OPTIMIZACIÓN EN LA PLANEACIÓN DE POZOS POR MEDIO DE LA PREDICCION DE TIEMPOS, COSTOS Y NPT´S, APLICANDO UN MODELO DE MACHINE LEARNING PARA LA CAMPAÑA DE PERFORACIÓN DE CASTILLA Y CASTILLA NORTE 2020

JUAN DAVID MARTINEZ ALONSO

JOHAN STEVEN POVEDA CRUZ

Proyecto integral de grado para optar al título de

Ingeniero de Petróleos

Orientador

Nelson Fernández Barrero

Ingeniero de Petróleos

FUNDACIÓN UNIVERSIDAD DE AMÉRICA

FACULTAD DE INGENIERÍAS

PROGRAMA DE INGENIERÍA DE PETRÓLEOS

BOGOTÁ D.C.

2021

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____ _____

Nombre Nombre

Firma del Director Firma del Presidente Jurado

_____

Nombre

Firma del Jurado

_____

Nombre

Firma del Jurado

Bogotá D.C. febrero de 2021.

DIRECTIVOS DE LA UNIVERSIDAD

Presidente Institucional y Rector del Claustro.

Dr. MARIO POSADA GARCIA PEÑA

Consejero institucional.

Dr. LUIS JAIME POSADA GARCIA PEÑA

Vicerrectoría Académica y de Investigaciones

Dra. MARIA CLAUDIA APONTE GONZALES

Vicerrector administrativo y financiero

Dr. RICARDO ALFONSO PEÑARANDA CASTRO

Secretaria General

Dra. ALEXANDRA MEJIA GUZMAN

Decano facultad de ingeniería.

ING. JULIO CESAR FUENTES ARISMENDI

Director de Programa Ingeniería de Petróleos.

ING. JUAN CARLOS RODRIGUEZ ESPARZA

DEDICATORIA

Quiero dedicar este trabajo de grado principalmente a mi familia

A mi madre Luz Marina por estar siempre cuando más lo necesite, por su amor incondicional y su apoyo constante en todo lo que me he propuesto.

A mi hermano Alejo que es una gran parte de mi vida y mi mejor amigo, gracias por siempre ayudarme cuando más lo necesite

A mi hermana Caro que sin su apoyo y su guía no estaría hoy donde estoy.

A mi padre Isauro por su confianza en mí y su apoyo incondicional.

A mi tío Daniel por ser siempre una persona con la que he podido contar y que siempre ha estado velando por nosotros.

A mi amigo y compañero de tesis Johan quien ha sido un gran soporte este último año y me alegra haberlo conocido, sin el nada de esto hubiera sido posible, fue un gran privilegio trabajar a su lado.

A mis amigos y a todas las personas que me acompañaron durante este viaje, aprendí algo diferente de cada uno de ellos.

¡Muchas Gracias!

Juan David Martinez Alonso

DEDICATORIA

Quiero dedicar este trabajo de grado a Dios principalmente, por haberme brindado salud, fuerza y sabiduría para culminar mi carrera profesional satisfactoriamente.

A mis padres José Joaquín Poveda y María Cristina Cruz, quienes han estado incondicionalmente apoyándome en cada decisión que he tomado, guiándome y brindándome lo mejor de ellos, siempre motivándome a dar lo mejor en cada aspecto de mi vida.

A mi hermano Rafael Poveda, por orientarme y guiarme, siendo un ser muy importante para mí, sin su ayuda y apoyo esto no hubiera sido posible.

A Jessica Serrano, por ser un gran apoyo cuando más lo necesite, brindándome su ayuda a lo largo de estos años.

A mi amigo y compañero de tesis Juan Martínez, a quien admiro y agradezco por estar en todo momento enseñándome y motivándome a ser siempre el mejor, por su compresión y dedicación a lo largo de este año para sacar este proyecto adelante. Fue un honor conocerlo y trabajar a su lado.

A mis amigos y a todas las personas que me acompañaron durante este recorrido, gracias por su ayuda y compañía.

¡Muchas gracias!

Johan Steven Poveda Cruz

AGRADECIMIENTOS

A la Fundación Universidad de América, por los conocimientos impartidos y a su cuerpo docente por orientarnos y guiarnos a lo largo de este proceso.

A la compañía Ecopetrol S.A, por brindarnos su apoyo y soporte para llevar a cabo este trabajo grado.

A nuestros directores de trabajo de grado, ing. Ricardo Andrés García Parra y Ing. Wilmar Osorio Quintero, por su paciencia, colaboración y tiempo dedicado para la ejecución del presente trabajo de grado.

A nuestros orientadores de tesis, Nelson Fernández Barrero, Jorge Andrés Tovar, Sebastián Alejandro Gomez Alba, por guiarnos y orientarnos en base a sus conocimientos durante la ejecución de este proyecto.

Al ingeniero Jhon Franklin Gonzales Gamboa por enseñarnos e incentivarnos a la mejora continua, por ser un gran punto de apoyo durante la ejecución de este trabajo de grado.

Las directivas de la Fundación 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

INTRODUCCIÓN 14

1. MARCO TEÓRICO 17

1.1. Generalidades del campo Castilla 17

1.2. Inteligencia artificial 19

1.3. Machine Learning 21

1.3.1. Tipos de machine Learning 21

1.3.2. Clasificación de machine Learning 21

1.3.3. Tipos de modelos de machine Learning 22

1.4. Herramientas digitales 25

1.4.1. OpenWells® 25

1.4.2. Power BI 25

1.4.3. Jupyter notebook 25

1.4.4. Python 26

1.5. Tiempos no productivos 27

2. METODOLOGÍA Y DATOS 28

2.1. Primera etapa: recopilación de información del 2019 y 2020 29

2.1.1. Selección de los pozos perforados del 2019 y 2020. 30

2.1.2. Extracción de información variables implicadas en la matriz de complejidad y respectiva implementación para la campaña del 2019 y 2020 32

2.1.3. Extracción de tiempos y costos 34

2.1.4. Extracción de los NPT'S 35

2.1.5. Creación de base de datos general y especificas 35

2.2. Segunda etapa: selección y evaluación del modelo predictivo en la campaña de perforación del 2019. 38

2.2.2. Búsqueda bibliográfica y selección de los modelos predictivos 45

2.2.3. Aplicación de los modelos predictivos en la campaña de perforación 2019, evaluación y selección del modelo predictivo con mayor desempeño 46

2.2.4. Selección del modelo predictivo con mayor desempeño 51

2.3. Tercera etapa: aplicación del modelo predictivo en la campaña de perforación del 2020. 52

2.3.1. Aplicación del modelo predictivo en la campaña de perforación 2020 52
2.4. Cuarta etapa: creación de la interfaz gráfica y evaluación del desempeño del modelo predictivo 53
2.4.1. Creación base de datos final con lo planeado, ejecutado y pronosticado para los pozos perforados en el 2020 54
2.4.2. Creación del tablero dinámico en Power BI y evaluación del desempeño del modelo predictivo seleccionado 55
3. ANÁLISIS Y RESULTADOS 56
3.1. Análisis de resultados de la aplicación de los modelos predictivos en la campaña de perforación 2019 56
3.1.1. Evaluación y selección de los mejores hiperparametros 58
3.1.2. Validación cruzada 65
3.2. Selección de los modelos predictivos 66
3.2.1. Selección del mejor modelo para días 67
3.2.2. Selección del modelo para la predicción de costos 69
3.2.3. Selección del mejor modelo para NPT’s 71
3.2.4. Análisis variables predictoras de los modelos seleccionados 73
3.3. Resultados aplicación de los modelos predictivos en la campaña de perforación del 2020. 77
3.4. Análisis y resultados creación de la interfaz gráfica y evaluación del desempeño del modelo predictivo 78
3.4.1. Análisis y resultados modelo DTR sobre la campaña 2020 para días. 79
3.4.2. Analisis y resultados modelo RFR sobre la camapaña 2020 para costos. 82
3.4.3. Analisis y resultados modelo DTR sobre la camapaña 2020 para NPT’s. 83
CONCLUSIONES 87
RECOMENDACIONES 90
BIBLIOGRAFÍA 91
TÉRMINOS TÉCNICOS 96
ANEXOS 98

LISTA DE FIGURAS

pág

Figura 1. Estado de mecánico general del campo Castilla y Castilla Norte 19
Figura 2. Clasificación de la inteligencia artificial. 20
Figura 3. Procedimiento interno árbol de decisión. 23
Figura 4. Procedimiento interno del modelo bosques aleatorios 24
Figura 5. Metodología para la implementación de los modelos predictivos 29
Figura 6. Procedimiento primera etapa. 30
Figura 7. Implementación y obtención del resultado de la matriz de complejidad. 32
Figura 8. Resultado final de la matriz de complejidad. 34
Figura 9. Procedimiento segunda etapa. 38
Figura 10. Información base de datos 2019. 40
Figura 11. Estadística descriptiva bases de datos 2019. 41
Figura 12. Funciones y códigos para estandarización de la base de datos 2019. 42
Figura 13. Diagrama de caja y bigotes. 42
Figura 14. Diagrama de mapa de correlación. 43
Figura 15. Diagrama de parejas. 44
Figura 16. Procedimiento para la búsqueda bibliográfica. 45
Figura 17. Procedimiento sección 2.2.3. 47
Figura 18. Proceso interno función cross_val_score. 51
Figura 19. Procedimiento tercera etapa. 52
Figura 20. Ingreso de nuevo valores para predicciones de días, costos y NPT’s. 53
Figura 21. Procedimiento cuarta etapa. 54
Figura 22. Diagrama de caja y bigotes. 57
Figura 23. MAE vs R2 para el modelo DecisionTreeRegressor. 60
Figura 24. MAE vs R2 para el modelo RandomForestRegressor. 61
Figura 25. MAE vs R2 para el modelo SupportVectorRegressor. 62
Figura 26. Valores de R2 más altos de cada modelo predictivo. 65
Figura 27. Representación del árbol de decisión. 68
Figura 28. Variables predictoras con sus respectivos pesos. 69
Figura 29. Muestra representativa de uno de los árboles de decisión del bosque aleatorio. 70
Figura 30. Variables predictoras con sus respectivos pesos. 71
Figura 31. Representación del árbol de decisión. 72
Figura 32. Variables más relevantes para el modelo de NPT´s. 73
Figura 33. Mapa de calor con las variables más relevantes del modelo DTR para días. 74
Figura 34. Mapa de calor correlacionando las variables más relevantes del modelo RFR para costos. 75
Figura 35. Mapa de calor correlacionando las variables más relevantes del modelo DTR para NPT’s. 76

Figura 36. Página de inicio del tablero dinámico. 79
Figura 37. Tablero dinámico para la variable días. 80
Figura 38. Tablero dinámico para la variable días. 81
Figura 39. Tablero dinámico para la variable costos. 82
Figura 40. Tablero dinámico para variable NPT’s. 84
Figura 41. Tablero dinámico para variable NPT’s. 85

LISTA DE TABLAS

pág

Tabla 1. Propiedades petrofísicas del yacimiento 18
Tabla 2. Campañas de perforación 2019 y 2020. 31
Tabla 3. Variables implicadas en la matriz de complejidad. 33
Tabla 4. Base de datos general. 36
Tabla 5. Base de datos 2019. 37
Tabla 6. Base de datos 2020. 37
Tabla 7. Equivalentes de las variables matriz complejidad en Python. 39
Tabla 8. Ventajas y desventajas de los modelos predictivos seleccionados. 46
Tabla 9. Parámetros y rangos determinados para cada modelo predictivo. 49
Tabla 10. Mejores parámetros modelo DecisionTreeRegressor para costos, días y NPT’s. 63
Tabla 11. Mejores parámetros modelo RandomForestRegressor para costos, días y NPT’s. 63
Tabla 12. Mejores parámetros modelo SupportVectorRegressor para costos, días y NPT's. 64
Tabla 13. Validación cruzada con los mejores parámetros de los modelos predictivos. 66
Tabla 14. Modelos predictivos seleccionados según variable objetivo 67
Tabla 15. Predicciones para tiempos, costos y NPT’s 2020. 77
Tabla 16. Base de datos 2020 final. 78

LISTA DE ABREVIATURAS

DDI: Directional Difficult index

DTR: Decision Tree Regressor

EDM: Engineers Data Model

MD: Measured Depth

NPT's: Non Productive Times

ODR: Operational Drilling Report

PPG: Pound Per Galon

RFR: Random Forest Regressor

SVR: Support Vector Regressor

TVD: True Vertical Depth

VS: Vertical Section

RESUMEN

Actualmente la compañía Ecopetrol S.A, tiene en cuenta los datos técnicos-históricos tales como los que se encuentran almacenados en OpenWells y Power BI, para evaluar el desempeño durante la fase de perforación de los pozos semana a semana, en vez de aprovechar dicha información junto con las variables implicadas en la matriz de complejidad, para optimizar la planeación de pozos mediante la implementación de un modelo predictivo generando así un valor agregado sobre la información almacenada.

Considerando lo anterior, el presente trabajo de grado se realizó con el fin de optimizar la planeación de pozos para la campaña de perforación de Castilla y Castilla Norte 2020 al aplicar los modelos de machine Learning seleccionados, los cuales predicen días costos y NPT's asociados a problemas en hueco abierto.

Por lo cual, se diseñó una metodología orientada a la implementación de tres modelos de machine Learning de tipo supervisado, basándose en la información de la campaña de perforación del Campo Castilla y Castilla Norte 2019. Posteriormente, se implementó y evaluó la predicción de los modelos para el mismo campo en el 2020.

Los resultados obtenidos durante la implementación de los modelos predictivos sobre el campo Castilla y Castilla Norte 2020, fueron los siguientes:

Para la predicción de días y NPT`s se seleccionó el modelo DecisionTreeRegressor, obteniendo una precisión de lo ejecutado con respecto a lo pronosticado del 77.1% y 30.72% respectivamente. Por otra parte, para la predicción de costos, seleccionó el modelo RandomForestRegressor obteniendo una precisión de lo ejecutado con respecto a lo pronosticado del 79.51%.

Palabras clave: Machine Learning, Método supervisado, matriz complejidad, OpenWells, Power Bi, Python.

INTRODUCCIÓN

Con base a los factores geopolíticos [1] y fluctuación en el precio del dólar, la industria petrolera está sujeta a crisis inesperadas donde la inversión económica para presentes y futuros proyectos se ven en riesgo. Lo anterior, puede afectar la viabilidad financiera de las operaciones como la exploración, perforación, completamiento y producción.

Ecopetrol S.A, actualmente opera los campos Castilla y Castilla Norte, ubicados en la Cuenca de los Llanos de Orientales, en el departamento del Meta. Durante la fase de planeación para dichos campos se evidencia una oportunidad de mejora a raíz de la información técnica e histórica como la que se encuentra consignada en OpenWells y Power BI. Sin embargo, estos datos solo se implementan para evaluar el desempeño periódico de los pozos, en vez de aprovechar dicha información junto a las variables que se encuentran en la matriz de complejidad, con la finalidad de optimizar la planeación de pozos o como se propone en este trabajo, predecir mediante una técnica innovadora días y costos relacionados con la operación de perforación y a su vez predecir tiempos no productivos (NPT'S) asociados a problemas de hueco abierto.

Este trabajo de grado fue uno de los primeros en implementar modelos predictivos para asistir al proceso de planeación de pozos, por medio de la predictibilidad de tiempos y costos basados en la información interna de la compañía. Este modelo de tipo supervisado pretenderá convertirse en un elemento indispensable a través del uso de Machine Learning, el cual se centrará en la aplicación de algoritmos que aprenderán de los datos y serán capaces de encontrar patrones, logrando así la toma de decisiones sobre nuevos conjuntos de datos, mejorando su precisión a medida que se adicione más información al mismo [2]. El uso de este tipo de modelos podrá implementarse en otras áreas como el completamiento y producción, generando así un valor agregado a la información de la compañía.

El uso de Machine Learning ha venido adquiriendo importancia a lo largo de los últimos años en la industria petrolera, esto se puede evidenciar mediante la investigación de Jessamyn Sneed, donde identifico estadísticamente los factores clave detrás de las fallas del BES y determino si era posible, predecir con precisión la vida útil de una BES

utilizando técnicas de modelos predictivos. Para la realización de los modelos predictivos el autor utilizó una variedad de algoritmos como: regresión lineal, arboles de decisión y bosques aleatorios de alto rendimiento, donde la mejor opción fue el modelo HP Random Forest, basándose en el error promedio cuadrado (MSE). Este modelo predijo que el 90% de error predictivo era alrededor de 30 días de la vida útil de la herramienta de levantamiento artificial. [3]

Por otro lado, los autores Xinxin Hou, Jin Yang, and Qishuai Yin, emplearon algoritmos de machine learning y tecnología big data para extraer y analizar datos de perforación de pozos en el mar de china meridional donde la perdida de circulación es importante, allí se consideraron características geológicas, parámetros de propiedad de los fluidos de perforación y parámetros operativos de perforación, adicionalmente los autores emplearon una red neuronal artificial para realizar el aprendizaje supervisado; para evaluar dicho modelo utilizaron métricas tales como exactitud, precisión, puntuación f1 y exhaustividad, donde se logró una exactitud del 92%, con un puntaje f1, precisión y exhaustividad hasta del 85%. Lo anterior demostró que el modelo tiene una buena capacidad de generalización y puede ser aplicado a otros campos, adicionalmente, puede predecir seis riesgos de perdida de circulación, cada uno de acuerdo con la tasa de perdida de perforación. [4]

El objetivo general de este trabajo de grado es la optimización en la planeación de pozos por medio de la predicción de tiempos, costos y NPT'S, aplicando un modelo de Machine Learning para la campaña de perforación de Castilla y Castilla norte 2020, donde los objetivos específicos son:

- Recopilar la información de entrada del algoritmo predictivo, a partir de la implementación de la matriz de complejidad, Open Wells y Power Bi para la campaña de perforación del 2019.
- Evaluar algoritmos predictivos que se ajusten a la data técnica recopilada para posterior aplicación de Machine Learning.
- Implementar el algoritmo predictivo para la campaña de perforación del 2019 de Castilla y Castilla Norte para entrenamiento del modelo y su primera validación.

- Aplicar el algoritmo seleccionado para su ejecución durante el año 2020 en la campaña de perforación en los pozos de Castilla y Castilla Norte.
- Diseñar una interfaz gráfica en el visualizador Power Bi para la evaluación del desempeño de la campaña de perforación de Castilla y Castilla norte 2020 aplicando el modelo seleccionado de Machine Learning.

El presente proyecto de grado plantea la aplicación de un modelo predictivo utilizando la técnica de Machine Learning, basándose en un modelo supervisado, con el propósito de lograr una optimización durante la fase de planeación para días y costos de perforación en la campaña de Castilla y Castilla norte 2020, asimismo se busca predecir los tiempos no productivos (NPT'S) asociados a problemas en hueco abierto en los anteriores campos mencionados.

Con el fin de dar cumplimiento a los objetivos específicos propuestos en el presente proyecto, en el capítulo de Metodología, en la sección 2.1, se describe los procedimientos para la recopilación de información del 2019 y 2020 en los Campos Castilla y Castilla Norte. Posteriormente, para cumplimiento al segundo y tercer objetivo específico, en la sección 2.2, se seleccionaron y evaluaron los modelos predictivos sobre la campaña de perforación del 2019 y se eligió el modelo con mayor desempeño. Más adelante, respecto al cuarto objetivo específico, en la sección 2.3, se aplicaron los modelos predictivos seleccionados sobre la campaña de perforación del 2020 para la obtención de las predicciones. Para dar cumplimiento al quinto objetivo específico, en la sección 2.4, se lleva a cabo la creación de una interfaz gráfica en el visualizador Power Bi para evaluar el desempeño del modelo predictivo seleccionado realizando una comparación entre lo planeado y ejecutado por la compañía versus lo pronosticado por los modelos predictivos.

1. MARCO TEÓRICO

Para llevar a cabo el desarrollo del presente proyecto de grado, se realizará una breve descripción de las generalidades del campo Castilla. Posteriormente, se introducirá al método de Machine Learning y herramientas claves que conducirán a la optimización deseada. Por último, se abordará los tiempos no productivos asociados a problemas en hueco abierto que se presentan dicho campo.

