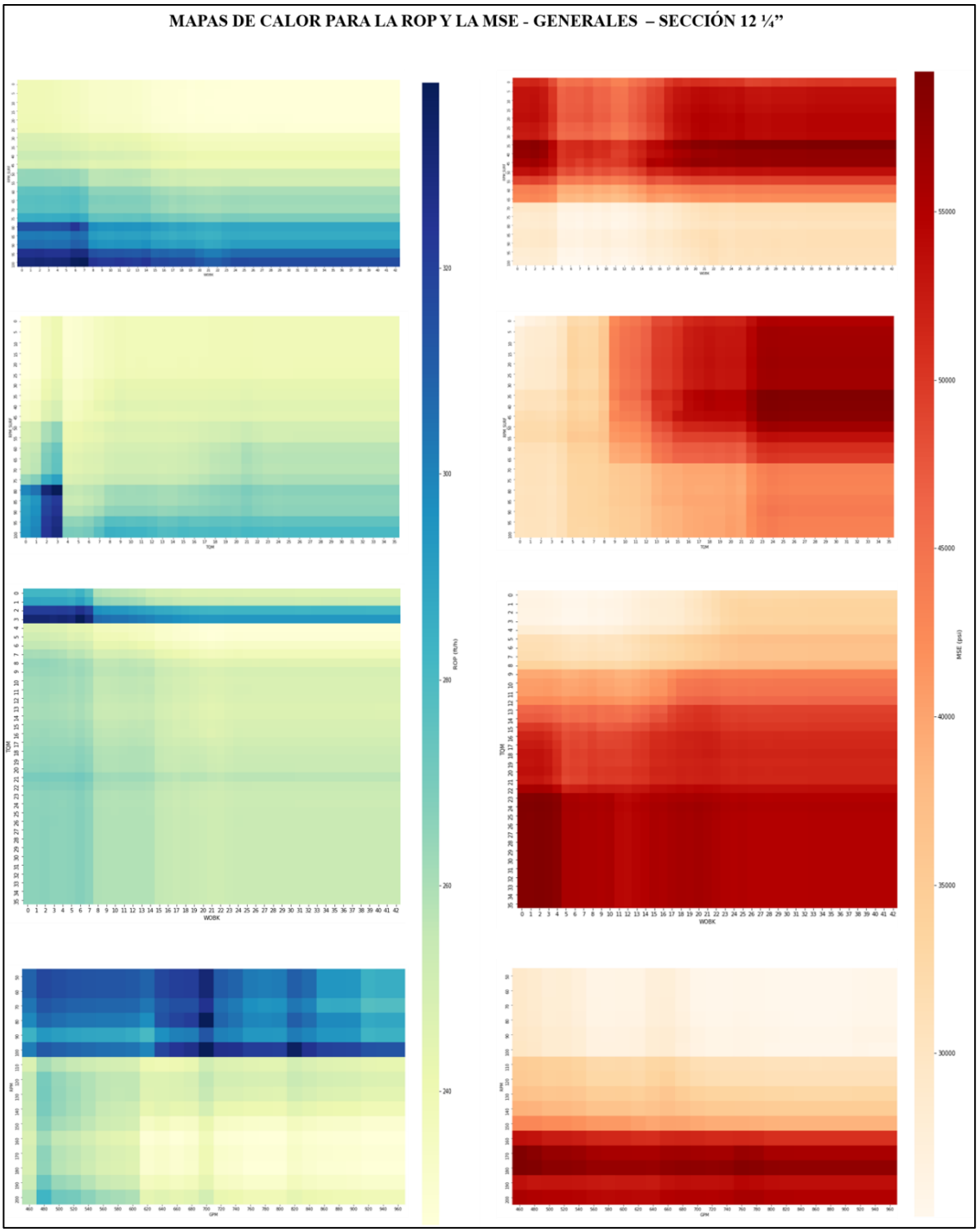
Figura 25.
Mapas de calor generales para la ROP y la MSE en la sección 8 ½’’



Nota: La figura muestra la izquierda los mapas de calor para la ROP y a la derecha para la MSE para la sección 8 ½’’.

Figura 26.

Mapas de calor generales para la ROP y la MSE en la sección 12 ¼’’'



Nota: La figura muestra la izquierda los mapas de calor para la ROP y a la derecha para la MSE para la sección 12 ¼’’.

Los mapas de calor para la ROP tienen un rango de lo 220-400 y 230-330 pies/hora, para la MSE de 68000-77000 y 29000-56000 psi para las secciones 8 ½'' y 12 ¼'' respectivamente. De manera general se puede observar la relación inversa entre ambos parámetros, al apreciarse colore azules en el mapa de la ROP (valores elevados), en los de la MSE se verán colores claros (valores menores), conservando la relación inversa entre las mismas establecida a partir de la Ecuación 1.

3.3.1. Análisis de resultados obtenidos en los mapas de calor

Una vez obtenidos los mapas de calor para las secciones en estudio, se realiza un análisis comparativo para ambos mapas (ROP y MSE) por formación geológica, acotando los rangos de las variables de entrada para obtener valores operacionales óptimos predichos por el modelo para la ROP y la MSE. A partir de estos valores se procederá a interpretar y analizar los rangos de los parámetros establecidos por el modelo para cada una de las variables y formaciones perforadas en las secciones en estudio.

Las matrices de correlación establecidas para cada una de las formaciones geológicas (Anexos 1 y 2) permiten establecer que para la sección 8 ½'' el torque es la variable con mayor incidencia sobre la ROP y MSE en las formaciones Real, La Cira Shale y La Paz CG con grados de correlación entre 0.39 y 0.45, mientras que para las formaciones Colorado, Mugrosa y La Paz C las revoluciones por minuto de fondo toman mayor relevancia con valores superiores a 0.4. Para la sección 12 ¼'' el caudal es el parámetro con mayor relación sobre las variables en estudio en la formación Real con una correlación de 0.23, y las revoluciones por minuto en superficie tienen mayor peso en Colorado, La Cira Shale y Mugrosa, con valores entre 0.45 y 0.65.

El caudal es un parámetro que se mantiene en un rango establecido para cada sección, con valores de 530-740 y 650-750 para las secciones 8 ½'' y 12 ¼'', a excepción de la formación La Paz – CG en los pozos Tipo S, ya que al ser productora debe perforarse con parámetros controlados para evitar fracturas en la misma [26], siendo la última sección perforada en los pozos con esta geometría, mientras que para los pozos horizontales será realizada en una nueva sección que corresponde a la geonavegación en la formación Mugrosa [4].

La formación Real es la primera presente en la columna litológica a perforar para ambas secciones. Posee un espesor promedio perforado de 1450 pies (Figura 16). Los rangos de las revoluciones por minuto son de 45-65 y 75-95 en superficie, y de 265-275 y 40-100 en fondo para las secciones 8 ½ y 12 ¼ respectivamente. La velocidad de rotación es inversamente proporcional a la resistencia de compresión de la roca, lo que afecta y cambia directamente el comportamiento de este parámetro. Para esta formación los valores de RPM son proporcionales al caudal utilizado al momento de perforar, cuando el caudal presenta un valor superior al planeado las RPM tienden a aumentar [26], la geología

de esta formación se basa principalmente en areniscas en capas delgadas con intercalaciones de arcillitas abigarradas [36], la cuales son formaciones poco consolidadas donde el valor optimo a utilizar de torque es entre 1-10 klbs-f y WOB entre 1- 15 klbs respectivamente.

El modelo predijo valores entre 360 – 750 para la ROP con los rangos de los parámetros mencionados anteriormente, sin embargo se debe prestar atención a no realizar cambios brucos que generen un incremento desproporcional en la ROP, ya que se puede generar el influjo de fluidos no deseados hacia pozo [19].

Por otra parte los valores arrojados de MSE están entre 15.000 y 37.500 (son los más bajos de todo el modelo). Esto debido a que es el inicio de la perforación para ambas secciones y debido a la poca profundidad no se genera un alto valor de energía mecánica especifica, el cual aumenta proporcionalmente con la profundidad [37].

La siguiente formación es La Cira Shale con un espesor aproximado de 1950 pies. En este punto los rangos operacionales de las revoluciones por minuto de fondo, de superficie y el torque incrementan a 70-100 , 40-300, y 2-9 respectivamente, debido a la complejidad de la formación y presencia de arcillas [38], razón por la cual se necesita un mayor esfuerzo por parte de los equipos al momento de perforar. Por consiguiente se puede apreciar una disminución en los rangos de la ROP para esta formación, que se encuentran entre 285-550 ft/h y a su vez un aumento en la energía mecánica especifica a 15000 – 59700 kpsi en los valores predichos por el modelo.

Se debe tener en cuenta al momento de la perforación no exceder los rangos operacionales para no recalentar la broca y aumentar su desgaste, también al existir la presencia de arcillas se debe tener cuidado al momento de perforar para no fracturar la formación, generando pérdidas de circulación [37].

La formación colorado tiene un espesor promedio de 1440 pies. Esta no representa una formación de interés para la producción de hidrocarburos debido a que por la geología consta principalmente de arcillas con intercalaciones de arenitas [36]. Las RPM de superficie arrojadas por el modelo varían en un rango entre 65-100 en esta sección debido a las condiciones litológicas, se debe tener especial cuidado con las pegas de tubería, al ser una formación con una gran presencia de arcillas las pegas y el embotamiento de la broca pueden ocurrir causando un aumento a la energía mecánica especifica que se encuentra en un rango operacional entre 20000 – 50000 psi , lo cual podría repercutir en retrasos en la operación , aumento de tiempos no productivos y costos. El torque y el peso sobre la broca se encuentran en un rango promedio de 7-23 klb-f y 2-15 klbs respectivamente.
Para esta formación el modelo arrojó como resultado una ROP óptima entre 400 – 505 ft/hr y 280 530 ft/hr para la sección 8 ½ y 12 ¼ respectivamente.

Seguido a esto se encuentra la formación Mugrosa, la cual posee un espesor promedio de 2880 pies y está compuesta de arenitas de grano fino a medio [36]. Esto la convierte en una formación atractiva para la acumulación de hidrocarburos, siendo el target de la sección 12 ¼, por lo cual es importante controlar los parámetros operacionales para no fracturar la formación y dañar una probable reserva. Los rangos operacionales arrojados por el modelo para la ROP fueron de 250 – 550 ft/hora, sin embargo, es necesario tener una broca de perforación adecuada para a las características geológicas de la formación, debido a que si se escoge una barrena errónea es posible que la ROP sea ineficiente y la perforación tome más tiempo de lo esperado, también se debe tener en cuenta en este punto la limpieza del pozo, el cual afecta directamente la velocidad de la operación [37].