1.1. Generalidades del campo Castilla

La operadora Chevron inicio el proceso de exploración en 1969 con el pozo Castilla 1, alcanzando una profundidad de 7347 pies, probando crudo pesado en las formaciones Mirador, Guadalupe y Une. [5] Mas adelante en 1973, se firmó un contrato de asociación entre las compañías Ecopetrol y Chevron, donde se pactó qué ambos tendrían una participación del 50%, con un término de 25 años desde 1975, año en el cual se declaró la comercialidad del campo e iniciaron las actividades de producción y explotación de este. En el año 2000, se declaró a Ecopetrol como operador total y directo del campo bajo un contrato de E&P. [6]

El campo Castilla se encuentra ubicado en la cuenca de los Llanos Orientales en el Departamento del Meta, limita con los municipios de Acacias y Castilla la Nueva, se encuentra a 30 km al sur de Villavicencio y a 167 km de la ciudad de Bogotá D.C. [7]

Geológicamente, la estructura del campo Castilla es un anticlinal asimétrico elongado con una orientación N60E de aproximadamente de 10 km de largo por 4 km de ancho, fallado en el flanco oriental. Contiene fallas internas normales e inversas y tiene un tipo de trampa estructural. [5]

Estratigráficamente, los pozos deben atravesar las siguientes formaciones: Fm. Guayabo, Fm. León, Fm. Carbonera, Fm. T2, Fm. K1, Fm. k2, siendo las dos últimas, las formaciones de interés.[7]

Tabla 1.

Propiedades petrofísicas del yacimiento.

Propiedades	Valores
Porosidad (%)	10
Permeabilidad (mD)	6874
Gravedad API (°API)	12.5 - 18
Espesor de interés (ft)	24 - 40
Presión Yacimiento (Psi)	2830
Temperatura Yacimiento (°F)	198
Reservas (BBL)	800.000 - 1.000.000
OOIP (BBL)	6.000.000 - 7.000.000
Índice de productividad	0.1 - 0.2
Mecanismo de producción	Empuje hidráulico

Nota. Generalidades del yacimiento para Campo Castilla 2017. Tomado de: Evaluación de los efectos del fracturamiento hidráulico sobre el comportamiento de producción en cuatro pozos del campo castilla norte mediante registros de producción, test de laboratorio y pruebas de productividad. [En línea] Disponible: https://hdl.handle.net/20.500.11839/7194 [Acceso: 17 de noviembre del 2020].

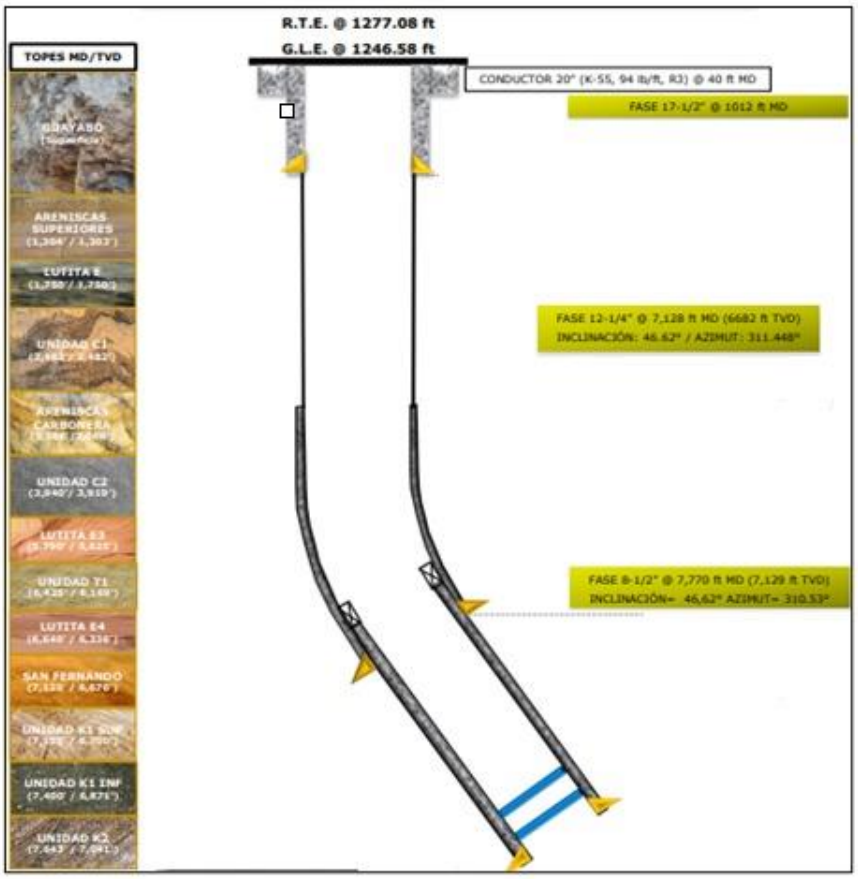
Los tipos de fluidos utilizados durante la fase de perforación para la fase de 17 ½" es lodo SPUD MUD con un peso entre 10-10,5 ppg, para la sección de 12 ¼" lodo polimérico semi disperso con un peso entre 10,5.12,2 ppg y finalmente, en la sección de 8 ½" Drill In desde la formación T2 hasta la formación de interés con un peso entre 8,6-8,9 ppg.[8]

Entre 1975 y 2000, se recuperaron 94 millones de barriles de crudo con un porcentaje de recobro del 4% de reservas.[6] El mecanismo de producción que predomina para las formaciones Une y Gacheta es un acuífero activo.[5]

A continuación, se mostrará el estado mecánico de perforación general para el campo Castilla:

Figura 1.

Estado de mecánico general del campo Castilla y Castilla Norte.



Nota. *Estado mecánico general del campo Castilla y Castilla Norte. Tomado de: Reporte final de perforación [Acceso: 17 de noviembre del 2020].*

1.2. Inteligencia artificial

La inteligencia artificial conocida por sus siglas IA, hace referencia a cualquier" inteligencia similar a la humana exhibida por una computadora o máquina. Es decir que tiene la capacidad para imitar las capacidades de la mente humana como aprender de

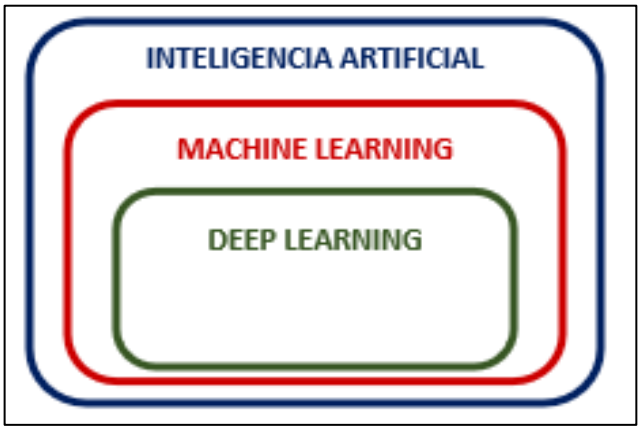
ejemplos y experiencias, reconocer objetos, comprender y responder al lenguaje, tomar decisiones y resolver problemas.” [8]

Esta tecnología viene desde la época de 1943, donde Warren McCulloch y Walter Pits propusieron un modelo de neuronas artificiales dando así inicio a IA. Más adelante en 1950, Alan Turing propuso la prueba de test de Turing, que consistía en verificar la capacidad que tenía la máquina para demostrar un comportamiento inteligente equivalente al de un ser humano. Mas tarde en 1956, se adopta la palabra IA por el científico informático Jhon McCarthy en una conferencia de Dartmouth.

Entre 1966 y 1972, Joseph Weizenbaum desarrollo y creo el primer chatbot que recibió el nombre de ELIZa. Por otra parte, se creó el primer robot humanoide inteligente en Japón que recibió el nombre de WABOT-1. Para el año 1980, la IA emula la capacidad de toma de decisiones de un experto humano. En la actualidad, el uso de IA a nivel mundial ha sido bastante notable, dada la aplicación de conceptos de aprendizaje profundo, big data y ciencia de datos. [9]

Figura 2.

Clasificación de la inteligencia artificial.



Nota. *División de los tipos de inteligencia artificial existentes.*

1.3. Machine Learning

Es una disciplina científica del ámbito de la Inteligencia Artificial y se define como un proceso automatizado que extrae patrones de datos. Los algoritmos de Machine Learning automatizan el proceso de aprendizaje de un modelo que captura la relación entre las características descriptivas y la característica objetivo en un conjunto de datos. [10]

1.3.1. *Tipos de machine Learning*

1.3.1.a. Modelo supervisado. "El aprendizaje supervisado normalmente comienza con un conjunto establecido de datos y un determinado nivel de comprensión sobre cómo se clasifican los datos. El aprendizaje supervisado tiene como objetivo encontrar patrones en los datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos." [11] Ejemplo de lo anterior, se pueden evidenciar en la industria oíl & gas al estimar la vibración en fondo basados en los parámetros de perforación de superficie o aplicando este modelo para la predicción de ocurrencia de pesca.

1.3.1.b. Modelo no supervisado. El aprendizaje no supervisado se implementa "cuando el problema requiere una gran cantidad de datos sin etiquetar, para comprender el significado detrás de estos datos se requiere algoritmos que clasifican los datos basándose en los patrones o clústeres que encuentra. Este tipo de aprendizaje dirige un proceso iterativo, en el cual se analizan datos sin intervención humana." [11] De lo anterior, se pueden evidenciar aplicaciones para la detección de fallas en equipos industriales o en estudios sobre el mecanismo de clasificación y formación de aceite microscópico remanente en la etapa de corte alto de agua.

1.3.2. *Clasificación de machine Learning*

Para efectos de este proyecto, solo se enfocará en dos principales clases. A continuación, se realizará una breve descripción de cada una de ellas y sus aplicaciones:

1.3.2.a. Clasificación. “El algoritmo prueba etiquetar cada ejemplo eligiendo entre dos o más clases diferentes, estos algoritmos crean modelos predictivos a partir de datos de entrenamiento que tienen características y etiquetas de clase. Estos modelos predictivos, a su vez usan las características aprendidas de los datos de entrenamiento sobre nuevos datos, no vistos previamente, para predecir sus etiquetas de clase. Elegir entre dos clases se denomina clasificación binaria y elegir entre más de dos clases se denomina clasificación multiclase.” [12]

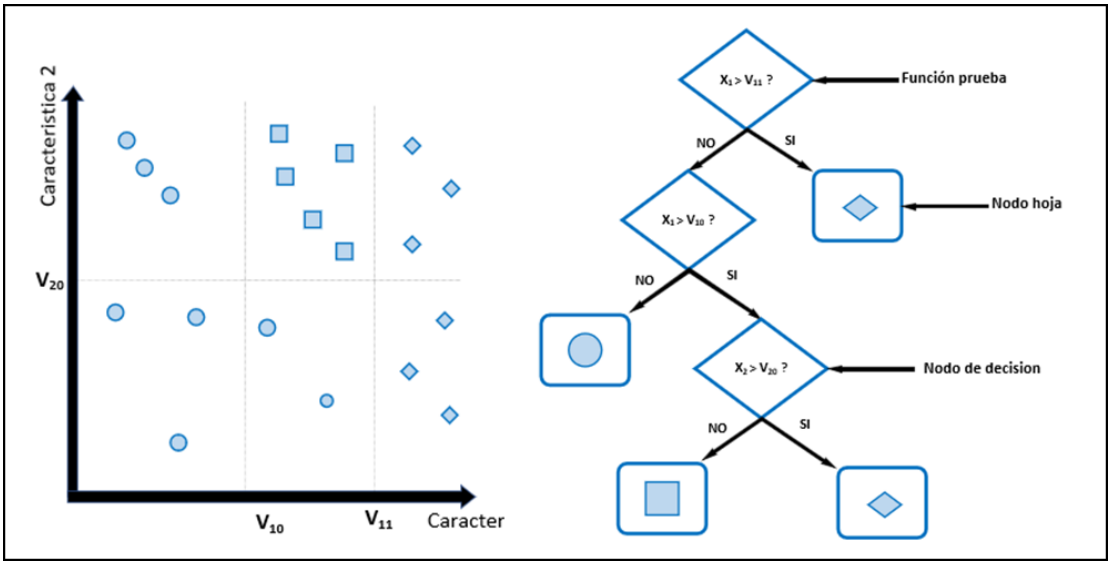
1.3.2.b. Regresión. Este tipo de «algoritmos se usan para predecir valores de salida basados en algunas características de entrada obtenidas de los datos. A esto, el algoritmo construye un modelo basado en las características y los valores de salida de los datos de entrenamiento y este modelo se usa para predecir los valores para nuevos datos. Los valores de salida en este caso son continuos y no discretos.» [12]

1.3.3. *Tipos de modelos de machine Learning*

1.3.3.a. Árbol de decisión. Los árboles de decisión son un método de aprendizaje supervisado no paramétrico, el cual es utilizado para regresión o clasificación. [13] El «objetivo de este algoritmo es crear un modelo que pronostique el valor de una variable objetivo aprendiendo reglas de decisión simples inferidas de las características de los datos. Este algoritmo establece un conjunto de reglas que pueden interpretarse como una estructura anidada Si, adicionalmente, el árbol de decisión utiliza un enfoque de “Caja-blanca” donde la toma de decisiones interna y la estructura del árbol son visibles para el usuario, lo cual hace que sean fáciles de interpretar.» [13] Lo anterior se puede evidenciar en la Figura 3:

Figura 3.

Procedimiento interno árbol de decisión.

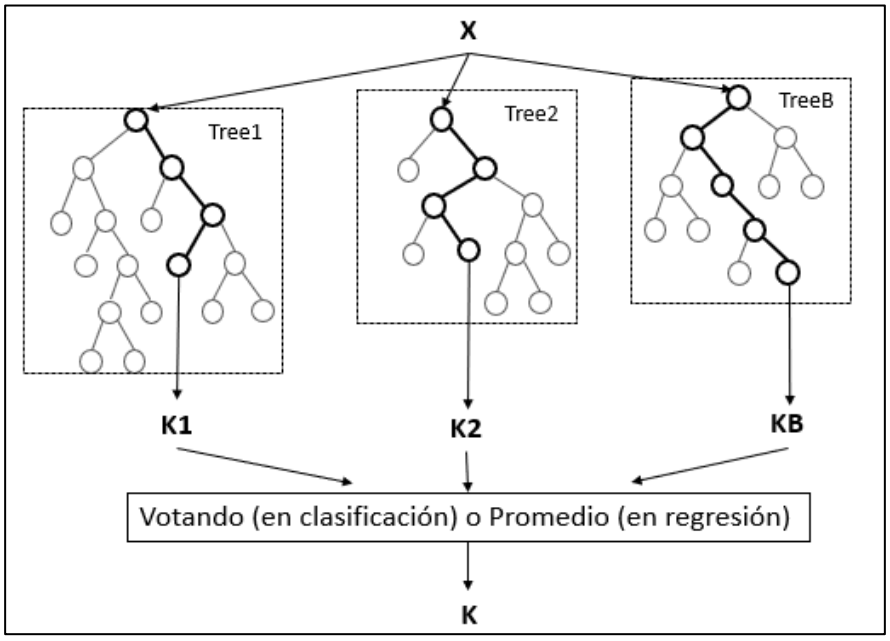


Nota. *Procedimiento interno del modelo árbol de decisión durante su ejecución. Tomado de: Machine Learning Methods for Fault Classification [En línea]. Disponible:* https://core.ac.uk/download/pdf/147542845.pdf *[Acceso: 18 de noviembre del 2020].*

1.3.3.b. Bosques aleatorios. Los bosques aleatorios son un algoritmo de aprendizaje de tipo supervisado el cual crea y fusiona de forma aleatoria varios árboles de decisión creando un bosque [15], este método puede ser implementando tanto para regresión como clasificación. Este algoritmo "funciona de manera eficiente en grandes volúmenes de datos, aporta estimaciones de que variables son relevantes en la clasificación, tienen un método eficaz para la estimación de los datos faltantes y mantiene la precisión cuando una gran parte de los datos faltan" [16], a su vez los bosques aleatorios no presentan sobreajuste dado que los árboles de decisión son independientes uno de los otros. [17] Lo anterior se ejemplifica en la Figura 4.

Figura 4.

Procedimiento interno del modelo bosques aleatorios.



Nota. *Funcionamiento interno del modelo bosques aleatorios durante su ejecución. Tomado de: ResearchGate [En línea]. Disponible:* https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643*. [Acceso: 17 de noviembre del 2020].*

1.3.3.c. Regresión de vectores de soporte. La regresión de vectores de soporte también conocido comúnmente en sus siglas SVR en inglés, «es una variante del modelo de análisis Support Vector Machine (SVM) utilizado para clasificar, sin embargo, con esta variante el modelo de vector soporte se utiliza como un esquema de regresión para predecir valores. En este caso, se establece un margen de tolerancia (épsilon) cerca del vector soporte con el fin de minimizar el error teniendo en cuenta que parte de ese error es tolerado.» [18] Dentro de SVR, se encuentran funciones de tipo lineal como no lineal, la selección del tipo de función dependerá de clase de datos que se van a implementar.

1.4. Herramientas digitales

Dentro de las herramientas digitales que permitieron realizar el presente proyecto se encuentran:

1.4.1. *OpenWells®*

OpenWells® es un aplicativo de Halliburton de la línea LandMark, en el cual se puede consignar las operaciones diarias que llevan a cabo en la fase de perforación y completamiento, para lograr de esta manera llevar un seguimiento de las operaciones que se realizan en campo. El software también ofrece la única interfaz de usuario interactiva en la industria, integrando la base de datos y las herramientas de ingeniería 'LandMark Engineer Data Model™'. Asimismo, esta herramienta agiliza los informes, facilita la recopilación y análisis de datos. [19]

1.4.2. *Power BI*

Es un software gratuito desarrollado por Microsoft, el cual brinda una solución de análisis empresarial que permite visualizar los datos y compartir información con toda la organización, con la opción de reflejar su información en la aplicación web. Conectándose a cientos de orígenes de datos y dando vida a los datos con paneles e informes dinámicos. Esta interfaz permite lograr la toma de decisiones de una manera mucho más rápida y eficaz, donde se podrán integrar con otros gestores de bases de datos como lo son Microsoft Excel y SQL. [20]

1.4.3. *Jupyter notebook*

Es una aplicación web de código abierto donde les permite a los usuarios crear y compartir documentos que contiene ya sea código en vivo, ecuaciones, visualizaciones o texto narrativo. Entre los usos más frecuentes para esta aplicación son la limpieza y transformación de datos, simulación numérica, modelo estadístico, visualización de datos y machine Learning, entre otros usos. [21] Los dos componentes principales de Jupyter Notebook son un conjunto de núcleos (Interpreter) y el Dashboard, cada núcleo

o kernel es un motor de ejecución para un lenguaje que se encarga de procesar las solicitudes y devolver las respuestas apropiadas. [22]

1.4.4. *Python*

«Python es un lenguaje de programación interpretado, orientado a objetos de alto nivel y con semántica dinámica. Su sintaxis hace énfasis en la legibilidad del código, lo que facilita su depuración y, por tanto, favorece la productividad. Ofrece la potencia y la flexibilidad de los lenguajes compilados con una curva de aprendizaje suave. Aunque Python fue creado como lenguaje de programación de uso general, cuenta con una serie de librerías y entornos de desarrollo para cada una de las fases del proceso de Data Science. Esto, sumado a su potencia, su carácter open source y su facilidad de aprendizaje le ha llevado a tomar la delantera a otros lenguajes propios de la analítica de datos por medio de Machine Learning como pueden ser SAS (software comercial líder hasta el momento) y R (también open source, pero más propio de entornos académicos o de investigación).» [23]

1.4.4.a. Scikit-learn. «Es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de mediana escala. Este paquete se centra en llevar el aprendizaje automático a los no especialistas mediante un lenguaje de alto nivel de uso general. Se hace hincapié en la facilidad de uso, el rendimiento, la documentación y la coherencia de la API. Tiene dependencias mínimas y se distribuye bajo la licencia BSD simplificada, lo que fomenta su uso tanto en entornos académicos como comerciales.» [24]

1.4.4.b. Numpy. “Es un proyecto de código abierto que tiene como objetivo permitir la computación numérica con Python. Fue creado en 2005, basándose en los primeros trabajos de las bibliotecas Numerical y Numarray”.[25]

“Es un paquete de Python que significa “Numerical Python”, es la librería principal para la informática científica, proporciona potentes estructuras de datos, implementando matrices y matrices multidimensionales. Estas estructuras de datos garantizan cálculos eficientes con matrices”. [26]

1.4.4.c. Pandas. Pandas es una herramienta desarrollada por Wes McKinney, hace parte de la librería de código abierto de Python que proporciona herramientas de análisis y manipulación de datos de alto rendimiento utilizando sus potentes estructuras de datos. El nombre de Pandas se deriva del término “Panel Data”. [27]

1.4.4.d. Matplotlib. Es una librería de Python desarrollada por Jhon Hunter y otros colaboradores, con el objetivo de “crear gráficos, tablas y figuras de alta calidad en 2D. Entre muchas de sus funciones encontramos la creación de histogramas, espectros de potencia, gráficos de barras, gráficos de error, gráficos de dispersión con tan solo pocas líneas de código”. [28]

1.4.4.e. Seaborn. «Es una librería de visualización de datos para Python desarrollada sobre Matplotlib. Ofrece una interfaz de alto nivel para la creación de atractivas gráficas. Además, está íntimamente integrada con las estructuras de datos de pandas, lo que permite utilizar el nombre de los DataFrames y campos directamente como argumentos de las funciones de visualización. Tiene como objetivo convertir la visualización en una parte central de la exploración y comprensión de los datos, generando atractivas gráficas con sencillas funciones que ofrecen una interfaz semejante, facilitando el paso de unas funciones a otras.» [29]

1.5. Tiempos no productivos

“Cada empresa o prestadora de servicios maneja una definición de Tiempos No Productivos la cual se acopla mejor a sus operaciones, sin embargo, un concepto que puede ser aplicado en el área de perforación sería el tiempo donde no existe un avance en la construcción del pozo o en el que la tasa de perforación es muy baja.” [30] Los tiempos no productivos identificados en el campo castilla corresponden a problemas asociados en hueco abierto, donde los principales problemas son: pegas (geométricas o empaquetamientos), perdidas de fluidos, puntos apretados, inestabilidad del hueco, entre otros.

2. METODOLOGÍA Y DATOS

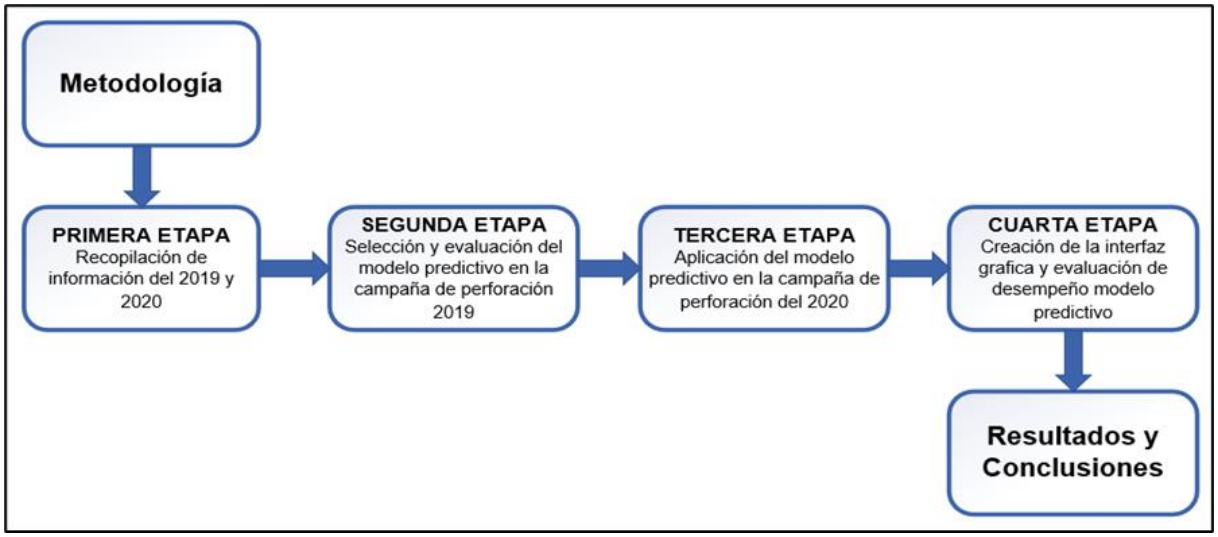
En esta sección se indica la metodología aplicada para el desarrollo del presente proyecto, en el cual se realizó la predicción de días, costos y NPT'S para la campaña de perforación del campo Castilla y Castilla Norte del 2020, basándose en la data técnica-histórica del campo Castilla y Castilla Norte del 2019.

El procedimiento aplicado consistió en la creación de bases de datos que funcionaron de recurso para el modelo predictivo, reuniendo información histórica de la campaña de perforación a partir de una muestra de 57 pozos de los campos Castilla y Castilla Norte 2019. Posteriormente, se llevó a cabo una búsqueda bibliográfica de diferentes clases de modelos predictivos de tipo supervisado, implementándolos sobre la información recopilada del 2019 y seleccionando los modelos con mayor desempeño con respecto a las métricas de regresión. Consecutivamente, se aplicaron los modelos seleccionados sobre la información de los campos Castilla y Castilla Norte del 2020 donde se obtuvieron las predicciones para días, costos y NPT'S. Más adelante, se creó una base de datos que recopilo las predicciones del 2020 y a su vez información técnica tanto planeada como ejecutada. Finalmente, se exportó la base de datos 2020 final al visualizador Power BI para la creación de un tablero dinámico donde se evaluó el desempeño del modelo predictivo comparando los valores de lo planeado versus lo ejecutado versus lo pronosticado.

Para dar cumplimiento a los objetivos específicos planteados, se dividió la metodología en 4 etapas como se observa en la **Figura 5.**

Figura 5.

Metodología para la implementación de los modelos predictivos.



Nota. La figura muestra el procedimiento para llevar a cabo la predicción de tiempos, costos y NPT's.

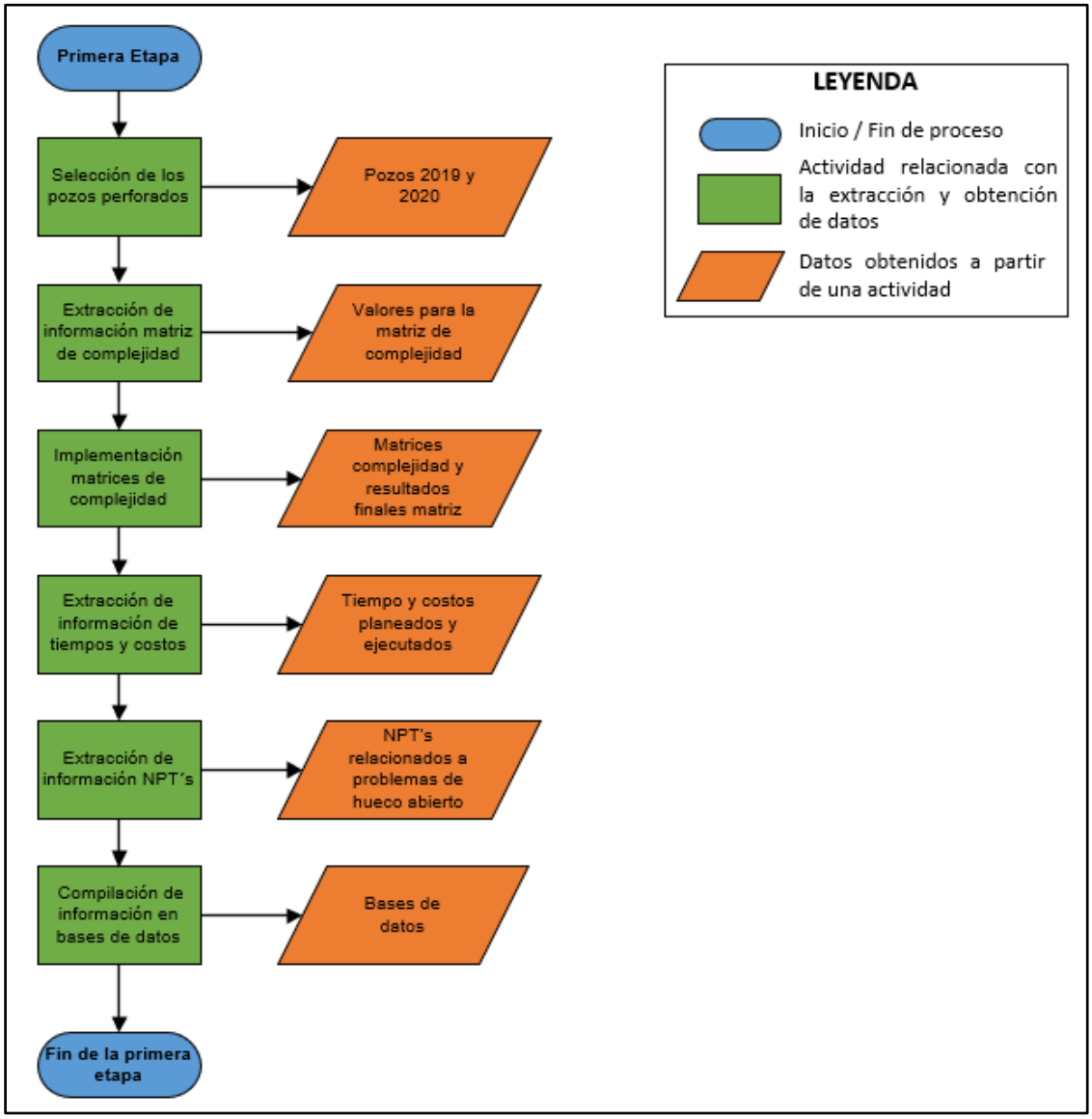
2.1. Primera etapa: recopilación de información del 2019 y 2020

Para dar cumplimiento al primer objetivo específico, se llevaron a cabo diferentes procesos (subetapas) con el propósito de recopilar y almacenar la información de los pozos perforados durante el 2019 y 2020, consecutivamente, se recolecto información relacionada con las variables de la matriz de complejidad para su posterior implementación. Más adelante, sé recolecto información de tiempos y costos tanto planeados como ejecutados y se extrajo información concerniente a tiempos no productivos asociados a problemas en hueco abierto.

Los diferentes datos recolectados anteriormente fueron almacenados en bases de datos, debido a que estas fueron implementadas para la aplicación de los modelos predictivos. En la **Figura 6**, se muestra el procedimiento llevado a cabo durante la primera etapa:

Figura 6.

Procedimiento primera etapa.



Nota. La figura muestra los procedimientos para llevar a cabo la primera etapa.

2.1.1. *Selección de los pozos perforados del 2019 y 2020.*

Para llevar a cabo la selección de pozos de la campaña de perforación de los campos Castilla y Castilla Norte, se utilizaron aplicaciones como OpenWells y DataAnalyzer de la línea LandMark-Haliburton. OpenWells es alimentado diariamente por el personal relacionado con la operación con el fin de llevar la trazabilidad de los proyectos, la

anterior información queda almacenada en la sección Wellbore, asociado al evento ODR; allí se consignan las diferentes actividades llevadas a cabo durante la fase de perforación. Por otro lado, DataAnalyzer permite a los ingenieros comparar y contrastar cualquier tipo de información que este almacenada en la base de datos EDM, en donde se encuentra la información compilada proveniente de OpenWells, mediante el aplicativo DataAnalyzer se realizaron consultas utilizando un filtro de tiempo sobre la base de datos EDM con el fin de obtener la lista de pozos perforados.

La campaña de perforación del año 2019 estuvo conformada por 72 pozos, de los cuales 6 pozos que eran Side-Track no se tuvieron en cuenta dado que se consideraron como valores atípicos, los cuales generaban dispersión en los datos. Adicionalmente, debido a la escasez de información, 9 pozos no se tuvieron en cuenta. Por lo tanto, solo se contó con 57 pozos para trabajar los modelos predictivos.

Por otra parte, en el 2020 se tenía contemplada una campaña de perforación más ambiciosa pero dada la contingencia sanitaria, solamente se lograron perforar 5 pozos.

En la **Tabla 2**, se observa la muestra de pozos para la campaña de perforación del 2019 y 2020.

Tabla 2.

Campañas de perforación 2019 y 2020.

Campos	Pozos 2019	Pozos 2020
Castilla	21	1
Castilla Norte	36	4
Total	**57**	**5**

Nota. *Esta tabla resume la cantidad de pozos que se trabajaron durante este proyecto de grado.*

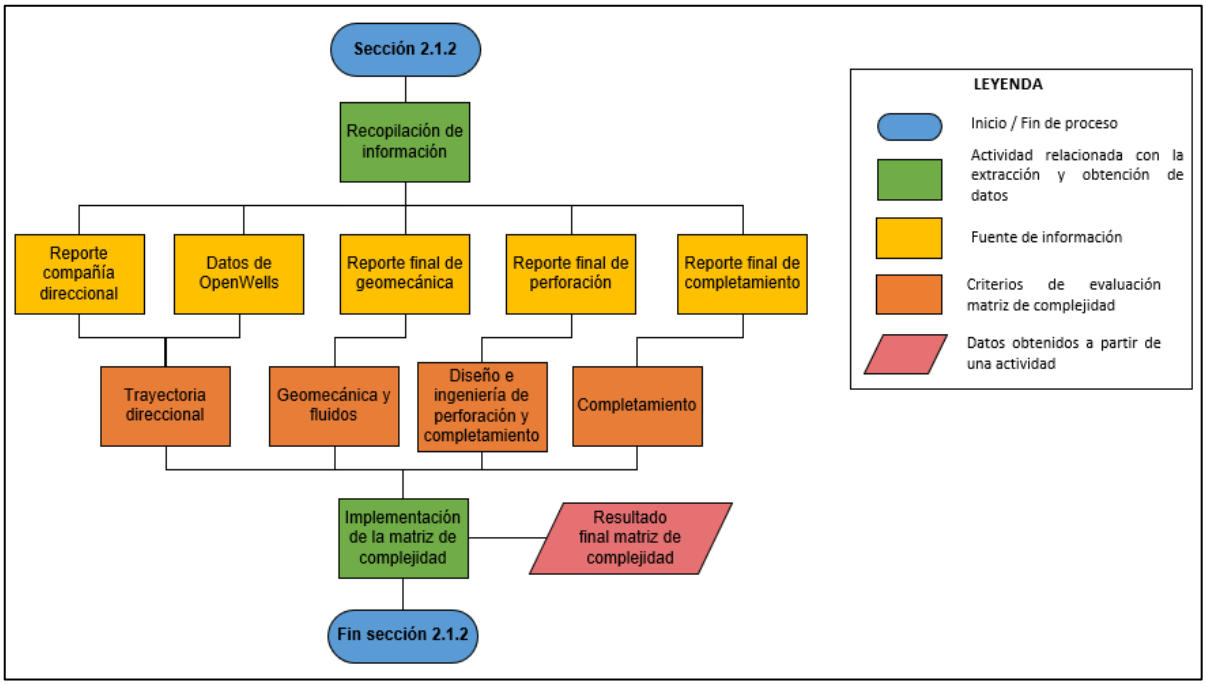
En el **ANEXO A** y **ANEXO B**, se encontrarán las listas de pozos discriminados para cada año y campo perforado respectivamente.

2.1.2. *Extracción de información variables implicadas en la matriz de complejidad y respectiva implementación para la campaña del 2019 y 2020*

Para llevar a cabo la sección 2.1.2 se desarrolló un proceso para la obtención y recopilación de información, posteriormente se implementó la matriz de complejidad con la finalidad de obtener el resultado para cada pozo y así lograr clasificar por categoría el grado de complejidad del pozo durante la fase de planeación en el área de perforación. En la **Figura 7**, se describe el proceso anteriormente mencionado:

Figura 7.

Implementación y obtención del resultado de la matriz de complejidad.



Nota. La figura representa el procedimiento para la implementación y obtención del resultado final de la matriz de complejidad.

La matriz de complejidad se encuentra conformada por las siguientes secciones y variables como se muestra en la **Tabla 3**:

Tabla 3.

Variables implicadas en la matriz de complejidad.

Sección	Variable	Unidad
Trayectoria direccional	Factor de separación	Adimensional
	DDI	Adimensional
	VS/TVD	Adimensional
Geomecánica y fluidos	Complejidad estructural	Adimensional
	Buzamiento capa geológicas	Grados
	Angulo de ataque	Grados
	Presión y Temperatura	Psi -°F
	Gradiente de presión de yacimiento	ppg
	Max densidad en el Overburden	ppg
	Régimen esfuerzos inSitu	Adimensional
	Dureza de la roca	Psi
Diseño e ingeniería de perforación y completamiento	Clasificación Lahee	Adimensional
	Presión parcial H2S & CO2	Psi
	Complejidad diámetro final del pozo	In
	Profundidad medida del pozo	Ft
	Servicio del Pozo	Adimensional
	Número de secciones del pozo	Adimensional
Completamiento y producción	Máxima densidad fluido de control	ppg
	Tipo de completamiento inferior	Adimensional
	Tipo de completamiento Superior	Adimensional
	GOR	scf/stb
	Pruebas de producción	Adimensional
Resultado final Matriz		Adimensional

Nota. La tabla muestra las variables que componen la matriz de complejidad con sus respectivas unidades.

Una vez seleccionados los pozos en la sección 2.1.1, se procedió a la recolección de las variables implicadas en la matriz de complejidad para cada pozo, esta información es de carácter privada y fue suministrada por la operadora y compañías de servicio. La variable factor de separación se tomó del reporte final de planeación de la compañía direccional, el DDI se extrajo de una consulta por medio de DataAnalyzer**,** las variables vertical section (VS) y profundidad vertical verdadera (TVD) se consiguieron del reporte final de

perforación, la relación VS/TVD se obtuvo mediante la división de VS sobre TVD, las variables geomecánicas (Complejidad estructural, buzamiento capa geológica, ángulo de ataque, densidad en el overburden, régimen de esfuerzos en situ, compresibilidad de la roca y datos de presión y temperatura) se extrajo del reporte final de geomecánica, las variables de diseño de ingeniería de perforación y completamiento (Diámetro final del pozo, longitud medida final, número de secciones, clasificación Lahee y datos de presión parcial para el H_2S y el CO_2) se tomaron del reporte final de perforación**,** las variables de completamiento (Densidad de fluido de control, completamiento a instalar, producción, pruebas de producción) se obtuvieron del reporte final de completamiento.

Con la información anteriormente recopilada, se realizó la implementación de la matriz de complejidad para cada pozo con el objetivo de obtener el resultado final de dicha matriz como se muestra en la **Figura 8**:

Figura 8.

Resultado final de la matriz de complejidad.

Nota. La figura representa el resultado final de la matriz de complejidad, categorizándolo según su complejidad en 4 intervalos.

2.1.3. *Extracción de tiempos y costos*

En esta sección se recopilaron los valores de días y costos ejecutados para los pozos del campo Castilla y Castilla Norte 2019 seleccionados previamente, luego se reunieron los días y costos tanto planeados como ejecutados para la campaña de perforación 2020 del anterior campo mencionado. Dicha información fue suministrada por la compañía y se obtuvieron mediante consultas del aplicativo DataAnalyzer, el cual extrajo la información que se encuentra en la base de datos de la herramienta OpenWells.

2.1.4. *Extracción de los NPT'S*

Para la extracción y recopilación de los NPT'S asociados a problemas en hueco abierto, se utilizó el programa de Microsoft Power BI, específicamente el archivo *GPN-GIF001-REPORTE OPTIMIZACIÓN ECOPETROL*, dado que la compañía implementa esta herramienta para visualizar información como: pozos, compañías de servicio, costos, tiempos y NPT'S concernientes al área de perforación para evaluar el desempeño de esta área y realizar posteriormente un análisis de dicha información.

Una vez dentro de la interfaz del reporte anteriormente mencionado, en su página de inicio, se buscó la opción de perforación, luego se seleccionó entre sus opciones la categoría de NPT'S donde se filtró la información por tiempos no productivos asociados a problemas en hueco abierto. Posteriormente, se extrajo información concerniente al tipo, descripción y duración NPT para cada pozo. Cabe resaltar que el archivo mencionado, es de uso exclusivo de la operadora.

2.1.5. *Creación de base de datos general y especificas*

Durante esta etapa, se invirtió gran cantidad del tiempo específicamente en la recopilación de la información para cada variable de la base de datos, debido a la complejidad para la obtención de estos datos. Adicionalmente, se garantizó la veracidad de esta base de datos al ser completada puntualmente pozo a pozo.

Después de haber recopilado la información en las secciones 2.1.1, 2.1.2, 2.1.3 y 2.1.4, se procedió a crear una base de datos general con dicha información. Luego, se seleccionaron las variables más representativas para la aplicación de los modelos predictivos, durante este proceso no se tuvieron en cuenta las variables de completamiento y producción, debido a que estas variables no tienen ningún tipo de influencia sobre la predicción de tiempos y costos sobre la *fase de perforación*, asimismo, no se tuvieron en cuenta aquellas variables que fueran constantes en la base de datos, debido a que estos al ser repetitivos no contribuyen al modelo predictivo.

Los datos recopilados fueron almacenados en una base de datos general como se muestra en la **Tabla 4**, que posteriormente se refino y dividió en 2 bases de datos específicas como se muestran en la **Tabla 5** y **Tabla 6**

Tabla 4.

Base de datos general.