El peso sobre la broca recomendado para la formación mugrosa se encuentra en un rango entre 1-20 klbs, es importante no exceder este valor para evitar fracturas y filtraciones indeseadas.

La energía mecánica especifica varia en cada sección, para la sección 12 ¼ se encuentra entre 25000 y 45000 psi y para la 8 ½ entre 45000 – 72500 psi respectivamente, la gran diferencia se debe a que como se mencionó anteriormente para la sección 12 ¼ esta formación es de interés mientras que para la sección 8 ½ no , lo cual hace que para la segunda sección mencionada se perfore de una manera más rápida y eficiente.

Finalmente la formación La paz (dividida en los mapas de calor en las unidades C y CG con un espesor estimado de 1420 y 1930 pies respectivamente) es el target de la sección 8 ½ , en el modelo se puede observar que los rangos operacionales disminuyen notablemente debido a que se debe tener cuidado al momento de perforar para no fracturar la formación y dañar las posibles reservas de hidrocarburos, ya que existe un riesgo de quebrantar la formación productora provocando que el contacto de agua petróleo se filtre directamente al pozo, por ello se opta por perforar controlando los parámetros [20]. De esta manera el modelo arrojo los rango operacionales para las variables teniendo en cuenta lo anterior dando los resultados mostrados en las Tablas 10 y 11.

Tabla 10.
Resultados de rangos de parámetros operacionales para la sección 8 ½”

Formación Geológica	RANGOS DE PARAMETROS OPERACIONALES PARA LA SECCIÓN 8 1/2"						
	RPM SUPERFICIE	RPM FONDO	TORQUE	WOBK	CAUDAL	ROP ESPERADA	MSE ESPERADA
REAL	45 - 65	265 - 275	4 -10	7 -15	530-590	460 - 560	20000 - 37500
LA CIRA SHALE	75 - 95	270 - 300	2 - 8	23 - 35	570-630	370 - 410	43000 - 57900
COLORADO	65 - 95	285 - 300	8 - 10	1 - 10	630-740	400 - 505	37000 - 50000
MUGROSA	75 - 95	285 - 300	4 - 8	4 - 21	570-650	375 - 430	45000 - 72500
LA PAZ C	25 - 65	190 - 200	2 - 20	11 - 35	660-740	250 - 280	97000 - 108000
LA PAZ CG	30 - 55	190 - 200	12 - 20	11 - 23	320-420	240 - 275	95000 - 115000

Nota: Rangos operacionales para una ROP y MSE óptimas en la sección 8 ½’’.

Tabla 11.
Resultados de rangos de parámetros operacionales para la sección 12 ¼ ”

Formación Geológica	RANGOS DE PARAMETROS OPERACIONALES PARA LA SECCIÓN 12 1/4"						
	RPM SUPERFICIE	RPM FONDO	TORQUE	WOBK	CAUDAL	ROP ESPERADA	MSE ESPERADA
REAL	75 - 95	40 - 100	1 - 4	1 - 13	660-700	360 - 750	15000 - 25000
LA CIRA SHALE	70 - 100	40 - 100	1 - 4	1 - 9	650-710	285 - 550	15000 - 25000
COLORADO	80 - 100	40 - 100	1 - 4	1 - 20	690-730	280 - 530	20000 - 29000
MUGROSA	70 - 100	40 - 100	1 - 4	1 - 7	690-750	275 - 550	25000 - 41000

Nota: Rangos operacionales para una ROP y MSE óptimas en la sección 12 ¼’’.

4. CONCLUSIONES

El Análisis Exploratorio de Datos (EDA) permite obtener una descripción estadística de las variables en estudio con el fin de establecer los rangos operacionales de los parámetros de perforación, con valores de 0-100 revoluciones por minuto en superficie, 100-300 y 60-250 revoluciones por minuto en fondo, 0-32 y 0-42 klbs-f de torque, 0-35 y 0-42 klbs de peso sobre la broca, 400-700 y 700-1500 galones por minuto de caudal en las secciones 8 ½'' y 12 ¼'' respectivamente, identificando así su distribución, valores atípicos, entre otros.

Las métricas de desempeño sobre los datos de entrenamiento del modelo son mejores comparadas a las de prueba, obteniendo resultados superiores al 90% y 70% respectivamente, ya que en el primer proceso se entrena y prueba el modelo con la misma data, buscando que el algoritmo identifique patrones en los datos para que sea capaz de reconocer nueva información.

La mejor división para el data set fue: 80% de los datos para entrenamiento y de 20% para la validación del modelo predictivo, debido a que con esta segmentación se obtuvieron los valores de error medio porcentual absoluto más bajos en el modelo, siendo este un 28,67% y 30,55% para la ROP y 30% y 31,57% para la MSE en las secciones 8 1⁄2'' y 12 1⁄4'' respectivamente.

El algoritmo de regresión supervisado Random Forest Regressor presenta altos desempeños (superiores al 70%) en la predicción de la ROP y la MSE debido a que su funcionamiento se basa en un comportamiento no lineal de las variables de entrada, siendo apto para los parámetros de perforación.

Se evidencio que luego de realizar el proceso de ajuste de hiperparámetros en el modelo se obtiene una mejora en el rendimiento del mismo, aumentando el número de árboles de decisión, se obtiene una mejoría entre 1,24% y 2,06% para las variables y secciones en estudio, sin embargo, llega a un punto en el cual se estabiliza la curva de mejora, (Figura 21), donde no vale la pena seguir aumentando el número de árboles de decisión debido a que no se obtiene mejora significativa.

Se determinó que al ejecutar el algoritmo Random Forest Regressor, teniendo en cuenta las formaciones geológicas y los parámetros de superficie como, la profundidad vertical verdadera, el torque, el peso sobre la broca y las revoluciones por minuto de superficie y fondo, presenta

un gran desempeño al momento de la predicción de las variables en estudio, logrando una precisión 72,57% y 71,51% para la ROP y 71,63% y 71,1% para las MSE en las secciones 8 1⁄2'' y 12 1⁄4'' del campo en estudio, en donde se obtuvieron mapas de calor para cada una de las secciones obteniendo los mejores rangos operacionales para cada una de las secciones presentes en el pozo.

La Tasa de penetración y la energía mecánica especifica tienen una relación inversa como se puede evidenciar en la Ecuación 1, esto puede ser corroborado en los mapas de calor proporcionados por el modelo predictivo, en donde a mayor tasa de penetración, se puede observar un poco uso de la energía mecánica específica, dando como resultado los óptimos rangos operacionales de ROP y MSE para cada una de las formaciones en estudio.

La profundidad vertical verdadera, el torque y las revoluciones por minuto de superficie son las variables con mayor incidencia sobre el modelo predictivo, con valores promedio de importancia de 70%, 35% y 22% respectivamente, caso contrario con las formaciones que tienen una baja incidencia (menor al 5%), esto se debe a que fueron tomadas como variables categóricas en el desarrollo del presente trabajo.

La acotación de parámetros de perforación teniendo en cuenta la caracterización de las formaciones geológicas permite un desarrollo eficiente de la actividad disminuyendo problemas operacionales como el embotamiento de la broca, el fracturamiento de arcillas, las pegas de tubería, entre otros, logrando reducciones en costos y tiempos no productivos cerca de un 20%.

BIBLIOGRAFÍA

[1] Management Solutions, «Machine Learning, una pieza clave en la transformación de los modelos de negocio,» Management Solutions, pp. 19,20, 2018.

[2] A. N. M. M. J. A. Luis Barbosa, «Machine learning methods applied to drilling rate of penetration prediction and optimization - A review,» Journal of Petroleum Science and Engineering, vol. 183, 2019.

[3] H. D. H. M. K. G. Chiranth Hegde, «Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models,» Journal of Petroleum Science and Engineering, vol. 159, pp. 259-306, 2017.

[4] Ecopetrol S.A. .

[5] V. M. Sebastian Raschka, Python Machine Learning: Aprendizaje automático y aprendizaje profundo con Python, ScikitLearn y Tensorflow, España: Marcombo, 2019.

[6] E. M. y. A. R. Rafael Caballero, Big data con Python: Recolección, almacenamiento y preproceso, Bogotá D.C.: Alfaomega, 2019.

[7] J. Luna, «Medium,» 8 Febrero 2018. [En línea]. Disponible en: https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2. [Último acceso: 18 Septiembre 2020].

[8] W. Koehrsen, «Towards Data Science,» 27 Diciembre 2017. [En línea]. Disponible en: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0. [Último acceso: 02 Agosto 2020].

[9] A. Botchkarev, Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and, Toronto, Canada: Department of Computer Science, Ryerson University.

[10] Enciclopedia económica, «Enciclopedia Económica,» 2018. [En línea]. Disponible en: https://enciclopediaeconomica.com/variable-estadistica/. [Último acceso: 05 08 2020].

[11] Y. Yang, A study of pattern recognition of Iris flower based on machine learning, Turku University of Applied Science, 2013.

[12] J. Bagnato, «Aprende Machine Learning - Árboles de decisión,» 13 Abril 2018. [En línea]. Disponible en: https://www.aprendemachinelearning.com/arbol-de-decision-en-python-clasificacion-y-prediccion/. [Último acceso: 05 08 2020].

[13] M. G. José Alonso, Árboles de decisión, Sevilla: Dpto. de Ciencias de la Computaci´on e Inteligencia Artificial. Universidad de Sevilla.

[14] J. Martinez, «IArtificial.net,» 06 Abril 2019. [En línea]. Disponible en: https://www.iartificial.net/arboles-de-decision-con-ejemplos-en
python/#:~:text=Los%20%C3%A1rboles%20de%20decisi%C3%B3n%20son,(las%20ramas)%20representan%20soluciones.. [Último acceso: 14 Septiembre 2020].

[15] J. Bagnato, «Aprende Machine Learning - Random Forest, el poder del ensamble,» 17 Junio 2019. [En línea]. Disponible en: https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/. [Último acceso: 05 08 2020].

[16] P. Patil, «Towards Data Science,» 23 03 2018. [En línea]. Disponible en: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. [Último acceso: 22 08 2020].

[17] A. M. Castro, «Aspectos de producción,» IMP, México, 2013.

[18] C. Jiménez, «Perforación direccional,» Universidad de Oriente, Maturín, 2009.
[19] D. S. R. C. Helmuth Portilla., «La optimización de parámetros de perforación a través de propiedades geomecánicas.,» El reventón energético, vol. 10, nº 2, pp. 5,6, 2012.
[20] R. Hamrick, «Optimization of Operating Parameters for Minimum Mechanical Specific Energy in Drilling,» Department of Energy, United States of America, West Virginia, 2011.
[21] M. Z. M.-N. Afshin Davarpanah, «Assessment of Mechanical Specific Energy Aimed at Improving Drilling,» Journal of Petroleum & Environmental Biotechnology , vol. 7, nº 6, pp. 1, 5, 2016.
[22] E. M. y. A. R. Rafael Caballero, Big Data con Python: Recolección, almacenamiento y proceso, Bogotá: Alfaomega, 2019.
[23] Anacond Inc., «Anaconda Documentation,» [En línea]. Disponible en: https://docs.anaconda.com/anaconda/navigator/. [Último acceso: 03 10 2020].
[24] V. M. Sebastian Raschka, Python Machine Learning: Aprendizaje automático y aprendizaje profundo con Python, ScikitLearn y TensorFlow., Marcombo, 2019.
[25] R. O. Calafati, «Estrategias para el tratamiento de datos faltantes (missing data) en estudios con datos longitudinales.,» Universidad de Cataluña, España, 2017.
[26] M. Y. S. D. Portilla Helmut, «Optimización de parámetros de perforación con MSE e impacto en la construcción de un poz en el campo Yariguí-Cantagallo,» El reventón energético, vol. 12, nº 2, pp. 45-54, 2014.
[27] D. Ayala, A. Benítez y R. Valencia, «OPTIMIZACIÓN DE LA TASA DE,» Optimización de la tasa de penetración mediante el análisis de las vibraciones al perforar, caso de estudio Ecuador, vol. 15, nº 1, pp. 27-40, 2017.
[28] D. Cardozo, Field Drilling Data Cleaning and Preparation for Data Analytics Applications, Lousiana State University: Master's Theses, 2019.
[29] Pandas Pydata, «Pandas,» pandas.get_dummies, [En línea]. Disponible en: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html. [Último acceso: 19 10 2020].
[30] B. Chen, «Towards Data Science,» What is hot encoding and how to use pandas get dummies function, 00 2020. [En línea]. Disponible en: https://towardsdatascience.com/what-is-one-hot-encoding-and-how-to-use-pandas-get-dummies-function-922eb9bd4970.
[31] R. Meinert, «Optimizing Hyperparameters in Random Forest Classification,» Towards Data Science - Medium, 05 06 2019. [En línea]. Disponible en: https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6. [Último acceso: 12 08 2020].
[32] StatQuest, «Random Forest Explanation,» 2018.
[33] W. Koehrsen, «Medium - Towards Data Science,» Hyperparameter Tuning the Random Forest in Python, 09 01 2018. [En línea]. Disponible en: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74. [Último acceso: 22 08 2020].
[34] Kaggle, «Kaggle,» High train score, 23 08 2015. [En línea]. Disponible en: https://www.kaggle.com/getting-started/14998. [Último acceso: 17 11 2020].
[35] X. Ying, «An Overview of Overfitting and its Solutions,» Research Gate, pp. 1-7, 2019.

[36] D. Ramos, Evaluación Técnico - Financiera de la Tecnología de Conformance Químico en un Campo Petrolero, Bogotá: Universidad de América, 2018.

[37] J. Carlos, «Streemit,» Conversatorio #stem-espanol: Importancia en la interpretación de los parámetros de perforación, 21 08 2019. [En línea]. Disponible en: https://steemit.com/steemstem/@carlos84/conversatorio-stem-e-1561407379#:~:text=Los%20par%C3%A1metros%20de%20perforaci%C3%B3n%20son,la%20perforaci%C3%B3n%20y%20compararlos%20con. [Último acceso: 15 10 2020].

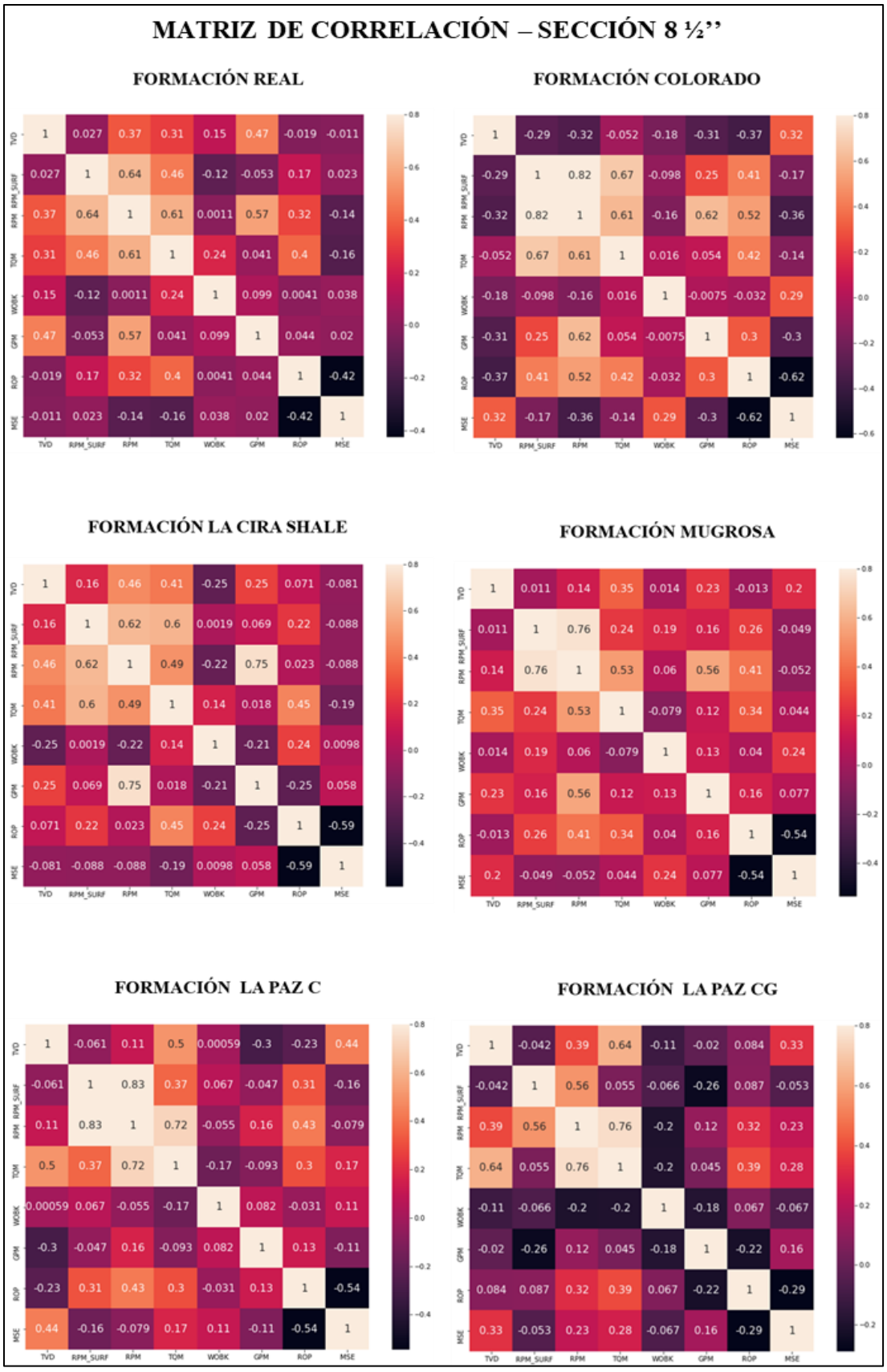
[38] W. C. M. G. Abraham Montes, «Aspectos de la perforación de pozos complejos en piedemonte en tiempos de crisis,» El reventón energético, vol. 16, nº 1, p. 87/97, 2018.

[39] Drilling Formulas, «Drilling Formulas,» Tangential Method Calculation, 22 08 2010. [En línea]. Disponible en: http://www.drillingformulas.com/tangential-method-calculation/. [Último acceso: 17 07 2020]

ANEXOS

ANEXO 1.

MATRIZ DE CORRELACIÓN DE VARIABLES PARA LA SECCIÓN 8 ½’’

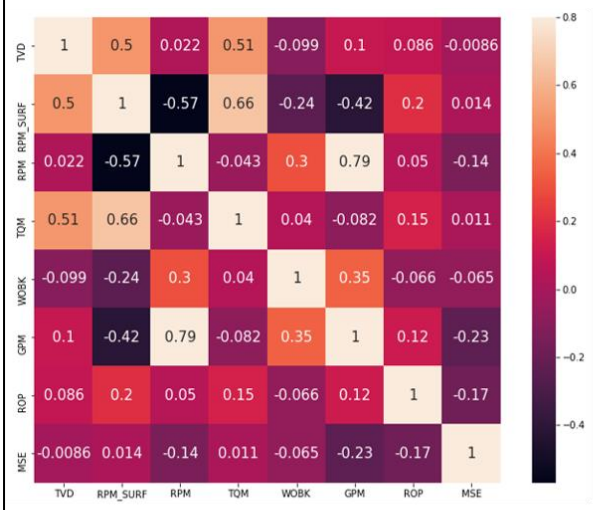


ANEXO 2.