Variable	Unidad
Well Name	Adimensional
Factor de separación	Adimensional
DDI	Adimensional
VS/TVD	Adimensional
Complejidad estructural	Adimensional
Buzamiento capa geológicas	Grados
Ángulo de ataque	Grados
Presión y Temperatura	Psi -°F
Gradiente de presión de yacimiento	ppg
Max densidad en el Overburden	ppg
Régimen esfuerzos inSitu	Adimensional
Dureza de la roca	Psi
Clasificación Lahee	Adimensional
Presión parcial H2S & CO2	Psi
Complejidad diámetro final del pozo	In
Profundidad medida del pozo	Ft
Servicio del Pozo	Adimensional
Número de secciones del pozo	Adimensional
Máxima densidad fluido de control	ppg
Tipo de completamiento inferior	Adimensional
Tipo de completamiento Superior	Adimensional
GOR	scf/stb
Pruebas de producción	Adimensional
Resultado final Matriz	Adimensional
Días ejecutados perforación	Días
Costo ejecutado perforación	US$
Tipo de NPT	Adimensional
Descripción del NPT	Adimensional
Duración NPT	Horas

***Nota*.** La tabla muestra las variables almacenadas para el 2019 y 2020.

Tabla 5.

Base de datos 2019.

Variable	Unidad
Factor de separación	Adimensional
DDI	Adimensional
VS/TVD	Adimensional
Complejidad estructural	Adimensional
Buzamiento capa geológicas	Grados
Angulo de ataque	Grados
Gradiente de Presión de yacimiento	ppg
Max densidad en el Overburden	ppg
Dureza de la roca	Psi
Profundidad medida del pozo	Ft
Resultado Final Matriz	Adimensional
Días ejecutados perforación	Días
Costo ejecutado perforación	US$
Duración NPT	Horas

Nota. La tabla muestra las variables seleccionadas para el año 2019.

Tabla 6.

Base de datos 2020.

Variable	Unidad
Factor de separación	Adimensional
DDI	Adimensional
VS/TVD	Adimensional
Complejidad estructural	Adimensional
Buzamiento capa geológicas	Grados
Angulo de ataque	Grados
Gradiente de Presión de yacimiento	ppg
Max densidad en el Overburden	ppg
Dureza de la roca	Psi
Profundidad medida del pozo	Ft
Resultado Final Matriz	Adimensional
Días planeados perforación	Días
Costos planeados perforación	US$
Días ejecutados perforación	Días
Costo ejecutado perforación	US$
Duración NPT	Horas

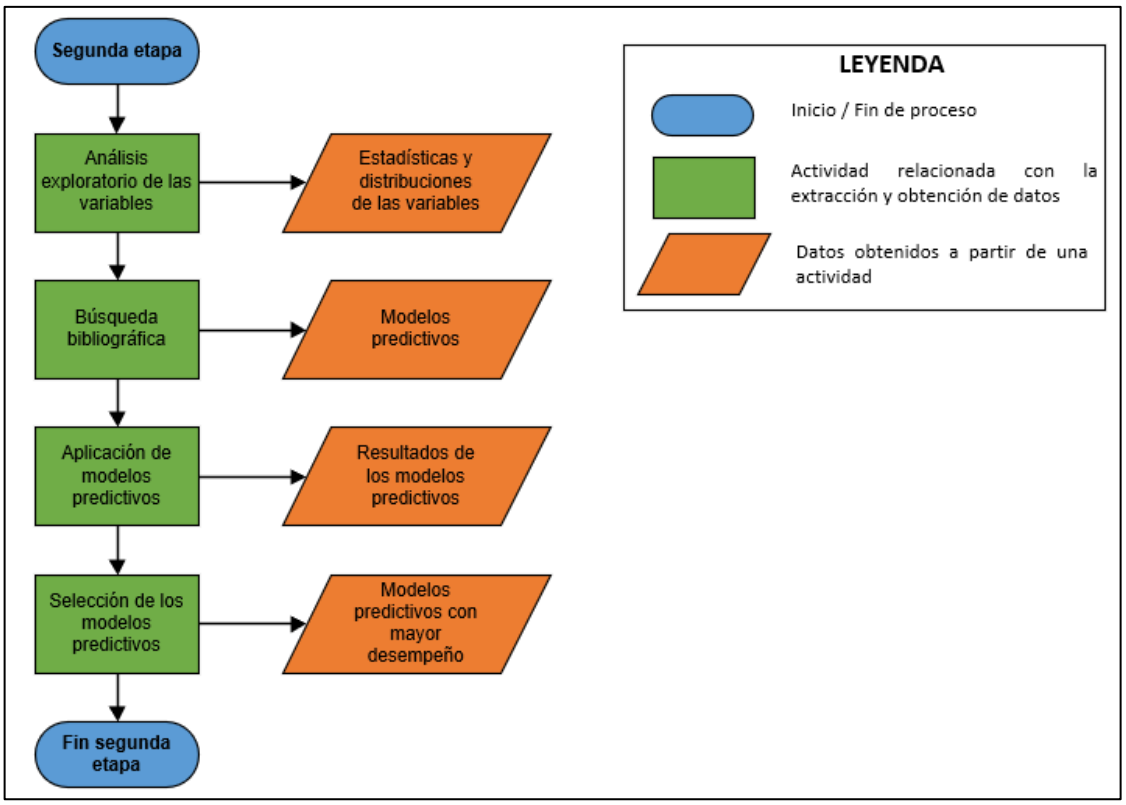
Nota. En esta tabla se muestran las variables seleccionadas para el año 2020.

2.2. Segunda etapa: selección y evaluación del modelo predictivo en la campaña de perforación del 2019.

Para dar cumplimiento al segundo y tercer objetivo del presente proyecto, se llevó a cabo un análisis exploratorio sobre la base de datos del 2019 en donde se encuentra las variables implicadas en la matriz de complejidad y el resultado de la misma, adicionalmente se encuentran los valores de días, costos y NPT's. Más adelante, se realizó una búsqueda bibliográfica concerniente a los modelos predictivos. Luego, se seleccionaron y evaluaron los modelos predictivos con el fin de comparar cuál de los tres fue el más eficiente. En la **Figura 9**, se muestra el procedimiento llevado a cabo durante la segunda etapa:

Figura 9.

Procedimiento segunda etapa.



Nota. La figura representa el procedimiento para llevar a cabo la segunda etapa.

2.2.1. *Análisis exploratorio de las variables*

De acuerdo a la información obtenida en la sección 2.1.5, se realizó un análisis exploratorio con el propósito de identificar el tipo de variables a manejar, asimismo criterios como: la mediana, desviación estándar, valor mínimo y máximo, rangos intercuartílicos, valores atípicos y la relación entre las variables. Lo anterior, se logró mediante gráficos tales como caja y bigotes, grafico de parejas y mapa de correlación.

Este procedimiento fue llevado a cabo importando la base de datos 2019 al entorno de Jupyter Notebook, haciendo uso de la librería Pandas de Python. Para cada variable que se recopilo en la base de datos 2019 se le asigno una variable equivalente en Python como se muestra en la siguiente tabla:

Tabla 7.

Equivalentes de las variables matriz complejidad en Python.

Variable	Variable equivalente Python	Unidad
Factor de separación	Fact_Separacion	Adimensional
DDI	DDI	Adimensional
VS/TVD	VS/TVD	Adimensional
Complejidad estructural	Fallas	Adimensional
Buzamiento capa geológicas	Buzamiento	Grados
Angulo de ataque	Angulo_Ataque	Grados
Gradiente de Presión de yacimiento	Grad_Yac	ppg
Max densidad en el Overburden	Maxd_Over	ppg
Dureza de la roca	Dureza_Roca	Psi
Profundidad medida del pozo	MD_Final	Ft
Resultado Final Matriz	Final_Matriz	Adimensional
Días ejecutados perforación	Dias_E	Días
Costo ejecutado perforación	Costos_E	US$
Duración NPT	NPT_E	Horas

Nota. Esta figura muestra los nombres equivalentes que se le asignaron a las variables de la matriz de complejidad con sus respectivas unidades.

Inicialmente se aplicó la función *.info* sobre el conjunto de datos, a partir de esto se obtuvo el total de filas y columnas, valores no nulos y el tipo de variable numérica (Entero o decimal). Lo anterior, se muestra en la **Figura 10**:

Figura 10.

Información base de datos 2019.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 57 entries, 0 to 56
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Fact_Separacion  57 non-null     float64
 1   DDI              57 non-null     float64
 2   VS/TVD           57 non-null     float64
 3   Fallas           57 non-null     int64
 4   Buzamiento       57 non-null     float64
 5   Angulo_Ataque    57 non-null     float64
 6   Grad_Yac         57 non-null     float64
 7   Maxd_Over        57 non-null     float64
 8   Dureza_Roca      57 non-null     float64
 9   MD_Final         57 non-null     int64
 10  Final_Matriz     57 non-null     float64
 11  Dias_E           57 non-null     float64
 12  Costos_E         57 non-null     int64
 13  NPT_E            57 non-null     float64
dtypes: float64(11), int64(3)
```

Nota. La figura muestra el manejo de 3 variables con números enteros y 11 variables con números decimales, finalmente muestra que no hay variables categóricas.

A partir de la **Figura 10** se puede observar que no se encuentran valores faltantes en la base de datos del 2019, esto se debe a que la base de datos mencionada se completó pozo a pozo, variable a variable con lo cual se garantizó la integridad del conjunto de datos.

A continuación, se implementó la función *.describe* que permitió obtener una estadística descriptiva de las variables a trabajar como se muestra en la **Figura 11.**

Figura 11.

Estadística descriptiva bases de datos 2019.

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Fact_Separacion	57.0	1.655614e+00	0.695280	3.400000e-01	1.250000e+00	1.590000e+00	1.950000e+00	3.920000e+00
DDI	57.0	5.350877e+00	0.217824	4.650000e+00	5.250000e+00	5.380000e+00	5.510000e+00	5.730000e+00
VS/TVD	57.0	3.462734e-01	0.110214	1.326778e-01	2.797760e-01	3.377789e-01	4.219937e-01	5.948437e-01
Fallas	57.0	5.789474e-01	0.754685	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	3.000000e+00
Buzamiento	57.0	6.946316e+00	6.265256	0.000000e+00	3.510000e+00	6.000000e+00	8.000000e+00	3.590000e+01
Angulo_Ataque	57.0	2.858211e+01	11.983267	9.800000e-01	1.970000e+01	3.080000e+01	3.650000e+01	4.970000e+01
Grad_Yac	57.0	8.368421e+00	0.714432	5.400000e+00	8.300000e+00	8.400000e+00	8.600000e+00	1.060000e+01
Maxd_Over	57.0	1.209860e+01	0.127775	1.200000e+01	1.200000e+01	1.200000e+01	1.220000e+01	1.240000e+01
Dureza_Roca	57.0	1.165711e+04	4081.490827	5.464250e+03	9.919450e+03	1.070703e+04	1.238285e+04	2.970110e+04
MD_Final	57.0	8.452982e+03	550.970264	7.300000e+03	8.043000e+03	8.435000e+03	8.850000e+03	9.750000e+03
Final_Matriz	57.0	8.066667e+00	1.269889	5.800000e+00	7.400000e+00	8.000000e+00	8.800000e+00	1.180000e+01
Dias_E	57.0	2.438070e+01	10.560108	1.420000e+01	1.850000e+01	2.100000e+01	2.580000e+01	6.790000e+01
Costos_E	57.0	3.111749e+06	862443.072228	2.228947e+06	2.609009e+06	2.872702e+06	3.263788e+06	7.329876e+06
NPT_E	57.0	4.456140e+01	103.004034	0.000000e+00	0.000000e+00	2.500000e+00	3.500000e+01	4.750000e+02

Nota. La figura muestra un resumen estadístico-descriptivo de las variables seleccionadas.

En la **Figura 11** se puede apreciar la cantidad de datos, mediana, desviación estándar, los valores mínimos y máximos, rangos intercuartílicos para cada variable respectivamente.

Para visualizar las variables, fue necesario realizar una estandarización de las mismas con el propósito de llevar todas a una misma escala o rango de medición determinado, lo anterior, se logró utilizando funciones y códigos como se muestra en la **Figura 12:**

Figura 12.

Funciones y códigos para estandarización de la base de datos 2019.

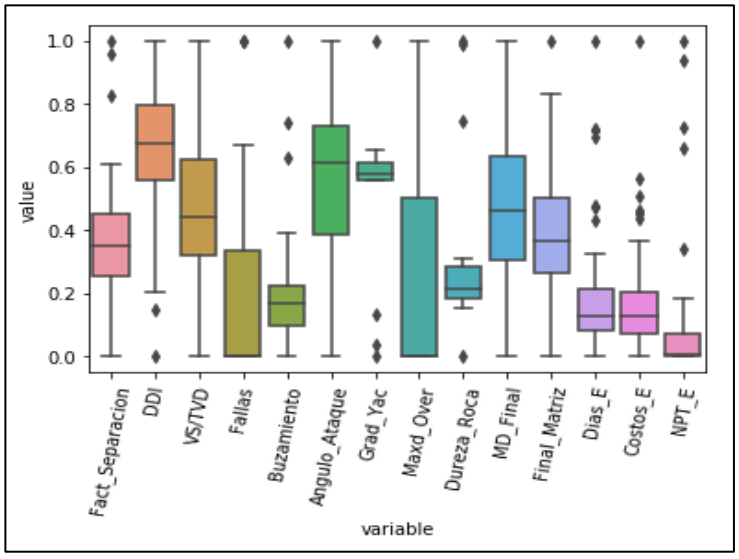
```
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing

columns_to_norm=['Fact_Separacion', 'DDI', 'VS/TVD', 'Fallas', 'Buzamiento',
       'Angulo_Ataque', 'Grad_Yac','Maxd_Over', 'Dureza_Roca', 'MD_Final', 'Final_Matriz',
        'Dias_E','Costos_E','NPT_E']