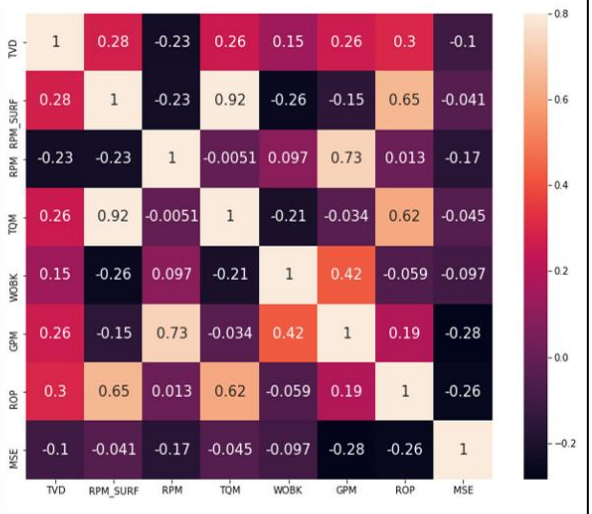
MATRIZ DE CORRELACIÓN DE VARIABLES PARA LA SECCIÓN 12 ¼’’

MATRICES DE CORRELACIÓN – SECCIÓN 12 ¼”

FORMACIÓN REAL



FORMACIÓN COLORADO

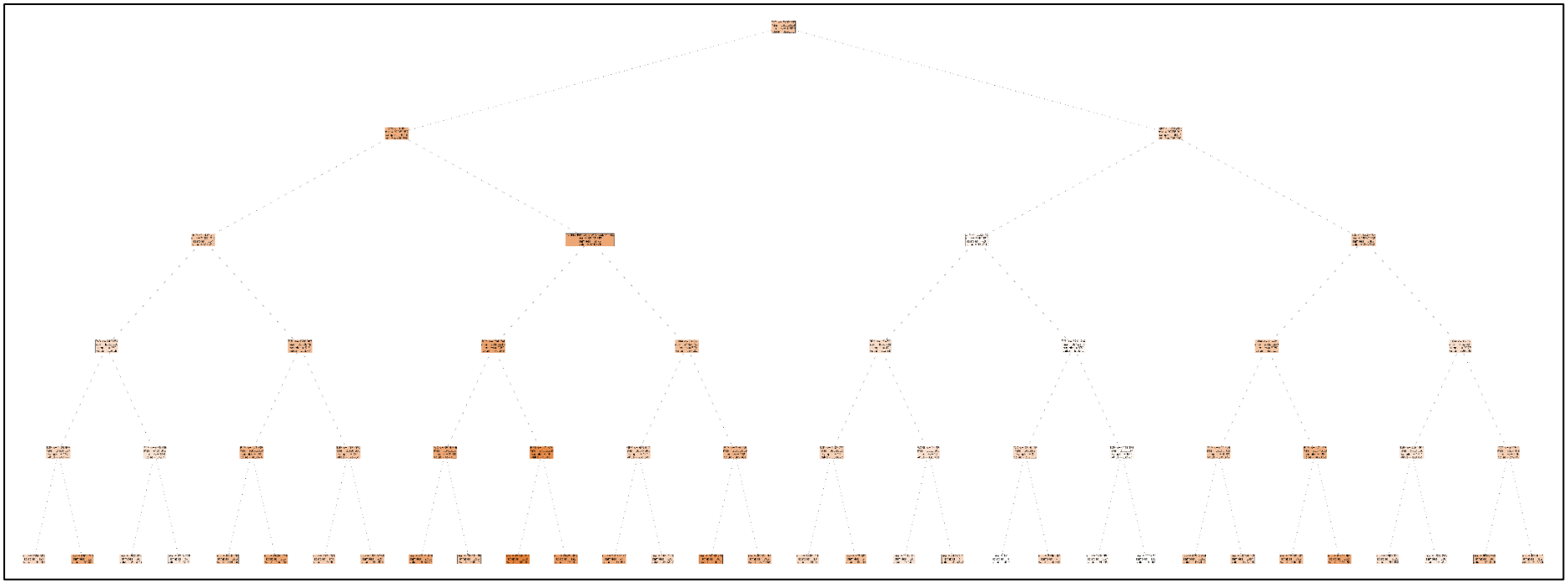


FORMACIÓN LA CIRA SHALE

FORMACIÓN MUGROSA

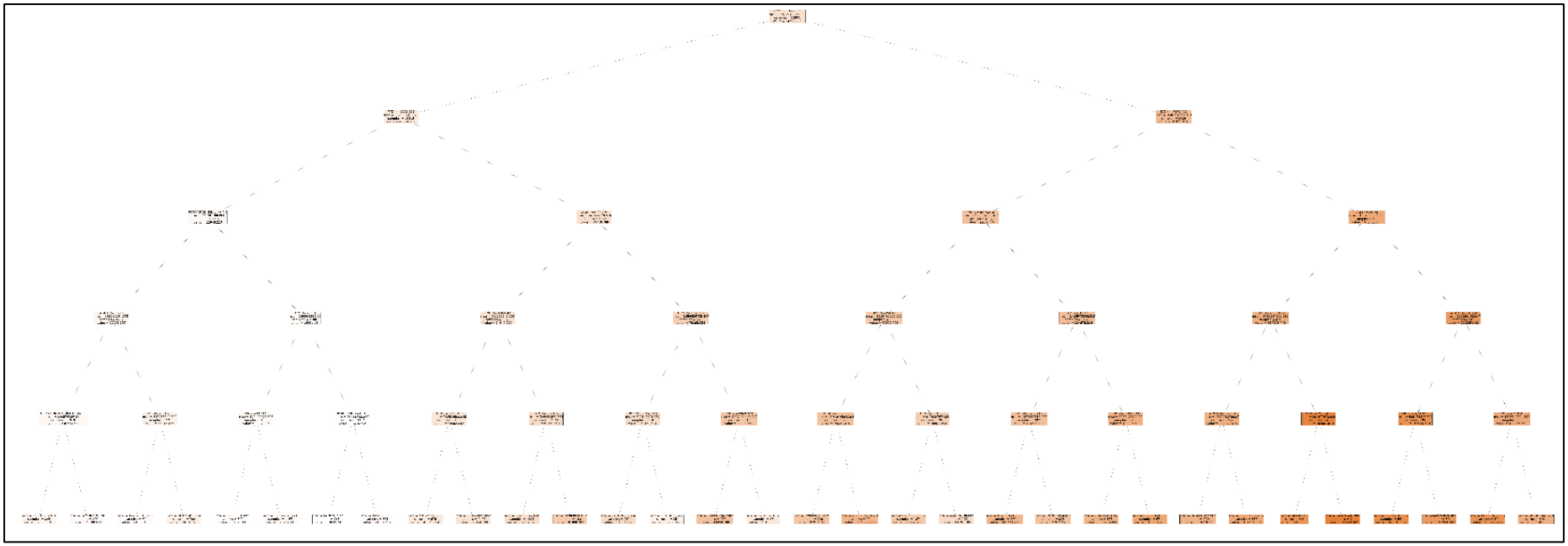
ANEXO 3.

ÁRBOL DE DECISIÓN PARA LA ESTIMACIÓN DE LA ROP EN LA SECCIÓN 8 ½’’



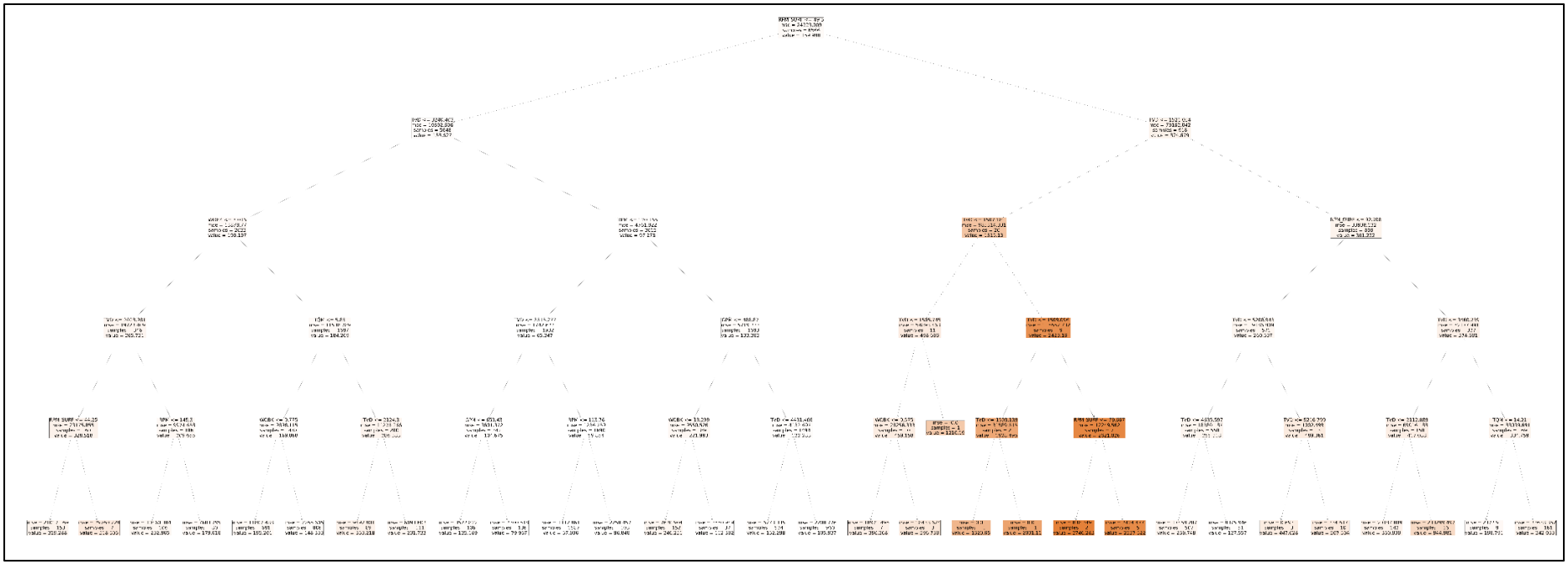
ANEXO 4.

ÁRBOL DE DECISIÓN PARA LA ESTIMACIÓN DE LA MSE EN LA SECCIÓN 8 ½’’



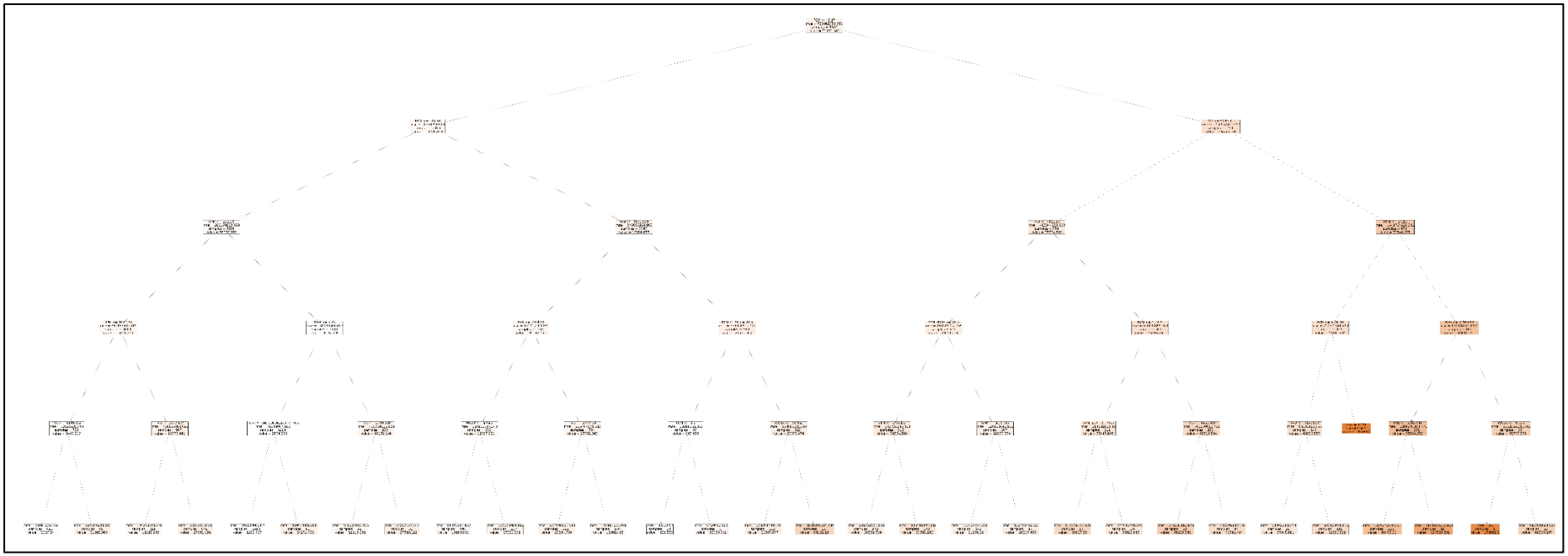
ANEXO 5. Á

RBOL DE DECISIÓN PARA LA ESTIMACIÓN DE LA ROP EN LA SECCIÓN 12 ¼’’