min_max_scaler = preprocessing.MinMaxScaler() # se encarga de normalizar una variable entre 0 y 1
df[columns_to_norm]=min_max_scaler.fit_transform(df[columns_to_norm])
dfn=df.iloc[:,0:14]
```

Nota. La figura muestra el procedimiento llevado a cabo para la estandarización la base de datos del 2019.

Con la base de datos del 2019 estandarizada, se procedió a representar los datos mediante gráficos estadísticos que permitieron una mejor comprensión del conjunto de datos, como se muestran en las siguientes figuras:

Figura 13.

Diagrama de caja y bigotes.



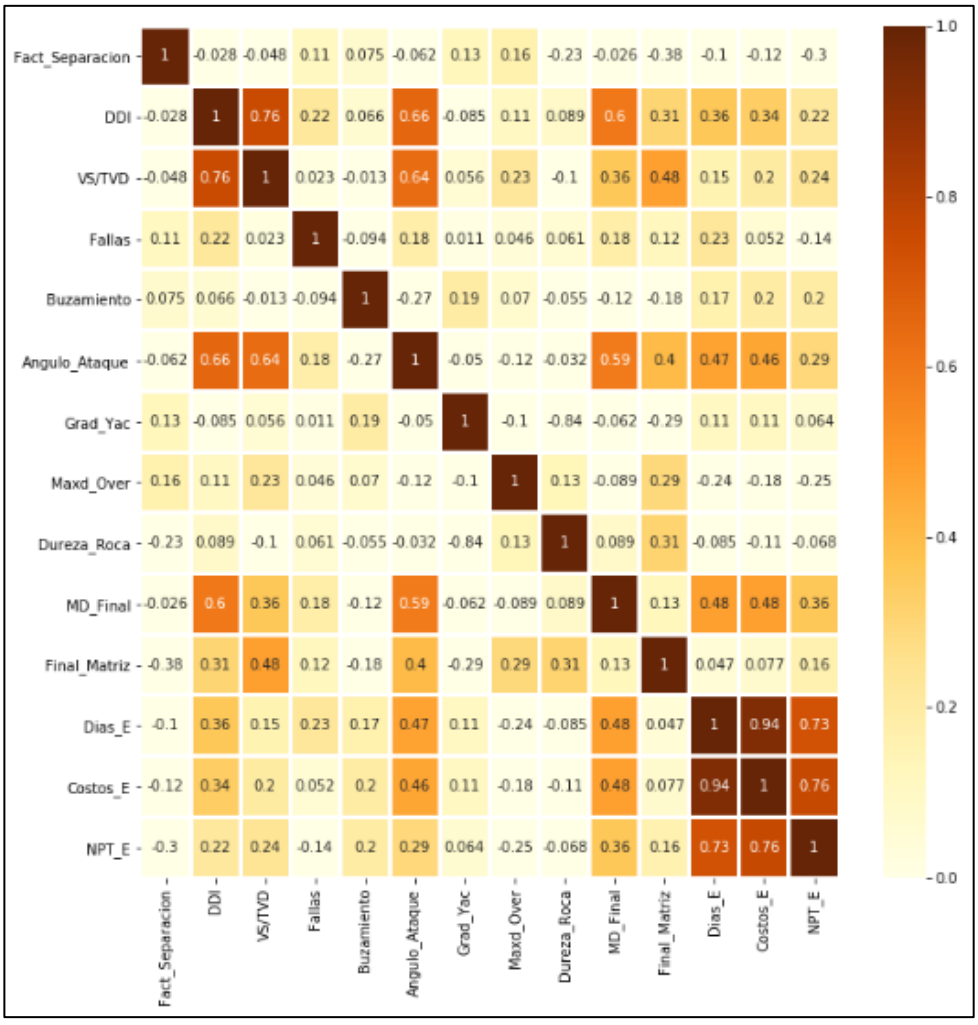
Nota. La figura representa los 3 rangos intercuartílicos, valores mínimos, máximos y valores atípicos de cada variable.

En la **Figura 13** se puede apreciar la distribución de los datos en donde los rombos representan los valores atípicos, las cajas representan los rangos intercuartílicos, los bigotes simbolizan los limites mínimos y máximos para cada variable respectivamente.

Posteriormente, se visualizó el conjunto de datos del 2019 mediante un mapa de correlación como se puede observar en la **Figura 14**, el cual nos permitió identificar la correlación que hay entre las variables.

Figura 14.

Diagrama de mapa de correlación.

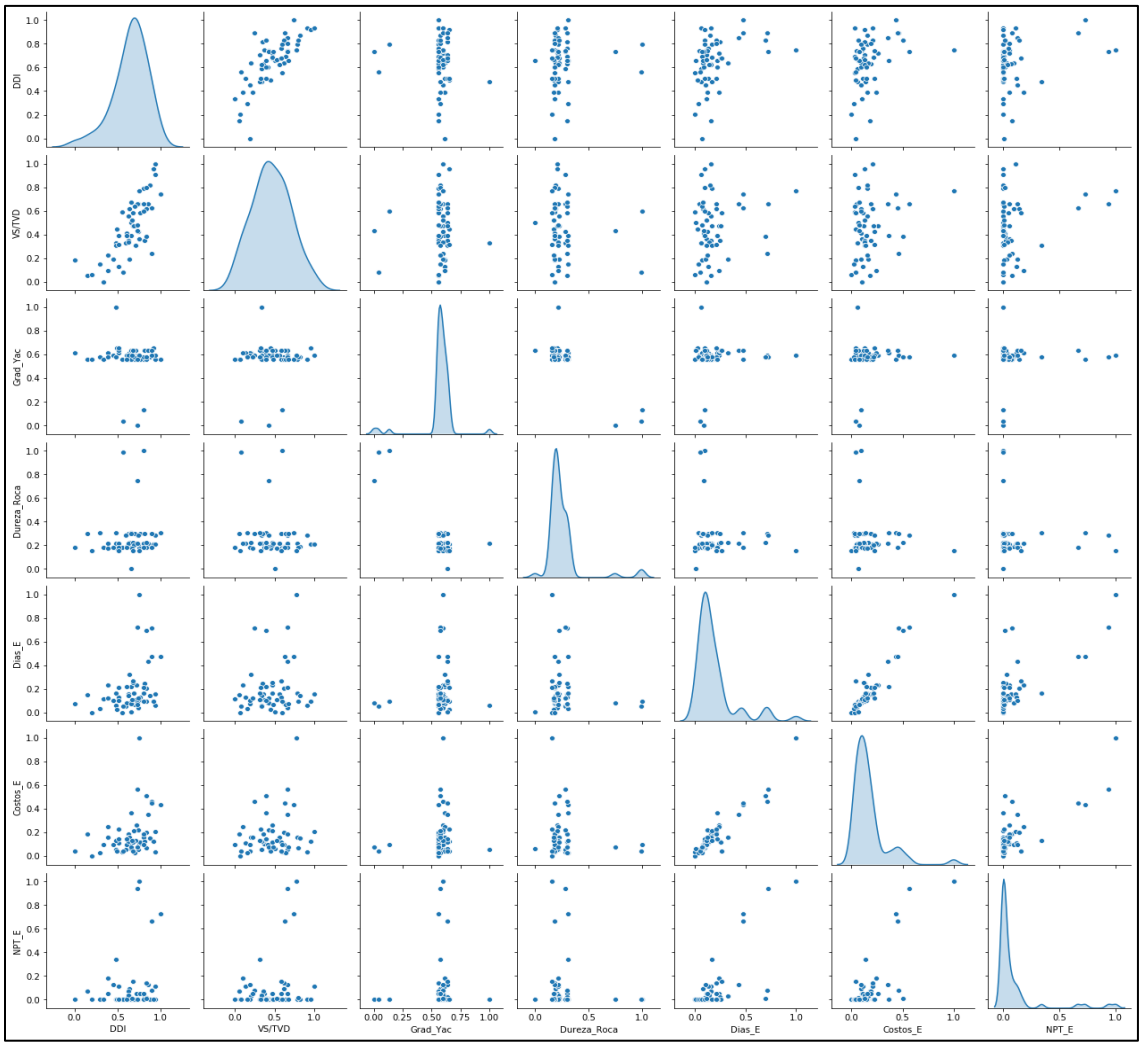


Nota. En esta figura se observa el cálculo de correlación entre cada una de las variables.

En la **Figura 14,** se observó que las variables predictoras DDI y VS/TVD presentaron una correlación positiva > 0.7, es decir que presentan una relación directamente proporcional, esto quiere indicar que una variable puede explicar la otra. Por otra parte, la variable Gradiente de Yacimiento y Dureza Roca presentaron una correlación negativa de -0.84 lo que nos indica que tienen una relación inversamente proporcional. Adicionalmente, se espera que los modelos predictivos presenten una alta correlación entre días, costos y NPT's dado que a mayor cantidad de días y horas de NPT's, mayor será el costo asociado.

Figura 15.

Diagrama de parejas.



Nota. La figura representa la relación entre cada una de las variables.

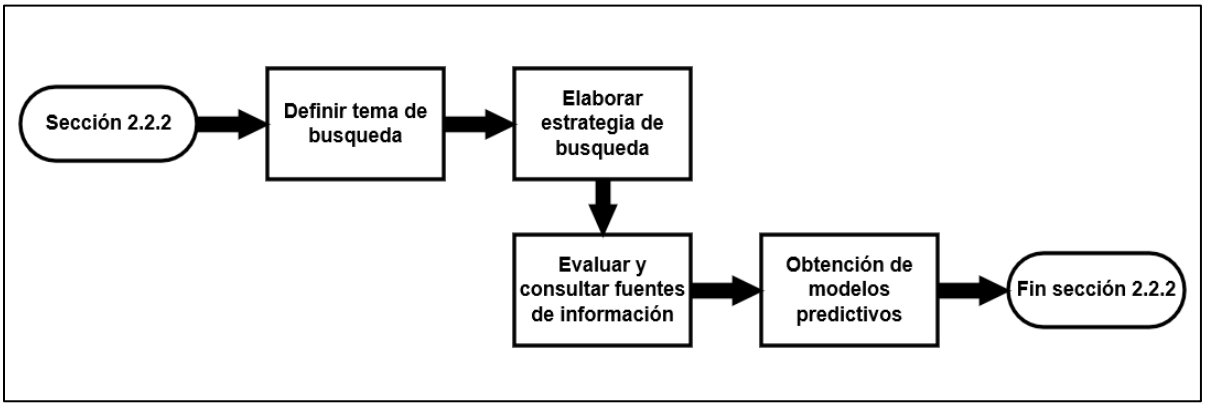
En la **Figura 15**, al comparar las variables de la base de datos del 2019, se observa el comportamiento entre estas y la distribución de los valores (Patrones), en donde en la diagonal se muestra una distribución univariante para mostrar los datos marginales en cada columna, adicionalmente nos muestra los datos atípicos.

2.2.2. *Búsqueda bibliográfica y selección de los modelos predictivos*

Con base a la información previamente analizada en la sección 2.2.1, se llevó a cabo una búsqueda bibliográfica como se muestra en la **Figura 16**:

Figura 16.

Procedimiento para la búsqueda bibliográfica.



Nota. La figura representa el procedimiento para la selección de los 3 modelos predictivos.

El primer paso consistió en la definición del tema sobre el cual se iba a investigar, para este proyecto fueron los modelos predictivos de regresión de tipo supervisado. Como segundo paso, se elaboró una estrategia de búsqueda mediante la identificación de palabras claves y operaciones lógicas (OR y AND) con la finalidad de encontrar los resultados acordes al tema en cuestión. Más adelante, en el tercer paso se evaluó y consulto fuentes de información basándose en criterios como: procedencia, veracidad, calidad y fecha de publicación de la información. Finalmente, en el cuarto paso se seleccionaron los 3 modelos y durante este paso se tuvo en cuenta las ventajas y desventajas de cada modelo seleccionado (DecisionTreeRegressor, RandomForestRegressor, SupportVectorRegressor) como se muestra en la **Tabla 8.**

Tabla 8.

Ventajas y desventajas de los modelos predictivos seleccionados.

Modelo	Ventajas	Desventajas
DecisionTreeRegressor	- Requiere poca preparacion de los datos. - Capaz de manejar datos tanto numericos como categoricos - Sencillo de entender e interpretar. - Es posbile validar el modelo mediante pruebas estadisticas. - Utiliza un modelo de caja blanca.	- Puede generar arboles complejos que no generalicen adecuadamente los datos. - Pueden ser inestables con pequeñas variaciones en los datos.
RandomForestRegressor	- Es un metodo preciso y robusto debido a la cantidad de arboles de decision que participan en el proceso - No se ve afectado por sobreajuste. - Puede ser implementado tanto para regresion como clasificacion. - Puede manejar valores perdidos. - Puede manejar un gran rango de variables e identificar cuales son las significativas.	- Son lentos para generar predicciones dado que tienen multiples arboles de decision. - El modelo es dificil de interpretar en comparacion con un arbol de decision.
SupportVectorRegressor	- Eficaz en espacio de gran dimension. - Utiliza vectores de soporte. - Son rapidos para generar predicciones debido al poco espacio que ocupa en memoria. - Puede especificar diferentes funciones kernel	- Si el numero de caracteristicas es mucho mayor que el numero de muestras, ocurre un sobreajuste. - No proporciona directamente estimaciones de probabilidad.

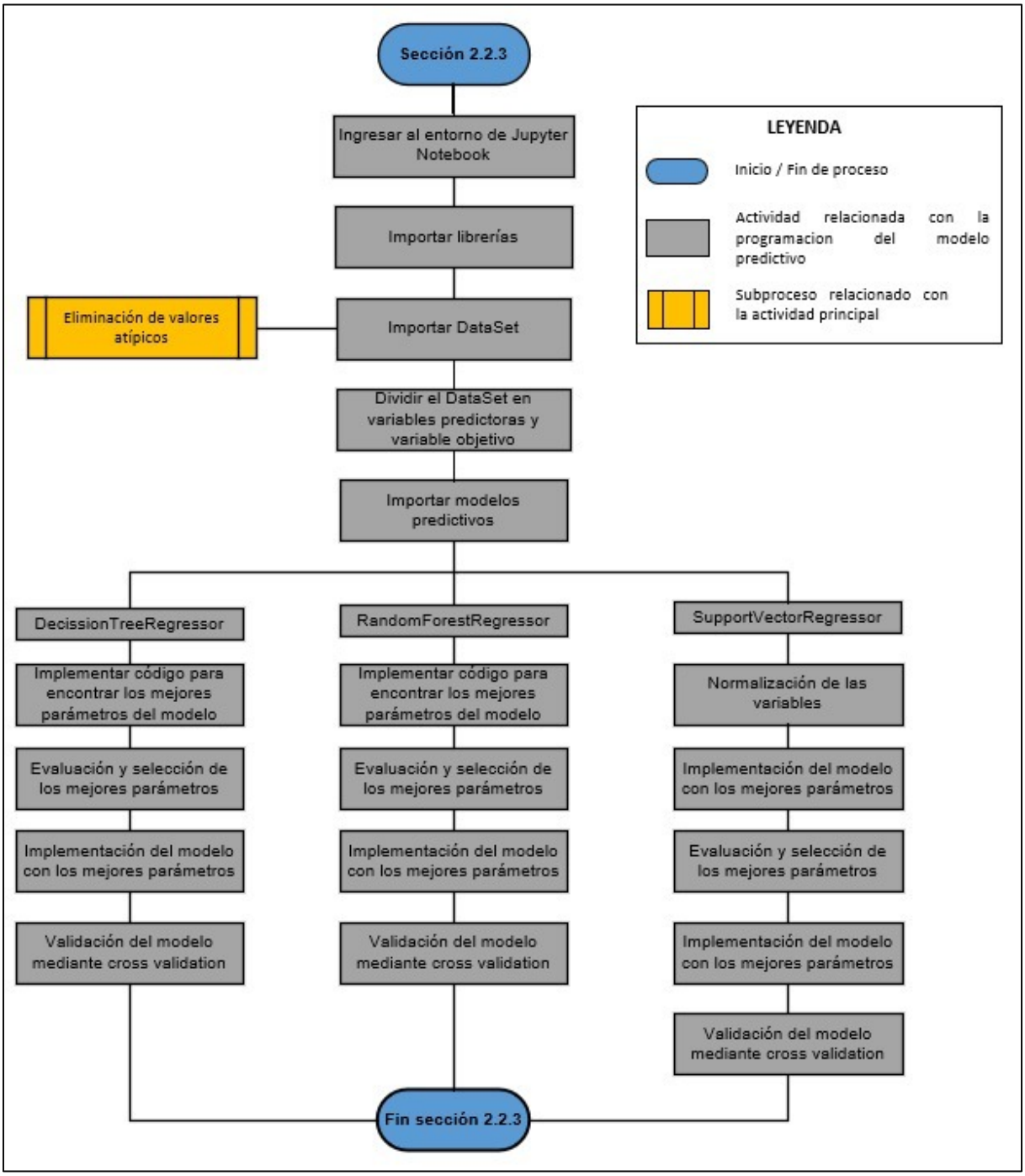
***Nota*.** En tabla se pueden apreciar las ventajas y desventajas de los modelos predictivos seleccionados.

2.2.3. *Aplicación de los modelos predictivos en la campaña de perforación 2019, evaluación y selección del modelo predictivo con mayor desempeño*

Con los tres modelos predictivos seleccionados en la sección 2.2.2, se realizaron los siguientes procedimientos como se muestra en la **Figura 17** para la creación e implementación de los modelos.

Figura 17.

Procedimiento sección 2.2.3.



Nota: La figura representa el procedimiento para llevar a cabo la sección 2.2.3.

Una vez dentro del entorno de Jupyter Notebook se procedió a importar las librerías que permitieron la implementación de los modelos predictivos. A continuación, se importó el

conjunto de datos y se eliminaron aquellos valores atípicos previamente identificados en el análisis exploratorio de la sección 2.2.1.

Posteriormente, se procedió a dividir la base de datos en 2 bloques, el primero estaría conformado por las variables predictoras (Factor de separación, DDI, VS/TVS, complejidad estructural, Buzamiento capa geológica, ángulo de ataque, gradiente de presión en el yacimiento, máxima densidad en el overburden, dureza de la roca, MD final y resultado final matriz) estas variables permitieron realizar las predicciones, mientras que el segundo bloque estaría conformado por la variable objetivo (días, costos y NPT's) que son aquellas las cuales se buscan predecir.

Continuando con el diagrama de proceso, se importaron los modelos (DTR, RFR, SVR) los cuales fueron seleccionados previamente durante la búsqueda bibliográfica.

Cabe resaltar que, para llevar a cabo el entramiento y la validación de cada modelo predictivo, se debe segmentar en dos las variables predictoras y la variable objetivo. Para llevar esto a cabo se utilizó la función Train_test_split que pertenece al método Model_Selection de la librería Sklearn, esta función permite seleccionar un porcentaje de datos para entrenar el modelo (x_train, y_train) mientras que el porcentaje restante se utilizara para probar el modelo (x_test, y_test). Asimismo, fue necesario especificar el parámetro Random_State dado que este realiza una división aleatoria de las variables x_train, y_train, x_test, y_test.

Una vez importado los modelos predictivos, se desarrolló un ciclo for que permite repetir una serie de instrucciones un numero especifico de veces. En esta parte el ciclo for realizara todas las combinaciones posibles entre los rangos definidos para los parámetros de cada modelo predictivo, a esto se le conoce como el *tuning de hiperparametros*, se debe aclarar que cada modelo predictivo maneja sus propios parámetros internos. Esto se realizó con la finalidad de encontrar los mejores parámetros, lo que a su vez permitió trabajar los modelos seleccionados con un mayor rendimiento y porcentaje de precisión.

Los parámetros modificados para los modelos predictivos y sus rangos se muestran en la **Tabla 9:**

Tabla 9.

Parámetros y rangos determinados para cada modelo predictivo.

Modelo Predictivo	Parámetro	Rangos
DecisionTreeRegressor	test_size	[0.15,0.2,0.25,0.3]
	max_depth	[1-10]
	min_sample_split	[2-10]
	min_sample_leaf	[1-10]
	random_state	[0-42]
RandomForestRegressor	test_size	[0.15,0.2,0.25,0.3]
	n_estimators	[10,50,100]
	max_depth	[1-10]
	min_sample_split	[2-10]
	min_sample_leaf	[1-10]
	random_state	[0-42]
SupportVectorRegressor	test_size	[0.15,0.2,0.25,0.3]
	Kernel	['rbf','linear','poly','sigmoid']
	degree	[1-11]
	gamma	['auto','scale']
	random_state	[0-42]

Nota. Esta tabla muestra los parámetros que fueron modificados en el Bucle For.

Inicialmente, en la delimitación de este proyecto de grado, se planeó manejar una partición del 70% para entrenar y 30% para validar (test_size) los modelos, pero durante la ejecución del mismo se decidió realizar más particiones con el propósito de conseguir modelos predictivos más óptimos y confiables.

Una vez ejecutado el *ciclo For* se obtuvieron las mejores particiones y parámetros según el modelo predictivo. Para la evaluación y selección de los mejores parámetros, se tuvieron en cuenta las métricas de regresión disponibles en la librería Sklearn como: mean absolute error, mean squared error y R^2.