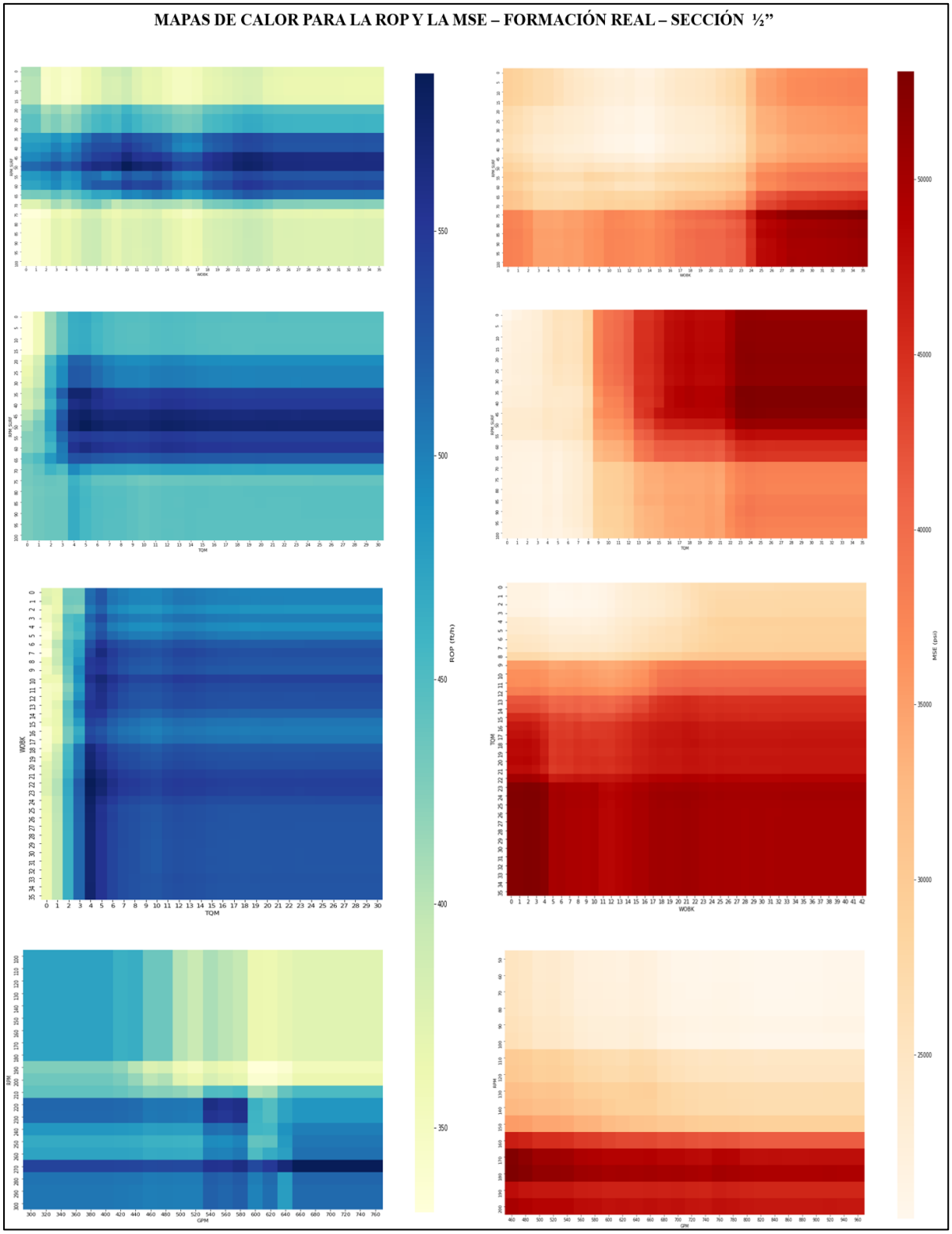
ANEXO 6.

ÁRBOL DE DECISIÓN PARA LA ESTIMACIÓN DE LA MSE EN LA SECCIÓN 12 ¼’’



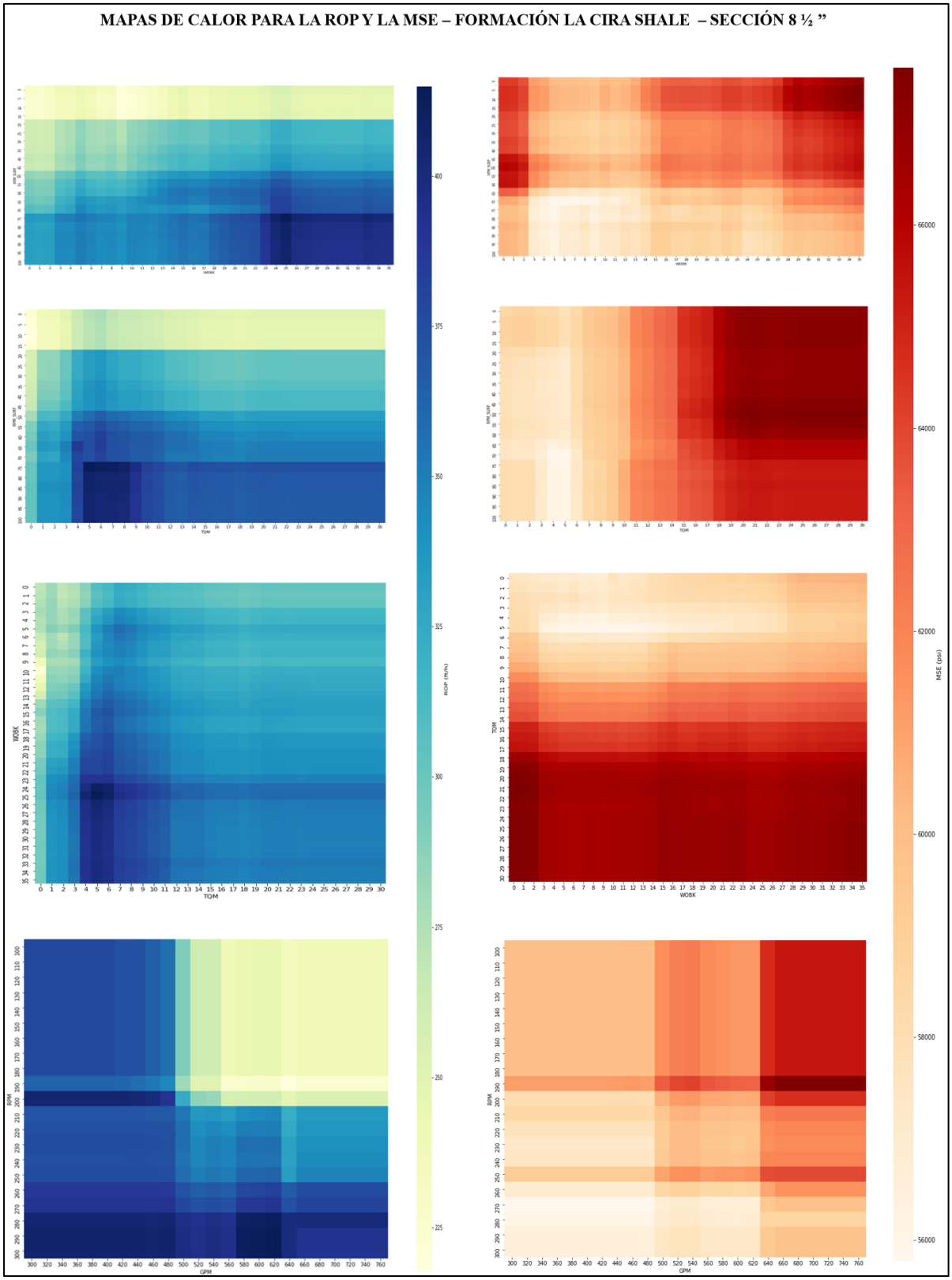
ANEXO 7.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN REAL



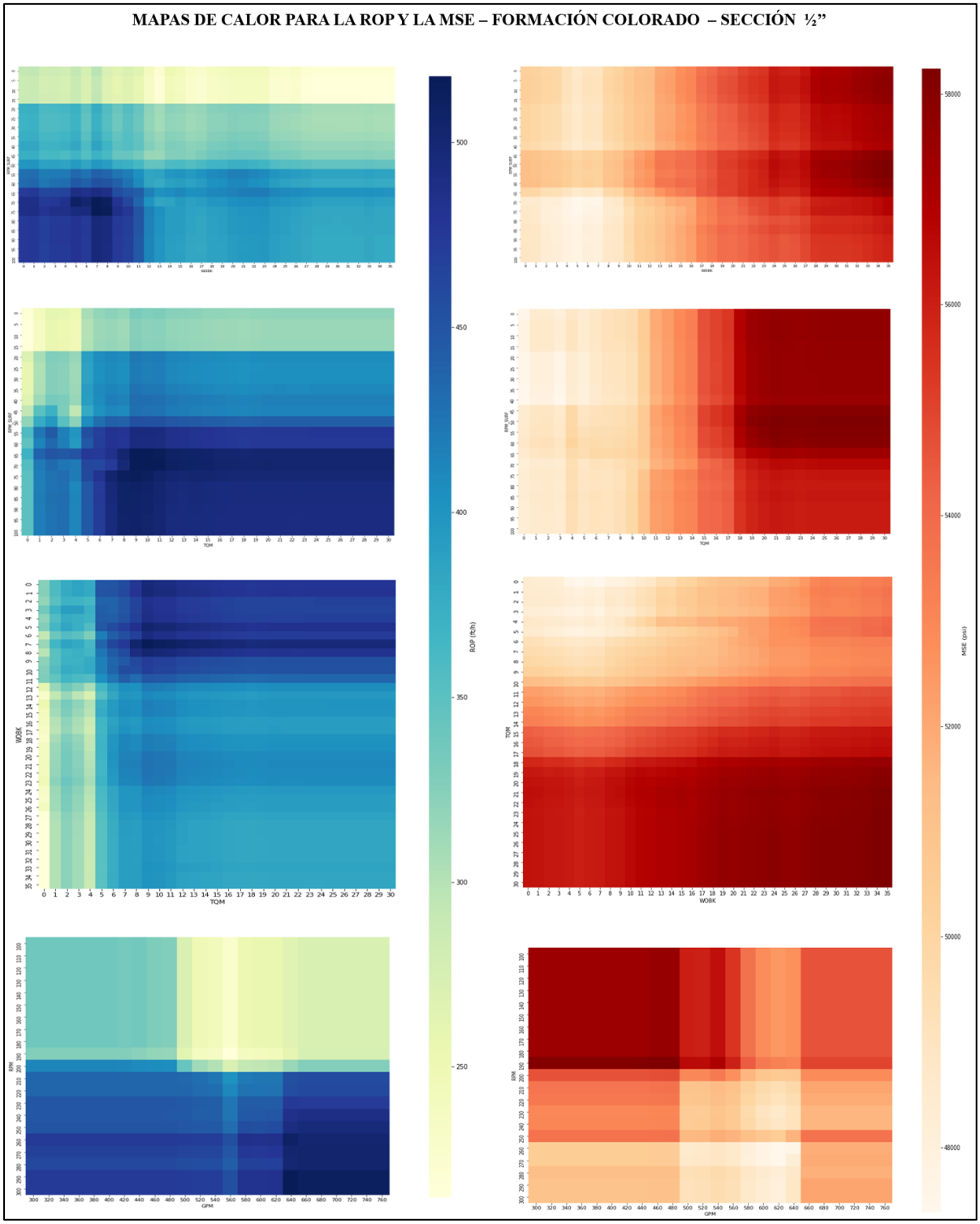
ANEXO 8.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN LA CIRA SHALE



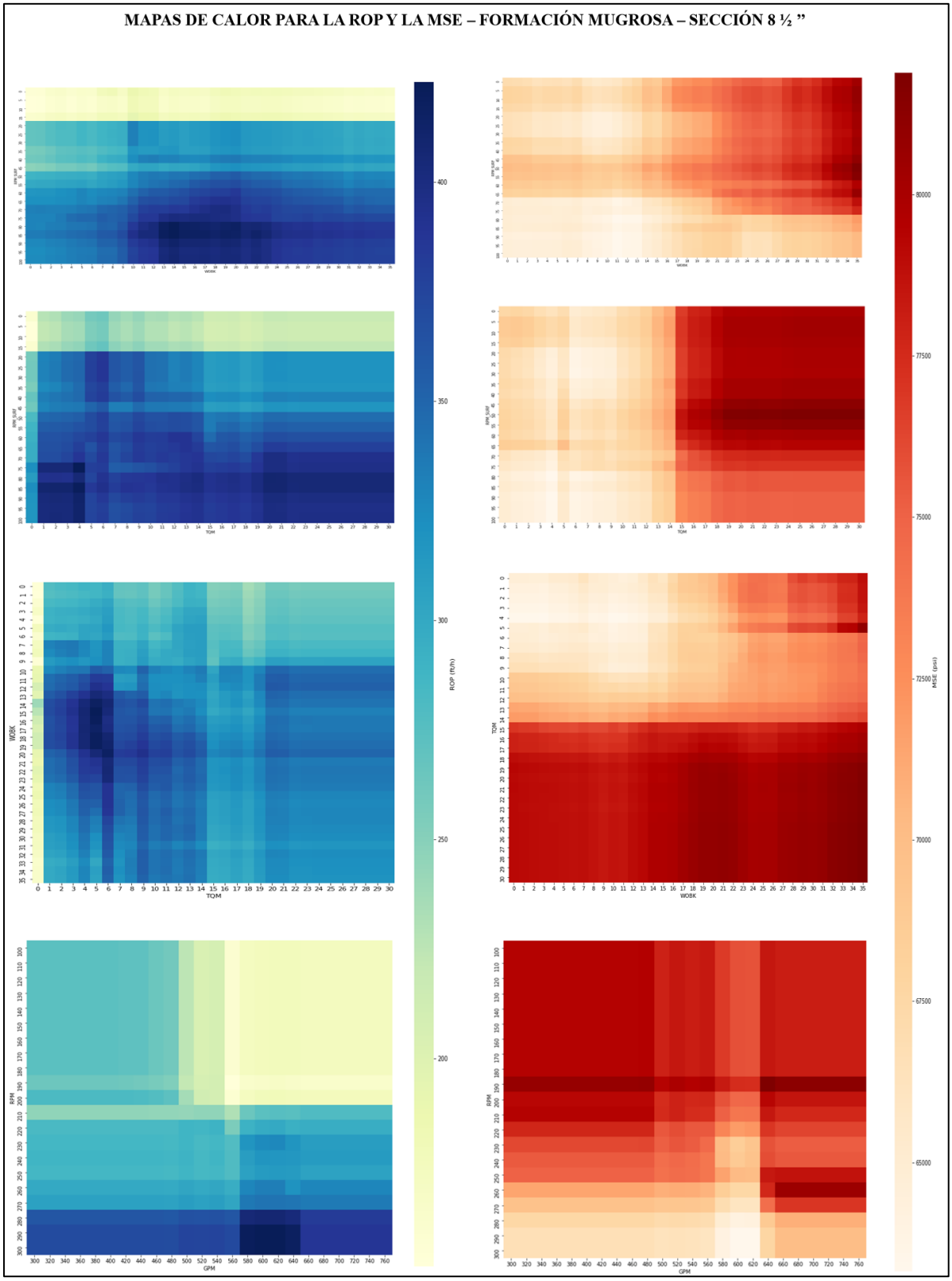
ANEXO 9.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN COLORADO



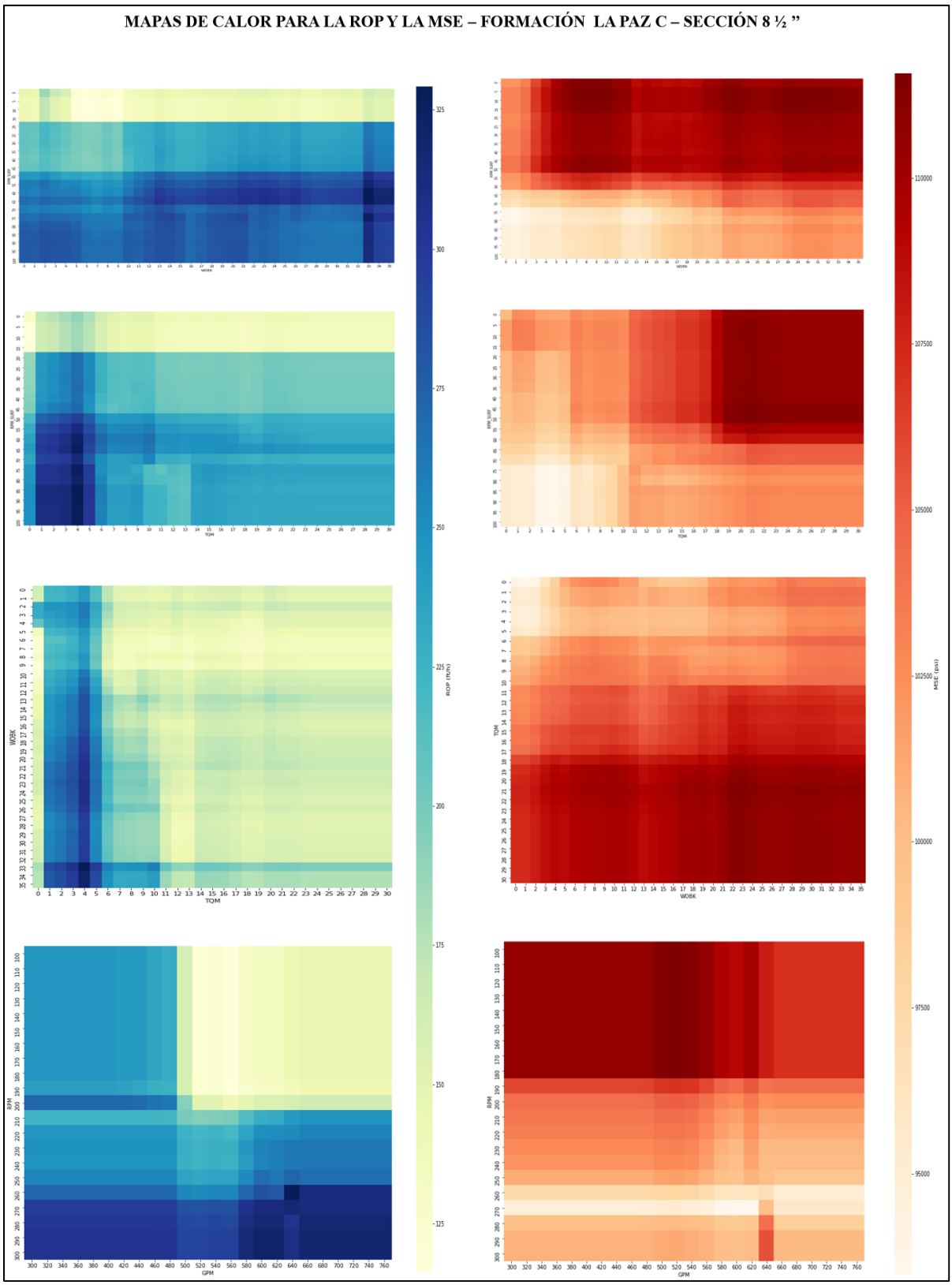
ANEXO 10.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN MUGROSA