La función *mean_absolute_error* calcula el error absoluto medio, esta es una métrica de riesgo correspondiente al valor esperado de la perdida de error absoluto o perdida normal [31], mientras que, la función *mean_squared_error* calcula el error cuadrático medio, esta es una métrica de riesgo correspondiente al valor esperado del error o perdida cuadrática

[31], por otra parte, la función *r2_score* calcula el coeficiente de determinación, generalmente denotado por R^2, representa la proporción de varianza (de Y) que ha sido explicada por las variables independientes en el modelo; proporciona una indicación de bondad de ajuste y, por lo tanto, una medida de qué tan bien es probable que el modelo prediga las muestras invisibles, a través de la proporción de varianza explicada. [31]

Más adelante, se creó una tabla donde quedaron consignados los resultados de las mejores particiones y los mejores parámetros junto con sus métricas de regresión para cada modelo. Seguidamente, se implementaron y entrenaron los modelos predictivos con estos resultados.

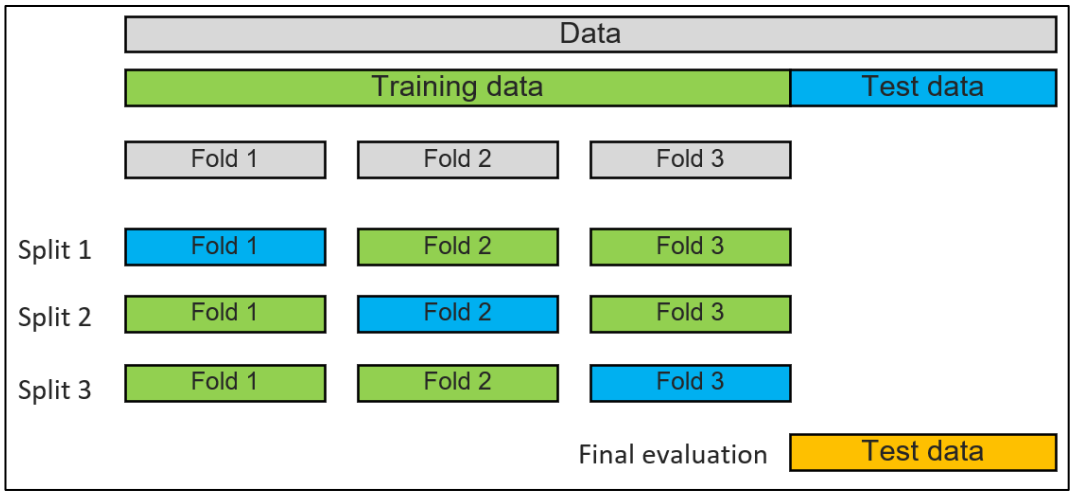
Finalmente, mediante la función Cross_Val_Score se realizó una validación cruzada con el objetivo de estimar la habilidad de generalización de los modelos predictivos frente a nuevos datos. Hasta el momento se ha trabajado dividiendo el conjunto de datos en dos secciones, la primera sección se utilizó para realizar el entrenamiento de los modelos y la segunda sección para validar los modelos. En cambio, cuando se aplica la validación cruzada se utiliza todo el conjunto de datos para entrenar y validar los modelos.

Esta función realizaría particiones según se le especifique, en este caso se realizó 3 particiones del conjunto de datos, donde cada división estuvo conformada por valores aleatorios iguales. Cabe resaltar que en cada iteración se utilizara un porcentaje para entrenar y el restante se utilizara para probar

En la **Figura 18** se muestra el proceso interno que lleva a cabo la función Cross_Val_Score.

Figura 18.

Proceso interno función cross_val_score.



Nota. Ejemplificación de las iteraciones realizadas entre el training data y test data. Tomado de: Scikit Learn. [En línea]. Disponible: https://scikit-learn.org/stable/modules/cross_validation.html [Acceso: 17 de noviembre del 2020].

2.2.4. *Selección del modelo predictivo con mayor desempeño*

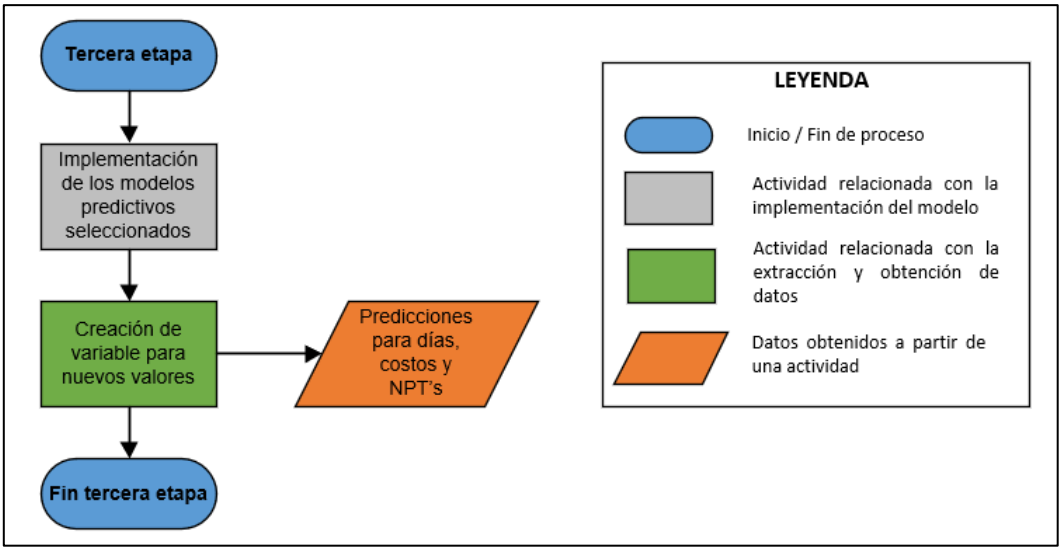
Con base a los resultados obtenidos en la sección 2.2.3. se seleccionaron los mejores modelos efectuando una comparativa entre sus métricas de regresión. Durante este proceso, el objetivo fue identificar aquel modelo con el mayor valor de R^2, menor MAE y consecutivamente el de menor MSE. Adicionalmente, se tuvo en cuenta los resultados obtenidos de la validación cruzada realizada en la anterior sección.

2.3. Tercera etapa: aplicación del modelo predictivo en la campaña de perforación del 2020.

Para dar cumplimiento al cuarto objetivo planteado, en la **Figura 19**, se muestra el procedimiento llevado a cabo para la tercera etapa:

Figura 19.

Procedimiento tercera etapa.



Nota. La figura representa los procedimientos para llevar a cabo la tercera etapa.

2.3.1. *Aplicación del modelo predictivo en la campaña de perforación 2020*

Una vez seleccionado el modelo predictivo, ya entrenado y probado con la información de la campaña del 2019, se llevó a cabo la creación de una variable “Predicciones_2020” en la cual se desarrolló un código con la finalidad de poder ingresar las variables requeridas para obtener las predicciones de días, costos y NPT’s del 2020. Lo anterior se muestra en la **Figura 20:**

Figura 20.

Ingreso de nuevo valores para predicciones de días, costos y NPT's.

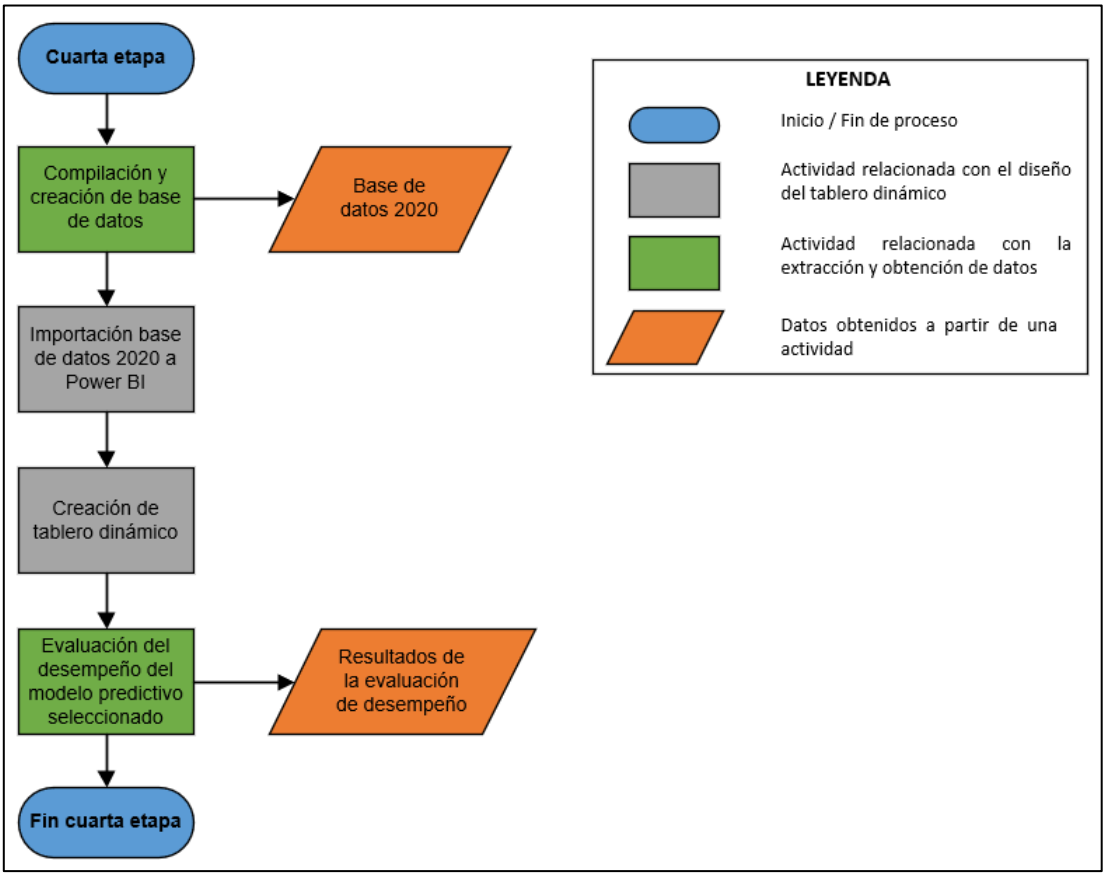
Nota. Esta figura muestra los datos de entrada que son requeridos para realizar las predicciones.

2.4. Cuarta etapa: creación de la interfaz gráfica y evaluación del desempeño del modelo predictivo

Para dar cumplimiento al último objetivo específico, la **Figura 21**, muestra el procedimiento llevado a cabo durante la cuarta etapa para la creación de un tablero dinámico en Power BI, donde se inició con la complicación de información en una base de datos 2020 final, luego se exporto estos valores al entorno de Power Bi, donde se evaluó la precisión de los modelos predictivos realizando una comparación entre lo planeado versus lo ejecutado versus lo pronosticado.

Figura 21.

Procedimiento cuarta etapa.



Nota. *La figura representa los procedimientos para llevar a cabo la cuarta etapa.*

2.4.1. *Creación base de datos final con lo planeado, ejecutado y pronosticado para los pozos perforados en el 2020*

De acuerdo con la información recopilada en la sección 2.1.4, 2.1.5, para días, costos, NPT's del 2020 y las predicciones realizadas por el modelo en la sección 2.3.1, se creó una base de datos 2020 final con la información anteriormente mencionada en el gestor de base de datos Excel para posterior importación al entorno de Power BI.

2.4.2. *Creación del tablero dinámico en Power BI y evaluación del desempeño del modelo predictivo seleccionado*

Con la base de datos 2020 final creada en la sección 2.4.1, se importó al visualizador Power BI, donde posteriormente, se diseñó un tablero dinámico enriquecido por la información de entrada, creando gráficos en los que se puedan apreciar la diferencia entre lo planeado, ejecutado y pronosticado.

3. ANÁLISIS Y RESULTADOS

En este capítulo se evidenciarán los resultados obtenidos durante la aplicación de un modelo predictivo para la predicción de días asociados a la fase de perforación, costos y NPT's. La metodología consistió en la recopilación de información de la campaña de perforación del 2019 y 2020, luego se llevó a cabo una búsqueda bibliográfica de diferentes clases de modelos predictivos de tipo supervisado, donde se escogieron y evaluaron 3 modelos con las variables seleccionadas de la matriz de complejidad relacionadas con perforación y su resultado final (Base de datos 2019), posteriormente se seleccionó el modelo predictivo con mayor desempeño entre ellos. Seguido a esto, se procedió aplicar el modelo seleccionado en la campaña de perforación 2020 con el fin de obtener las predicciones y así evaluar su rendimiento mediante un tablero dinámico en Power BI, comparando lo planeado vs lo ejecutado vs lo pronosticado. Dado lo anterior, los resultados parten desde la segunda etapa de la metodología.

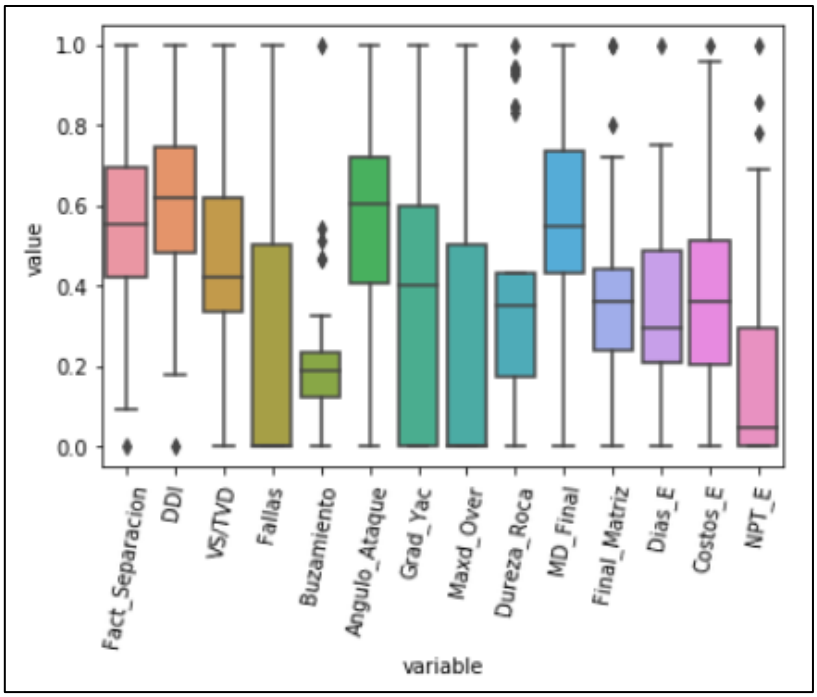
En base a lo anteriormente mencionado, se debe aclarar que la información obtenida para un modelo predictivo debe ser robusta, es decir, que cuente con una gran cantidad de datos estructurados o no estructurados, debido a que, al tener una gran cantidad de datos, conllevara a modelos más confiables y por lo tanto resultados más asertivos, siempre y cuando los datos sean verídicos y representativos.

3.1. Análisis de resultados de la aplicación de los modelos predictivos en la campaña de perforación 2019

Una vez importada la base de datos 2019 y luego de haber eliminado los valores atípicos dentro del entorno de Jupyter Notebook previamente identificados durante el análisis exploratorio, se procedió a visualizar la nueva distribución de los datos como se muestra en la **Figura 22**, obteniendo una base de datos que permite realizar predicciones más asertivas.

Figura 22.

Diagrama de caja y bigotes.



Nota. Esta figura muestra la nueva distribución de las variables, con los rangos intercuartílicos, valores mínimos y máximos.

Durante la limpieza de datos, solo se eliminaron aquellos pozos que presentaron un 50% o más de valores atípicos en sus variables, por lo tanto, en la anterior figura, se puede apreciar que en el diagrama de caja y bigotes todavía residen valores atípicos luego de haber llevado a cabo la limpieza de datos.

No se procedió a eliminar los valores atípicos que se encontraban cerca a los límites dado que estos no generarían afectaciones sobre el desempeño del modelo. adicionalmente, se debe tener en cuenta la baja cantidad de datos que se tienen para entrenar y probar los modelos predictivos.

Posteriormente al realizar la limpieza de valores atípicos de la base de datos se procedió a ejecutar el *ciclo for* con la finalidad de obtener los mejores parámetros para cada modelo predictivo seleccionado. Durante esta ejecución solo se tuvieron en cuenta los

resultados que obtuvieron un valor mayor o igual al 70% sobre el conjunto de datos de prueba.

A continuación, se mostrarán los resultados obtenidos durante esta sección:

3.1.1. *Evaluación y selección de los mejores hiperparametros*

Para la elección de los mejores parámetros se analizó la relación entre los resultados de precisión para entrenamiento (Train) y prueba (Test) de los modelos predictivos, identificando dos casos, en el primer caso se observó que al tener valores de Train superiores al 97%, los modelos estarían incurriendo en un sobreajuste, mientras que, en el segundo caso con valores de Train inferiores al 70%, se estaría incurriendo en un subajuste. Asimismo, se logró identificar las iteraciones que mostraron un excelente porcentaje de precisión sobre la variable Test, sin embargo, presentaban alguno de los dos casos mencionados anteriormente por lo tanto estas iteraciones no se tuvieron en cuenta.