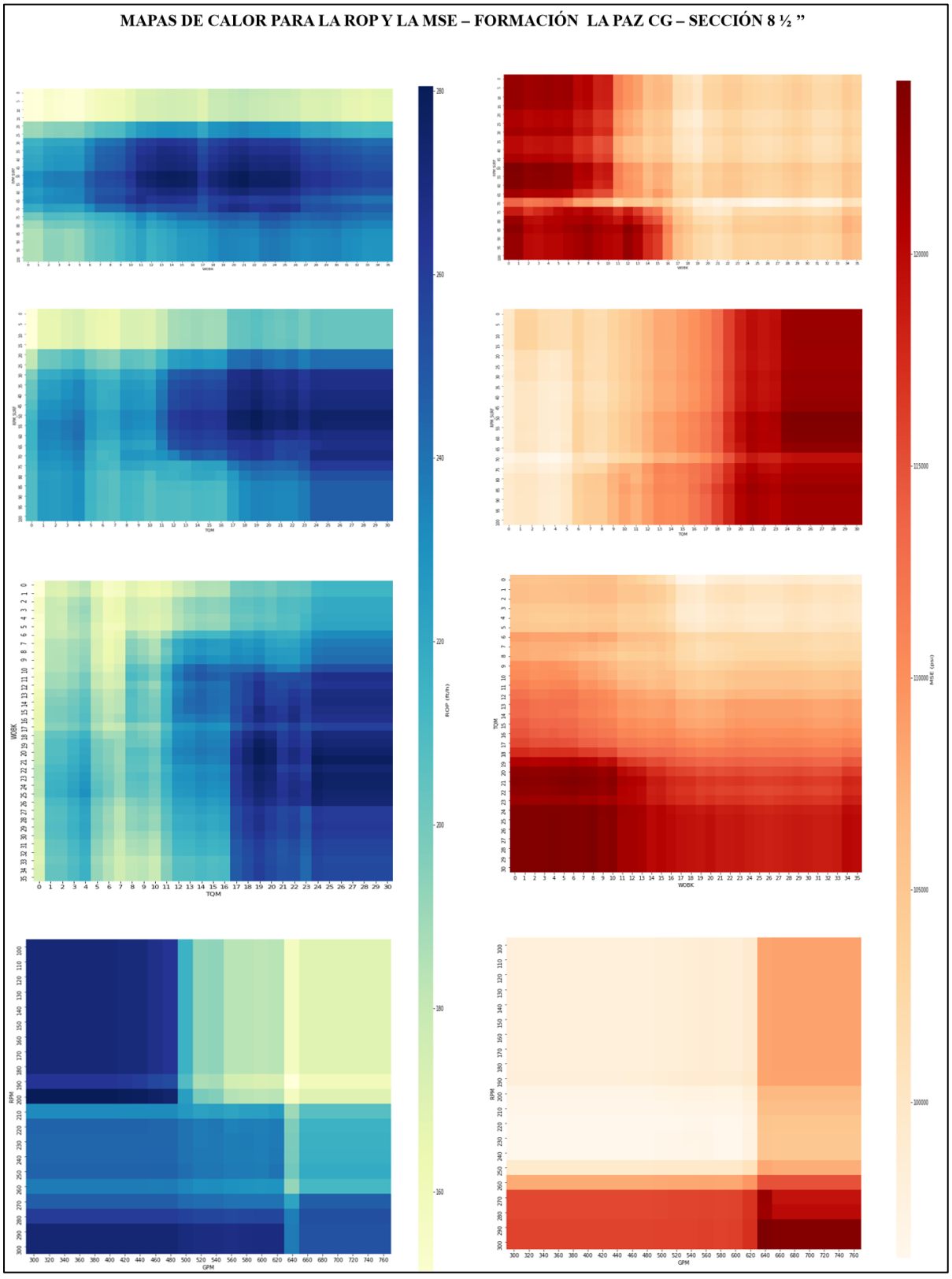
ANEXO 11.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN LA PAZ C



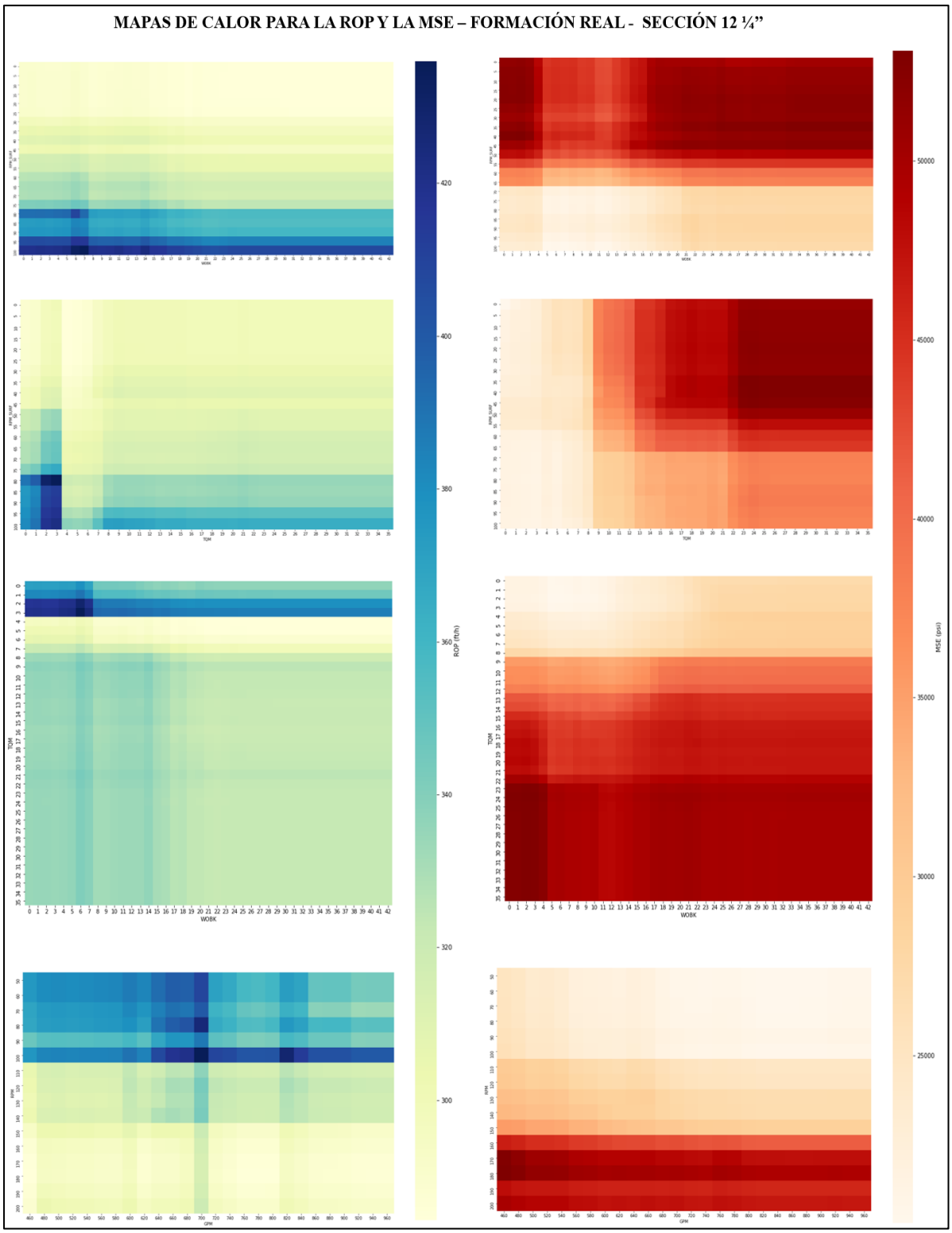
ANEXO 12.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 8 ½'' – FORMACIÓN LA PAZ CG



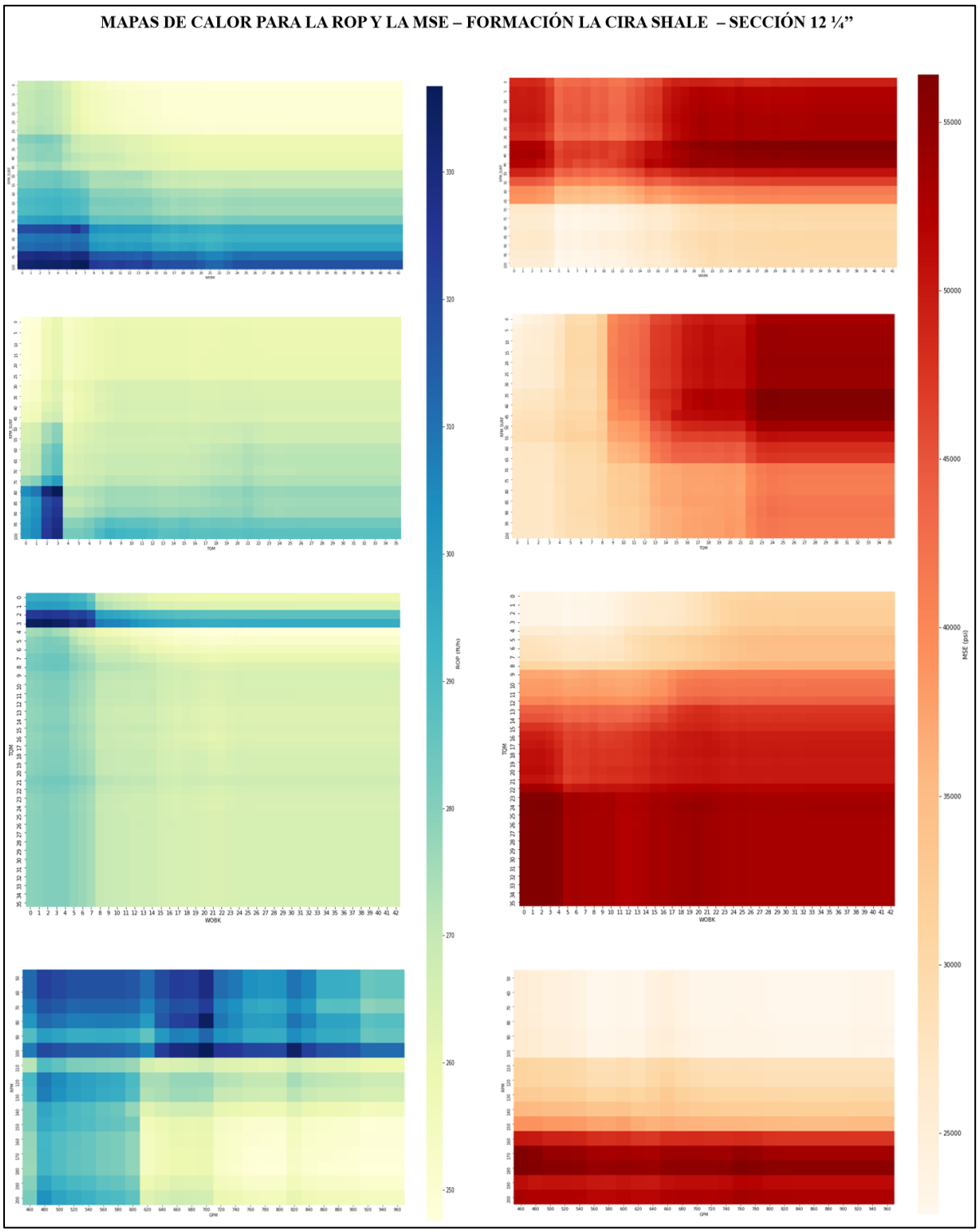
ANEXO 13.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 12 ¼'' – FORMACIÓN REAL