Por otro parte, se realizó un análisis sobre el parámetro Max_depth para los modelos DecisionTreeRegressor y RandomForestRegressor, obteniendo 2 casos; en el primer caso cuando se obtenía un valor de profundidad inferior a 3, los modelos predictivos al no contar una gran cantidad de nodos de decisión solo tendrían pocos criterios de selección realizando finalmente predicciones constantes y obteniendo un gran porcentaje de error, mientras que en el segundo caso al asignar una profundidad superior a 8, se estaría incurriendo en un sobreajuste sobre los datos de entramiento, lo cual aumentaría el porcentaje de error de los modelos frente a nuevos datos.

A continuación, en la **Figura 23, Figura 24, Figura 25,** se muestran los resultados obtenidos luego de haber aplicado el Ciclo For donde se obtuvieron las siguientes iteraciones: 151.200, 20.160 y 14.784, entre los valores preestablecidos en la sección 2.2.3 para cada modelo respectivamente (DTR, RFR, SVR), adicionalmente se incluyeron las métricas de MAE, R^2 y MSE para cada iteración, con el propósito de obtener los mejores parámetros que permitieron realizar una predicción más asertiva sobre las variables objetivos: días, costos y NPT'S.

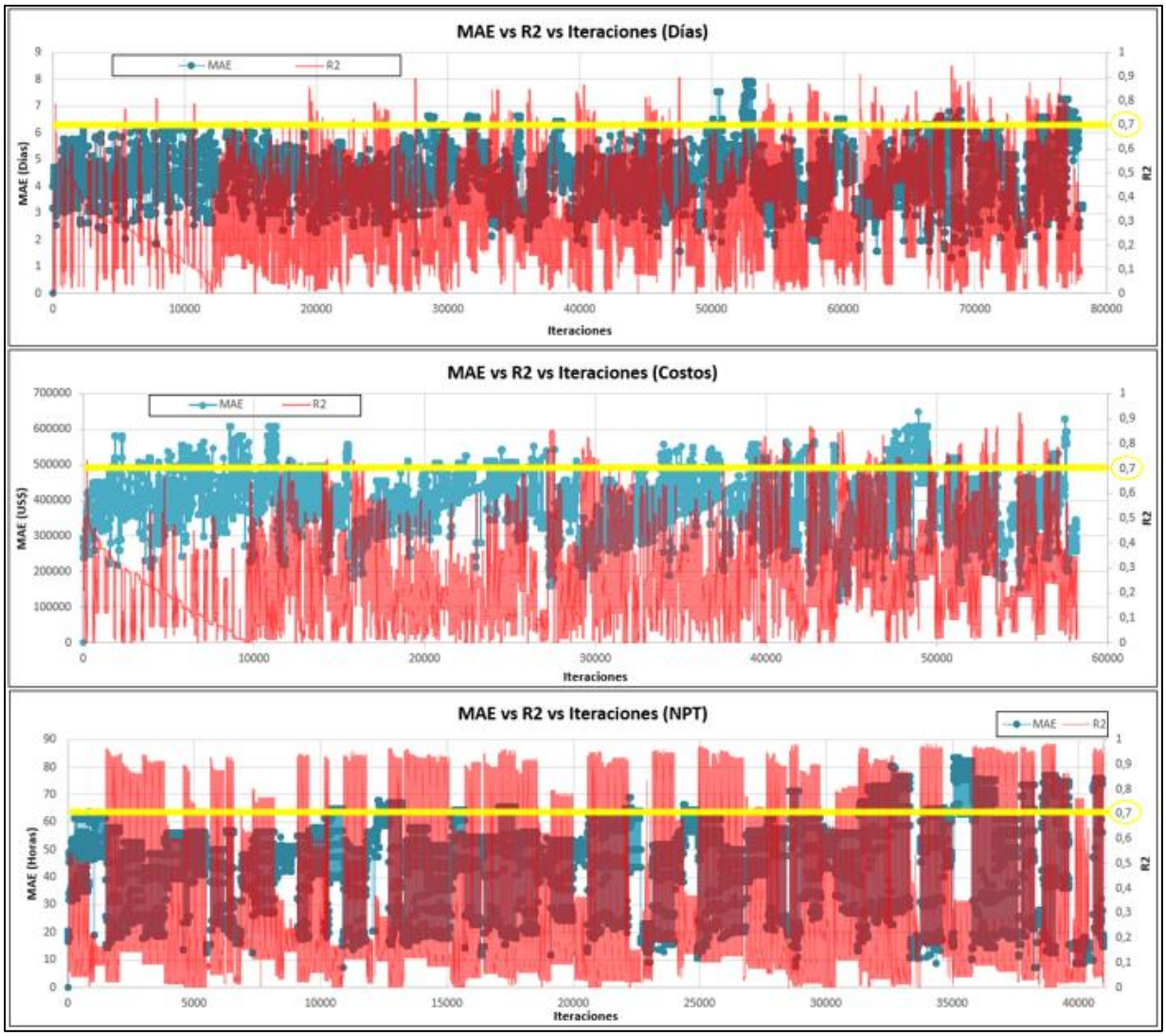
Cabe aclarar que no se tuvieron en cuenta todas las iteraciones obtenidas para cada modelo, dado que solo se seleccionaron las iteraciones de los análisis anteriormente realizados, adicionalmente no se tuvieron en cuenta las iteraciones que presentaron valores negativos de MAE. Por otra parte, se identificó que los resultados de obtenidos para MAE y MSE presentaron comportamientos similares, por lo cual, solo se utilizó la métrica MAE y R^2 para realizar las figuras comparativas junto con las iteraciones.

En las siguientes figuras**,** se observa la relación que hay entre las métricas de regresión (MAE y R^2) versus las iteraciones de los modelos predictivos para valores de R^2 superiores a cero.

Al costado izquierdo de las gráficas se encuentran los valores de MAE, en la parte inferior se encuentran la cantidad de iteraciones obtenidas durante la ejecución del Ciclo For y finalmente en el costado derecho se encuentran los valores de R^2. Para los 3 modelos predictivos se tuvieron en cuenta las iteraciones que presentaron valores de R2 superiores a 0.7 y a su vez bajos valores de MAE.

Figura 23.

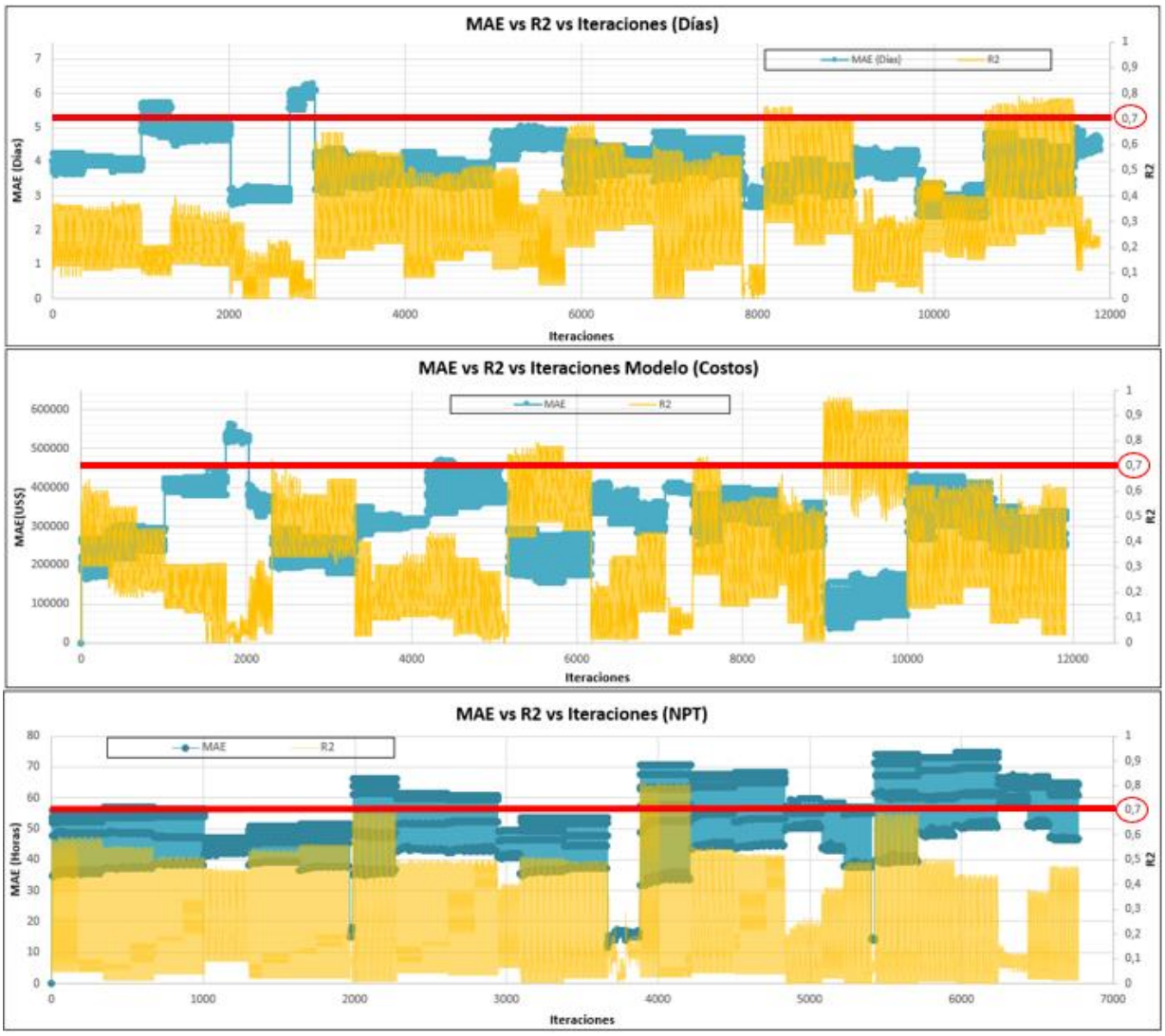
MAE vs R2 para el modelo DecisionTreeRegressor.



Nota. En la figura se puede apreciar las métricas MAE y R^2 para cada una de las iteraciones que realizo el *Ciclo For* para la predicción de días, costos y NPT's.

Figura 24.

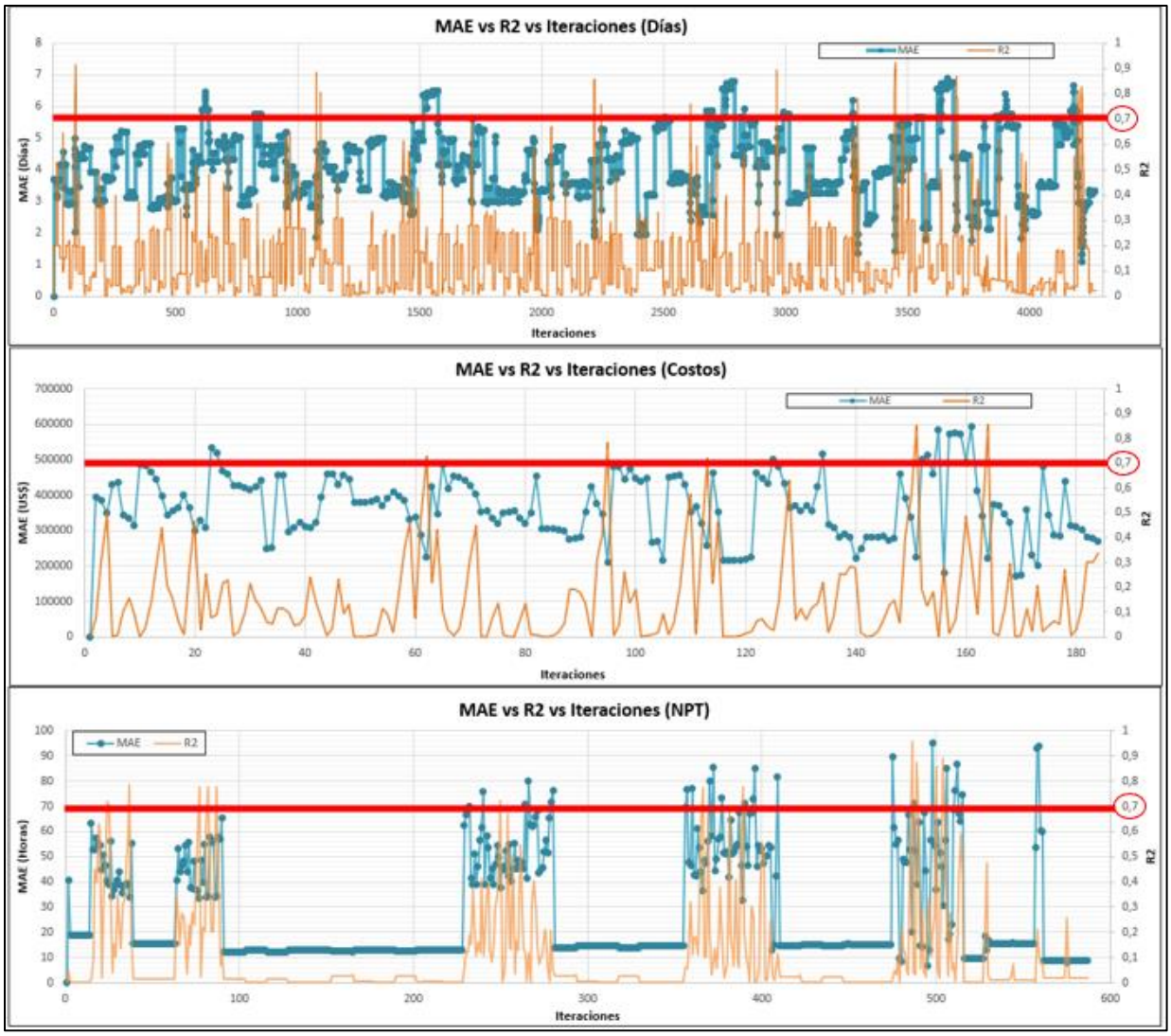
MAE vs R2 para el modelo RandomForestRegressor.



Nota. En la figura se puede apreciar los MAE y los R^2 con cada una de las iteraciones que realizo el *Bucle For* para la predicción de días, costos y NPT's.

Figura 25.

MAE vs R2 para el modelo SupportVectorRegressor.



Nota. En la figura se puede apreciar los MAE y los R^2 con cada una de las iteraciones que realizo el *Bucle For* para la predicción de días, costos y NPT's

En las anteriores graficas se puede observar la relación inversamente proporcional que existe entre las métricas MAE y R^2, dado que a un mayor valor de R^2 tenemos un menor valor de MAE, lo que permitió identificar las mejores iteraciones con sus respectivos parámetros. Estas métricas fueron un punto clave para la selección de los mejores modelos predictivos, aunque se le dio más relevancia a la métrica R^2 dado que esta es una medida de la relación lineal entre variables.

En la **Tabla 10, Tabla 11, Tabla 12,** se puede apreciar los mejores parámetros para cada una de las diferentes particiones que se llevaron a cabo previamente establecidas para cada modelo en la sección 2.2.3

Tabla 10.

Mejores parámetros modelo DecisionTreeRegressor para costos, días y NPT's.

Variable Objetivo	Train-Test Size	Train	Test	Max_depth	min samples split	min samples leaf	Random state	MSE	MAE	R2
Costos	70%-30%	0,899	0,649	6	2	2	29	$ 180.847.998.910	$ 372.610	0,648
	75%-25%	0,883	0,677	7	2	2	27	$ 114.751.386.483	$ 291.079	0,676
	80%-20%	0,922	0,809	5	2	2	26	$ 73.899.940.219	$ 242.642	0,809
	85%-15%	**0,926**	**0,842**	5	5	1	26	$ 67.501.353.696	**$ 205.973**	**0,841**
Días	70%-30%	0,929	0,790	5	5	1	18	10,594 Días	2,545 Días	0,789
	75%-25%	0,905	0,858	6	5	2	27	6,647 Días	2,402 Días	0,857
	80%-20%	0,815	0,834	5	9	1	36	8,119 Días	2,390 Días	0,834
	85%-15%	**0,911**	**0,909**	5	2	2	13	4,702 Días	**1,795 Días**	**0,909**
NPT's	70%-30%	0,930	0,878	5	8	1	22	1068,999 Horas	23,222 Horas	0,878
	75%-25%	0,927	0,893	5	7	1	22	1099,680 Horas	23,567 Horas	0,893
	80%-20%	**0,922**	**0,962**	5	4	1	36	456,925 Horas	**19,658 Horas**	**0,962**
	85%-15%	0,918	0,952	5	3	1	18	734,710 Horas	22,681 Horas	0,952

Nota: La tabla muestra los mejores parámetros para las diferentes particiones.

Tabla 11.

Mejores parámetros modelo RandomForestRegressor para costos, días y NPT's.

Variable Objetivo	Train-Test Size	Train	Test	n_estimators	Max depth	min samples split	min samples leaf	Random state	MSE	MAE	R2
Costos	70%-30%	0,744	0,648	10	5	2	1	0	$ 64.421.371.747	$ 176.202	0,648
	75%-25%	0,818	0,639	10	5	2	1	0	$ 70.265.035.355	$ 216.362	0,639
	80%-20%	0,761	0,662	10	5	2	1	0	$ 81.852.130.585	$ 206.431	0,662
	85%-15%	**0,801**	**0,972**	10	6	6	2	0	$ 2.748.196.581	**$ 40.640**	**0,972**
Dias	70%-30%	0,912	0,381	100	6	2	1	0	18,826 Días	3,764 Días	0,381
	75%-25%	0,829	0,647	10	7	2	2	0	12,258 Días	3,101 Días	0,648
	80%-20%	0,822	0,744	10	7	2	2	3	10,704 Días	2,861 Días	0,744
	85%-15%	**0,814**	**0,759**	50	5	3	2	3	12,851 Días	**3,008 Días**	**0,759**
NPT's	70%-30%	0,830	0,787	10	7	4	1	1	1500,947 Horas	26,011 Horas	0,787
	75%-25%	0,925	0,866	10	5	3	1	3	1103,849 Horas	23,719 Horas	0,866
	80%-20%	**0,931**	**0,807**	10	5	6	1	17	2458,339 Horas	**32,63 Horas**	**0,807**
	85%-15%	0,756	0,756	10	6	4	1	3	3233,085 Horas	31,026 Horas	0,757

Nota. La tabla muestra los mejores parámetros para las diferentes particiones.

En la anterior tabla se identificó que para el parámetro n_estimators, la cantidad de árboles necesarios para construir el modelo RFR es de 10 basado en el rango

preestablecido en la sección 2.2.3, adicionalmente, se logró determinar que así se aumente el número de estimadores solo se estará afectando el tiempo de corrida del modelo predictivo.

Tabla 12.

Mejores parámetros modelo SupportVectorRegressor para costos, días y NPT's.

Variable Objetivo	Train-Test Size	Train	Test	Gamma	Kernel	Degree	Random state	MSE	MAE	R2
Costos	70%-30%	0,708	0,465	scale	poly	10	15	$ 171.659.931.661	$ 298.382	0,465
	75%-25%	0,709	0,465	scale	poly	10	11	$ 183.499.626.954	$ 333.073	0,465
	80%-20%	0,741	0,786	scale	poly	11	3	$ 78.304.874.518	$ 210.405	0,786
	85%-15%	**0,723**	**0,853**	scale	poly	11	3	$ 66.970.910.672	**$ 223.938**	**0,853**
Dias	70%-30%	0,876	0,918	scale	poly	4	3	5,008 Días	2,027 Días	0,918
	75%-25%	0,865	0,808	scale	poly	3	5	7,448 Días	2,358 Días	0,808
	80%-20%	0,829	0,848	scale	poly	3	3	6,335 Días	2,108 Días	0,848
	85%-15%	**0,847**	**0,874**	scale	poly	3	13	6,549 Días	**2,103 Días**	**0,874**
NPT's	70%-30%	0,986	0,788	scale	poly	8	13	1474,019 Horas	33,775 Horas	0,788
	75%-25%	0,992	0,724	scale	poly	9	11	2317,840 Horas	37,575 Horas	0,724
	80%-20%	**0,963**	**0,778**	scale	poly	7	22	2729,419 Horas	**32,827 Horas**	**0,778**
	85%-15%	0,994	0,958	scale	poly	9	3	561,932 Horas	20,320 Horas	0,958

Nota. La tabla muestra los mejores parámetros para las diferentes particiones del train_test_split.

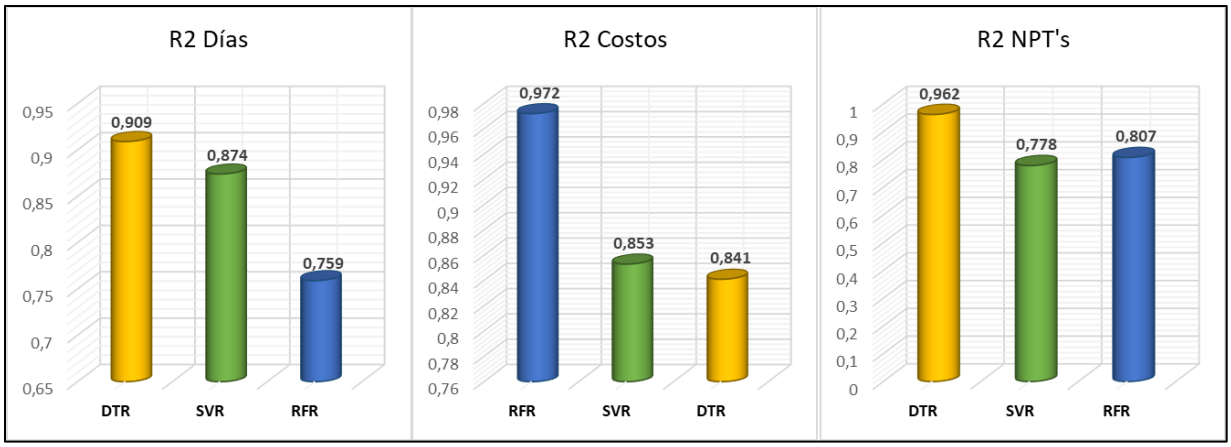
En las anteriores tablas, se estableció que la partición más adecuada para los modelos de días y costos obtuvieron un valor de train_size del 85% para entrenar y un valor test_size del 15% para probar, mientras que para el modelo de NPT’s la mejor partición obtuvo un valor train_size del 80% para entrenar mientras que se obtuvo un valor test_size del 20% para probar. Se debe aclarar que se escogieron las anteriores particiones partiendo de la relación de mayor R^2 y menor MAE.

Los anteriores resultados se deben a la baja cantidad de datos recopilados en la sección 2.1., por lo tanto, para los modelos predictivos fue primordial utilizar un mayor conjunto de datos para entrenar en vez de utilizarlos para validar.

A continuación, se muestra a manera de resumen los R^2 para cada modelo predictivo obtenidos para las variables objetivos: días, costos y NPT’s.

Figura 26.

Valores de R^2 más altos de cada modelo predictivo.



Nota. La figura muestra los valores de R^2 de los mejores parámetros de cada modelo respectivamente.

En este diagrama de barras, se puede apreciar los resultados obtenidos de R^2; para la variable días el modelo con mayor desempeño fue el DecisionTreeRegressor con un 90.9% de precisión mientras que el modelo RandomForestRegressor presento la mejor precisión para la variable costos con un 97.2%. Por otra parte, para la variable NPT's el modelo con mejor desempeño fue el DecisionTreeRegressor con un 96,2% de precisión. Cabe aclarar que estos resultados fueron obtenidos sobre la base de datos 2019.

3.1.2. *Validación cruzada*

Luego de haber realizado la implementación y entrenamiento de los modelos predictivos con las mejores particiones y parámetros, se llevó a cabo la validación cruzada con la finalidad de estimar el compartimiento de los modelos predictivos mediante el promedio de los valores calculados en cada partición. Durante la ejecución de la función cross_val_score de la librería de Sklearn, solo se llevaron a cabo 3 particiones aleatorias debido a la limitada cantidad de información que se cuenta para la base de datos 2019.

Los resultados obtenidos para cada modelo se muestran en la **Tabla 13**.

Tabla 13.