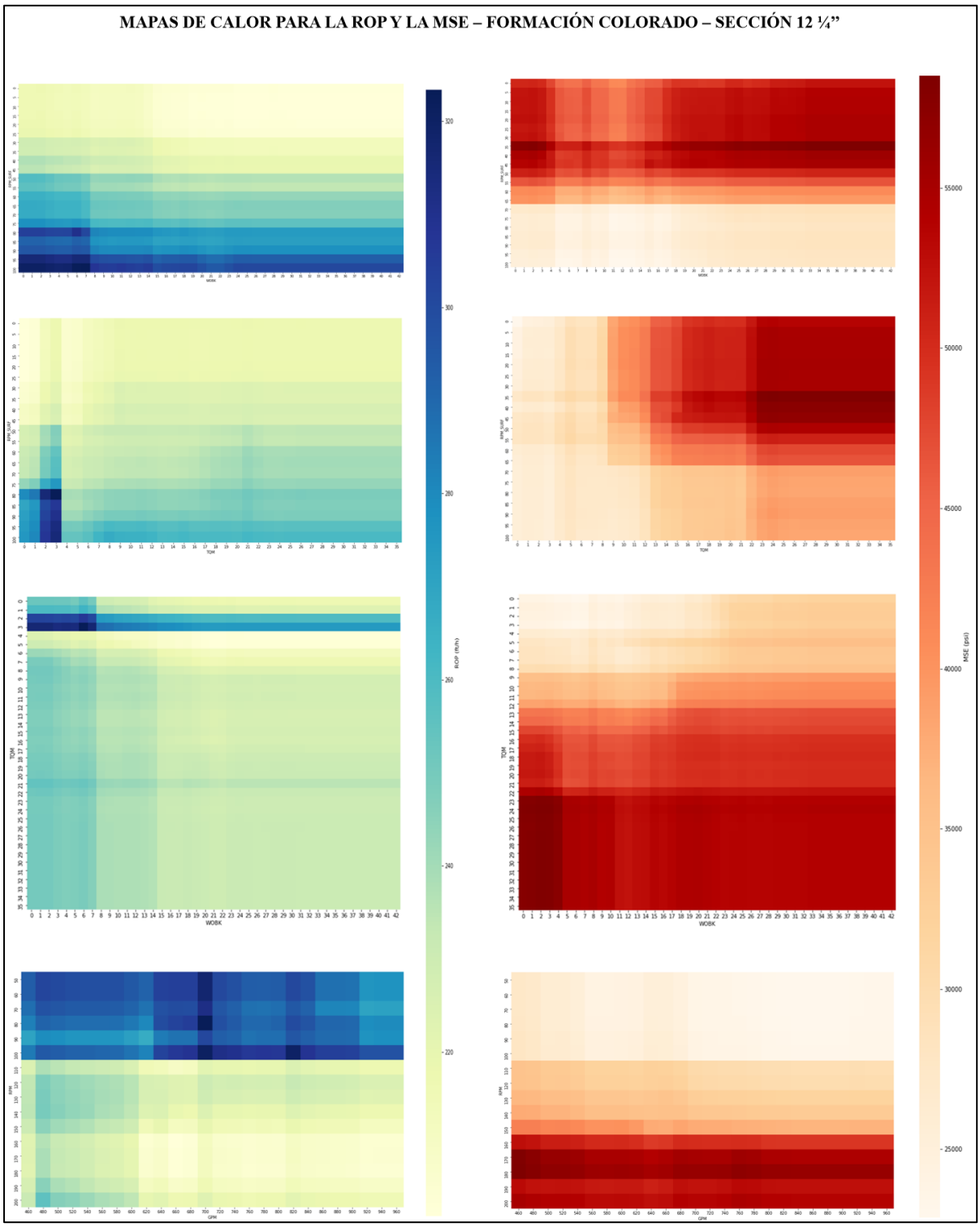
ANEXO 14.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 12 ¼'' – FORMACIÓN LA CIRA SHALE



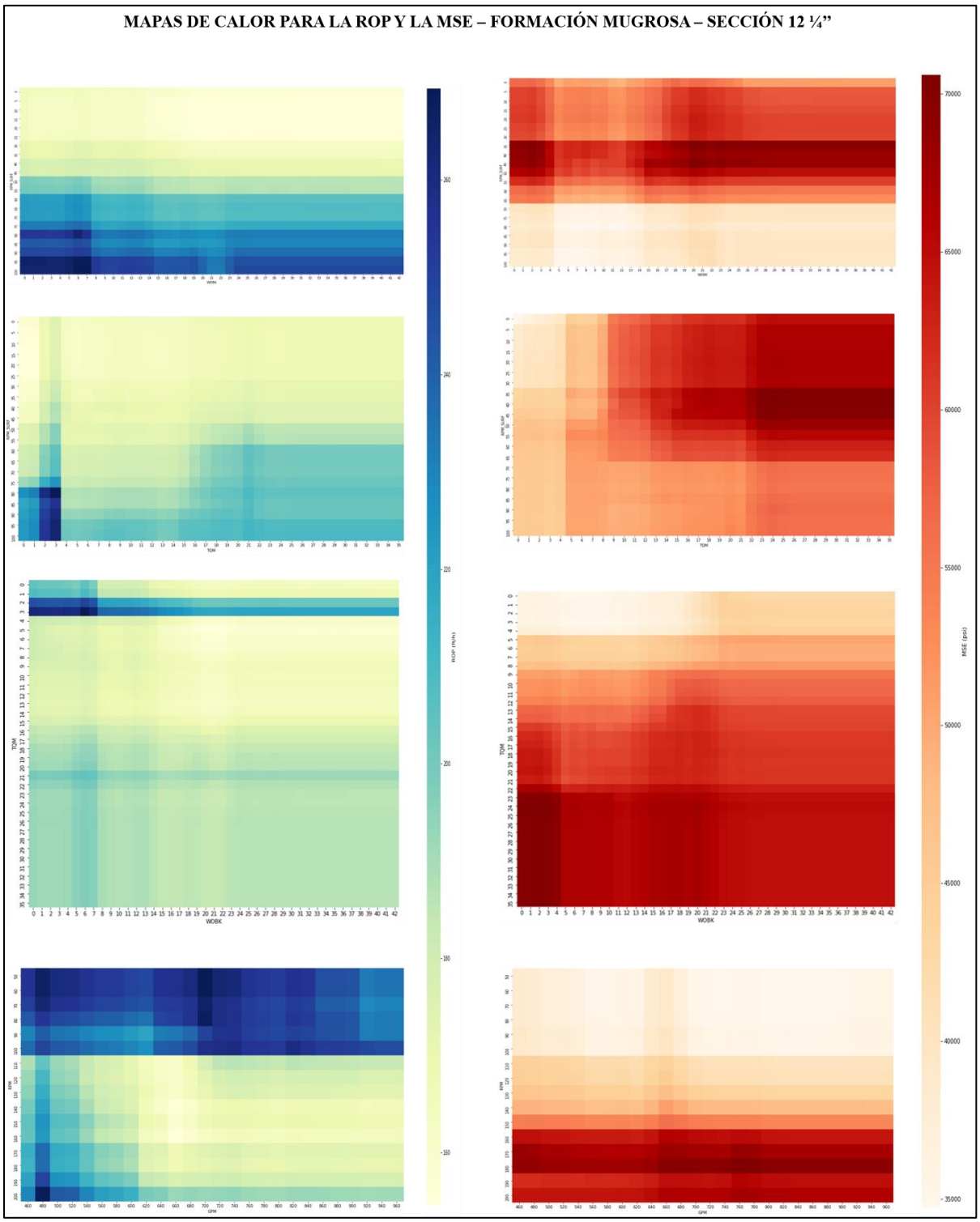
ANEXO 15.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 12 ¼'' – FORMACIÓN COLORADO



ANEXO 16.

MAPAS DE CALOR GENERALES PARA LA ROP Y LA MSE – SECCIÓN 12 ¼'' – FORMACIÓN MUGROSA



ANEXO 17.

GLOSARIO

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

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 qu
e 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 Pandas. 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.

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

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 conjuntos 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 [33, 5].

ANEXO 18.

RECOMENDACIONES

Se recomienda evaluar la integración de otras variables al modelo predictivo como puede ser la compresibilidad de la formación (UCS – CCS), con el fin de darle mayor generalidad al algoritmo, de esta forma podrían descartarse las variables categóricas de las formaciones geológicas.

Ejecutar un algoritmo predictivo sin la profundidad del pozo, con el fin de obtener un modelo generalizado para el Campo en estudio, buscando trabajar las formaciones y propiedades geológicas únicamente, para analizar la respuesta del algoritmo en esta situación.

Se recomienda agregar una variable categórica al modelo llamada geometría de pozo, que permita diferenciar entre los tipos de pozo perforados y las secciones en estudio, con el fin de evaluar el desempeño del algoritmo frente a la misma.

Aplicar otros modelos predictivos de machine learning para la elaboración de la presente investigación, con el fin de analizar diferentes escenarios, contrastando la efectividad entre los mismos para establecer mejores resultados.

Utilizar otras técnicas para la codificación de variables categóricas, como el algoritmo “ *label-encoder*, con la finalidad de codificar las variables categóricas, buscando así establecer la técnica se adapta mejor a la información suministrada.

Se recomienda ingresar información del mayor número de pozos posibles, con el propósito tener más datos al momento de entrenar el modelo, logrando obtener resultados más precisos al momento de ejecutar el algoritmo.

Implementar el modelo predictivo utilizado del presente proyecto en el campo en estudio con el fin de comparar y realizar un seguimiento en tiempo real de los
parámetros de perforación

Se recomienda extrapolar la metodología planteada en este proyecto, a diferentes campos en estudio, partiendo con información de fiable y de calidad.