Validación cruzada con los mejores parámetros de los modelos predictivos.

Variable Objetivo	Modelo	1ra Partición	2da Partición	3ra Partición	Presición
Costos	DecissionTreeRegressor	-0,34	0,187	-0,688	-0,28
	RandomForestRegressor	**-0,105**	**0,295**	**0,526**	**0,24**
	SupportVectorRegressor	-6,717	0,497	0,473	-1,92
Días	DecissionTreeRegressor	-0,242	-0,18	0,325	-0,03
	RandomForestRegressor	**0,108**	**0,177**	**0,365**	**0,22**
	SupportVectorRegressor	-1,11	-0,216	0,623	-0,23
NPT's	**DecissionTreeRegressor**	**-2,288**	**0,182**	**0,904**	**-0,4**
	RandomForestRegressor	-3,327	0,429	0,558	-2,11
	SupportVectorRegressor	-109.88	-0.656	-0.991	-37.18

Nota. En esta tabla se observa los resultados obtenidos para la validación cruzada de cada modelo.

Con base a los resultados de precisión mostrados en la anterior tabla, se evidencio que el modelo RandomForestRegressor obtuvo la precisión más alta con respecto a la predicción de las variables objetivo días y costos con un valor del 24% y 22% respectivamente, por otra parte, el modelo que tuvo una mejor puntuación sobre los datos para la predicción de NPT's fue el DecisionTreeRegressor con un -40%. Adicionalmente, se puede apreciar que todos modelos implementados no superan el de 70% de precisión, por lo tanto, estos modelos no son generalizables frente a nuevos datos y se requiere de mayor información para la obtención de predicciones más asertivas.

3.2. Selección de los modelos predictivos

En esta sección se identificaron los modelos predictivos de mayor eficiencia teniendo en cuenta las métricas de regresión tales como: coeficiente de determinación (R^2), error medio absoluto (MAE) y error cuadrático medio (MSE). Asimismo, se consideró el puntaje de la validación cruzada, dado que este permitió identificar qué modelo se encontraba realizando las mejores predicciones frente a nuevos datos.

Con base a los resultados obtenidos durante la ejecución de sección 3.2, se eligieron los siguientes modelos para la predicción de días, costos y NPT's respectivamente como se muestra en la **Tabla 14**:

Tabla 14.

Modelos predictivos seleccionados según variable objetivo

Variable objetivo	Modelo seleccionado
Días	DecisionTreeReegressor
Costos	RandomForestRegressor
NPT's	DecisionTreeReegressor

Nota. La tabla muestra los modelos con mayor desempeño para la predicción de costos, días y NPT´s respectivamente.

A continuación, se especificarán los modelos elegidos con mayor detalle para cada variable objetivo:

3.2.1. *Selección del mejor modelo para días*

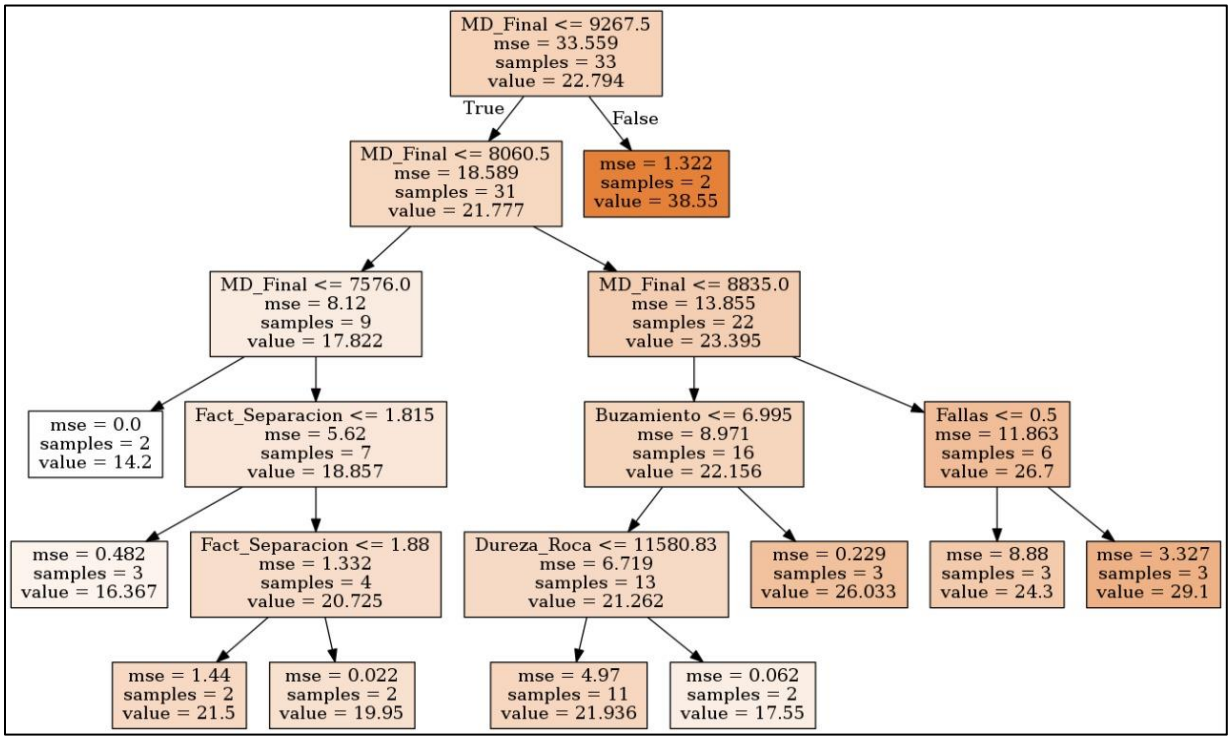
En base a los resultados reflejados en la **Tabla 11**, **Tabla 12**, se evidencio que para la predicción de días los modelos RandomForestRegressor *y* SupportVectorRegressor, obtuvieron valores de MAE altos y bajos valores en R^2 entre los modelos evaluados, por lo cual, estos no se tuvieron en cuenta para la predicción de días.

Por otro lado, el modelo *DecisionTreeRegressor* obtuvo el mayor porcentaje de R^2 con un valor del 90% y el menor MAE que corresponde a 1,79 días evidenciado en la **Tabla 10**. Adicionalmente, se comprobó que el modelo DTR no tuvo un gran puntaje en la validación cruzada por lo cual este podría no desempeñarse apropiadamente frente a nuevos datos, sin embargo, el DTR fue el modelo seleccionado para la predicción de la variable objetivo días.

Una vez con el modelo seleccionado para la predicción de la variable días, se visualizó el árbol y como este realiza la toma decisiones durante su ejecución, luego se mostraron las variables que tuvieron un peso significativo para realizar las predicciones y su influencia sobre la variable objetivo. Lo anterior se muestra en las siguientes figuras:

Figura 27.

Representación del árbol de decisión.



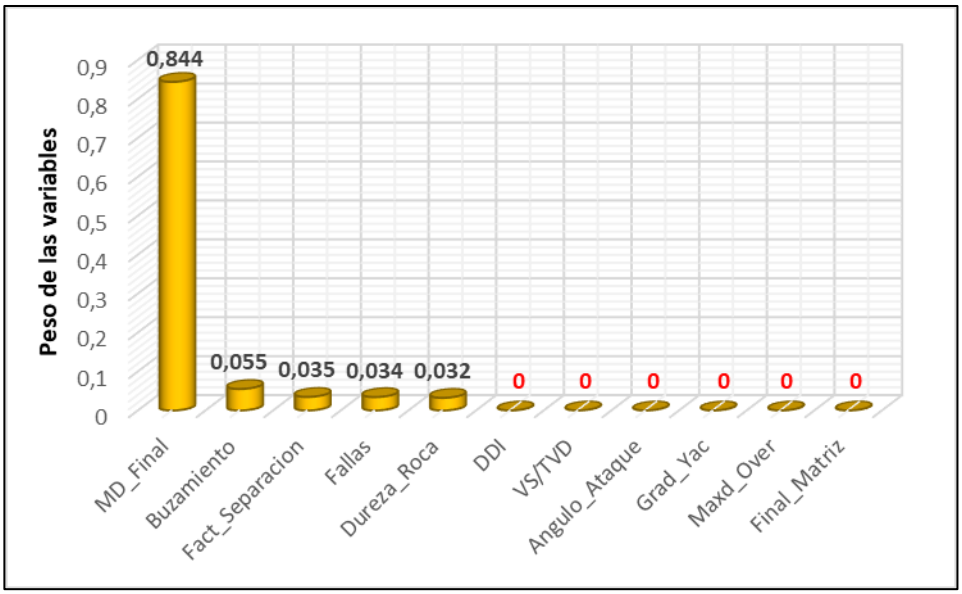
Nota. La figura muestra las bifurcaciones que se realiza durante la ejecución del modelo DTR.

En la anterior figura, se evidencio que el DecisionTreeRegressor tuvo en cuenta características como fallas, factor de separación, buzamiento entre otras para llevar a cabo sus predicciones.

Con base en lo anterior, se muestra en la **Figura 28** la importancia de las variables predictoras y sus pesos durante la ejecución del modelo predictivo.

Figura 28.

Variables predictoras con sus respectivos pesos.



Nota. Esta figura resalta aquellas variables predictoras que tuvieron influencia durante la ejecución del modelo DTR.

En la anterior figura se observa que las variables más importantes para el modelo predictivo días fueron MD_Final, Buzamiento y Fact_Separacion con un valor de 0.84, 0.05 y 0.03 respectivamente. Estas variables explican el 92% del modelo, en cambio las variables DDI, VS/TVD, Angulo_Ataque, Grac_Yac, Maxd_Over y Final_Matriz, no aportaron para la toma de decisiones en la ejecución del modelo para la predicción de días.

3.2.2. *Selección del modelo para la predicción de costos*

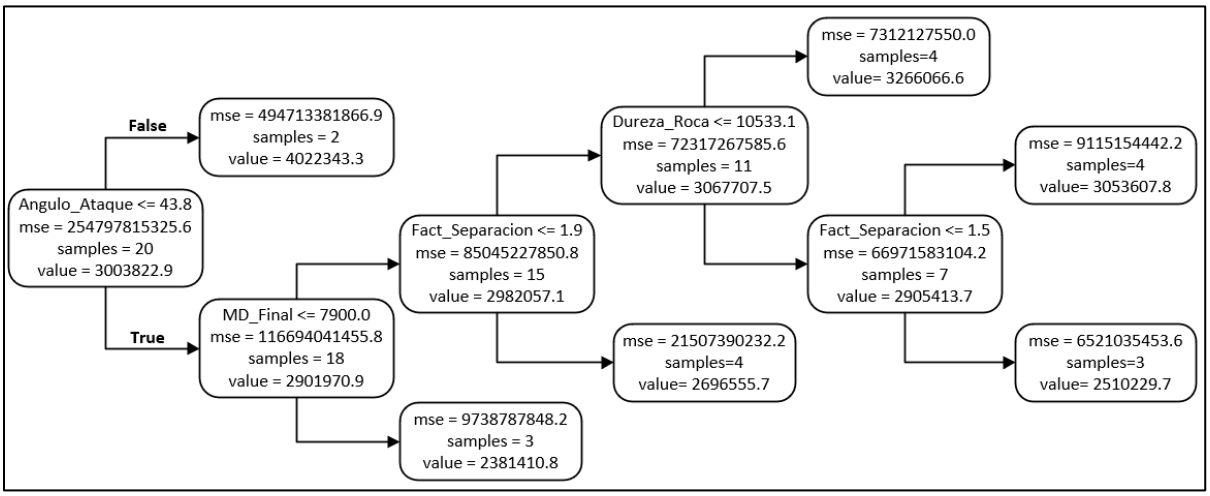
En base a los resultados reflejados en la **Tabla 10**, **Tabla 12**, se evidenció que para la predicción de costos los modelos *DecisionTreeRegressor* y *SupportVectorRegressor* obtuvieron los valores más altos de MAE entre los tres modelos evaluados, además de obtener porcentajes de validación cruzada muy bajos, por lo cual, estos modelos no se tuvieron en cuenta para la predicción de costos.

Por otra parte, el modelo RandomForestRegressor obtuvo el mayor porcentaje de R^2 con un 97,2% y el menor MAE con $40.640 evidenciado en la **Tabla 11**, de la misma manera obtuvo el mayor puntaje en la validación cruzada con un 24%. Por lo anterior, el modelo seleccionado para la predicción de la variable objetivo costos fue RandomForestRegressor.

Con el modelo seleccionado para la predicción de costos, se procedió a visualizar a manera de ejemplo 1 de los 10 que toma RFR. Asimismo, se plasmaron las variables que tuvieron relevancia al momento de realizar la predicción y ver su influencia sobre la variable objetivo, lo anterior se muestra en las siguientes figuras:

Figura 29.

Muestra representativa de uno de los árboles de decisión del bosque aleatorio.

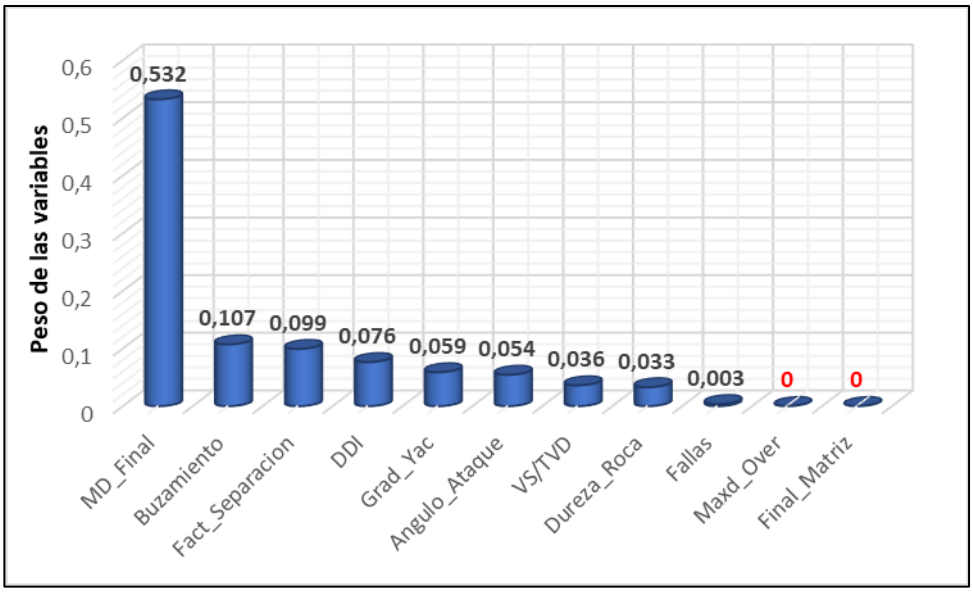


Nota. Esta figura muestra las bifurcaciones de uno de los 10 árboles con los que se ejecuta el modelo RFR.

En la anterior figura, se evidencio que el modelo RandomForestRegressor tuvo en cuenta características como: ángulo de ataque, factor de separación, profundidad medida entre otras para llevar a cabo sus predicciones.

Figura 30.

Variables predictoras con sus respectivos pesos.



Nota. Esta figura resalta aquellas variables predictoras que tuvieron relevancia durante la ejecución del modelo RFR.

En la anterior figura se observa que las variables más importantes para el modelo predictivo de costos fueron MD_Final, Buzamiento y Fact_Separacion con un valor de 0.53, 0.10 y 0.09 respectivamente, estas variables explican el 72% del modelo predictivo. Adicionalmente, se puede evidenciar que al tener en cuenta las predicciones de 10 árboles diferentes se obtuvo una mayor cantidad de variables para realizar la predicción de costos, sin embargo, las variables Maxd_Over y Final_Matriz, no aportaron para la toma de decisiones en la ejecución del modelo para la predicción de costos.

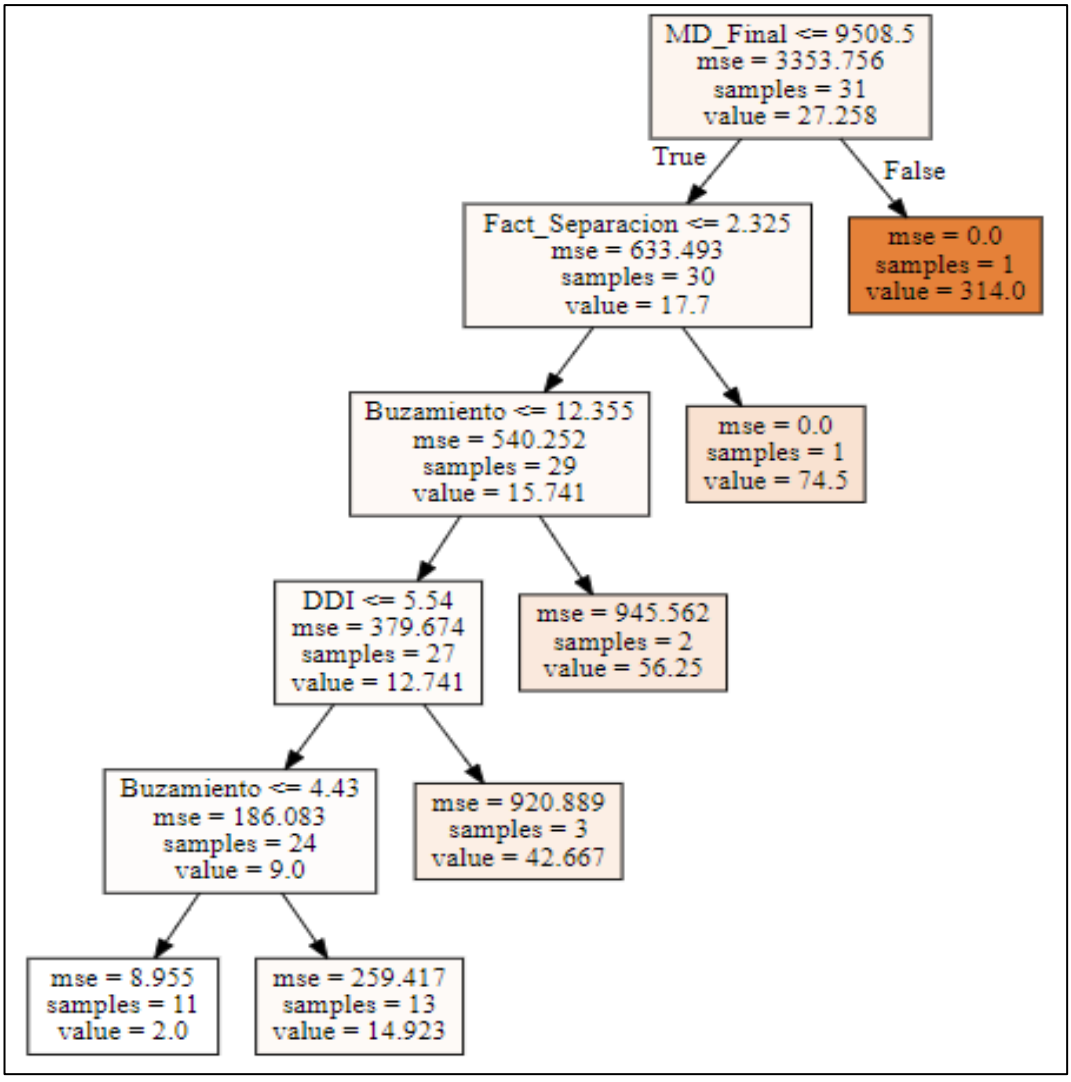
3.2.3. *Selección del mejor modelo para NPT's*

En base a los resultados reflejados en la **Tabla 11** y **Tabla 12**, para la predicción de NPT's se evidencio que los modelos RFR y SVR, obtuvieron valores altos de MAE y asimismo valores bajos en R^2 entre los tres modelos evaluados. Por lo cual, estos no se tuvieron en cuenta para la predicción. Por el contrario, el modelo DTR obtuvo el mayor porcentaje de R2 con un 96,2% y el menor MAE de 19,65 horas, lo cual lo hace ideal para la predicción de la variable objetivo NPT's.

Una vez con el modelo seleccionado para la predicción de NPT's, se visualizó el árbol y como esta toma decisiones durante su ejecución, luego se mostraron las variables que tuvieron peso para realizar las predicciones y su influencia sobre la variable objetivo. Lo anterior se muestra en las siguientes figuras:

Figura 31.

Representación del árbol de decisión.

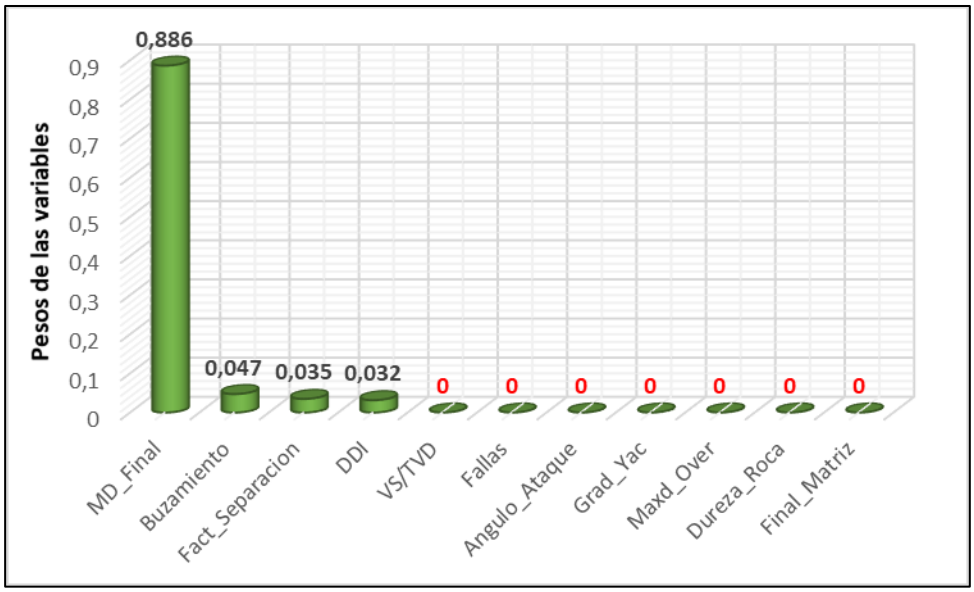


***Nota*.** Esta figura muestra las bifurcaciones que se realiza durante la ejecución del modelo DTR.

En base a lo anterior, se muestra en la **Figura 32** las variables predictoras y sus pesos durante la ejecución del modelo predictivo**:**

Figura 32.

Variables más relevantes para el modelo de NPT´s.



Nota. En esta figura se puede apreciar las variables que tuvieron mayor influencia sobre el modelo DTR.

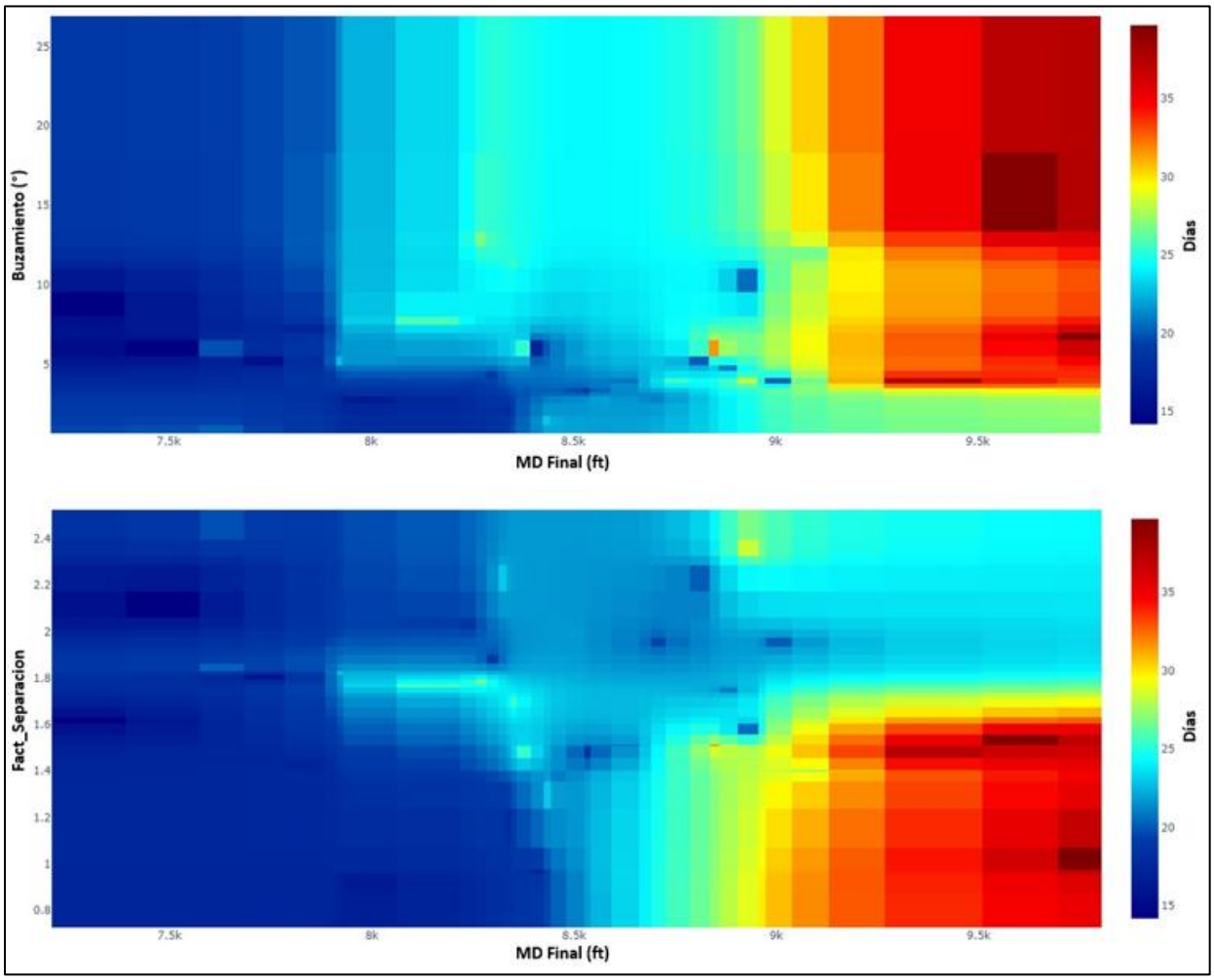
En la anterior figura se observa que las variables más importantes para el modelo predictivo de NPT's fueron MD_Final, Buzamiento y Fact_Separacion con un valor de 0.88, 0.04 y 0.035 respectivamente, estas variables explican el 95% del modelo.

3.2.4. *Análisis variables predictoras de los modelos seleccionados*

Finalmente, en base a los resultados obtenidos durante las secciones 3.2.1, 3.2.2 y 3.2.3, se identifica que para los 3 modelos predictivos seleccionados se tienen inicialmente las mismas 3 variables predictoras que son MD_Final, Buzamiento y Fact_Separacion. Por lo tanto, se procedió a representar mediante un mapa de calor las variables más representativas en los ejes X y Final Matriz donde se muestra la influencia de estas sobre las variables objetivos días, costos y NPT's en el eje z.

Figura 33.

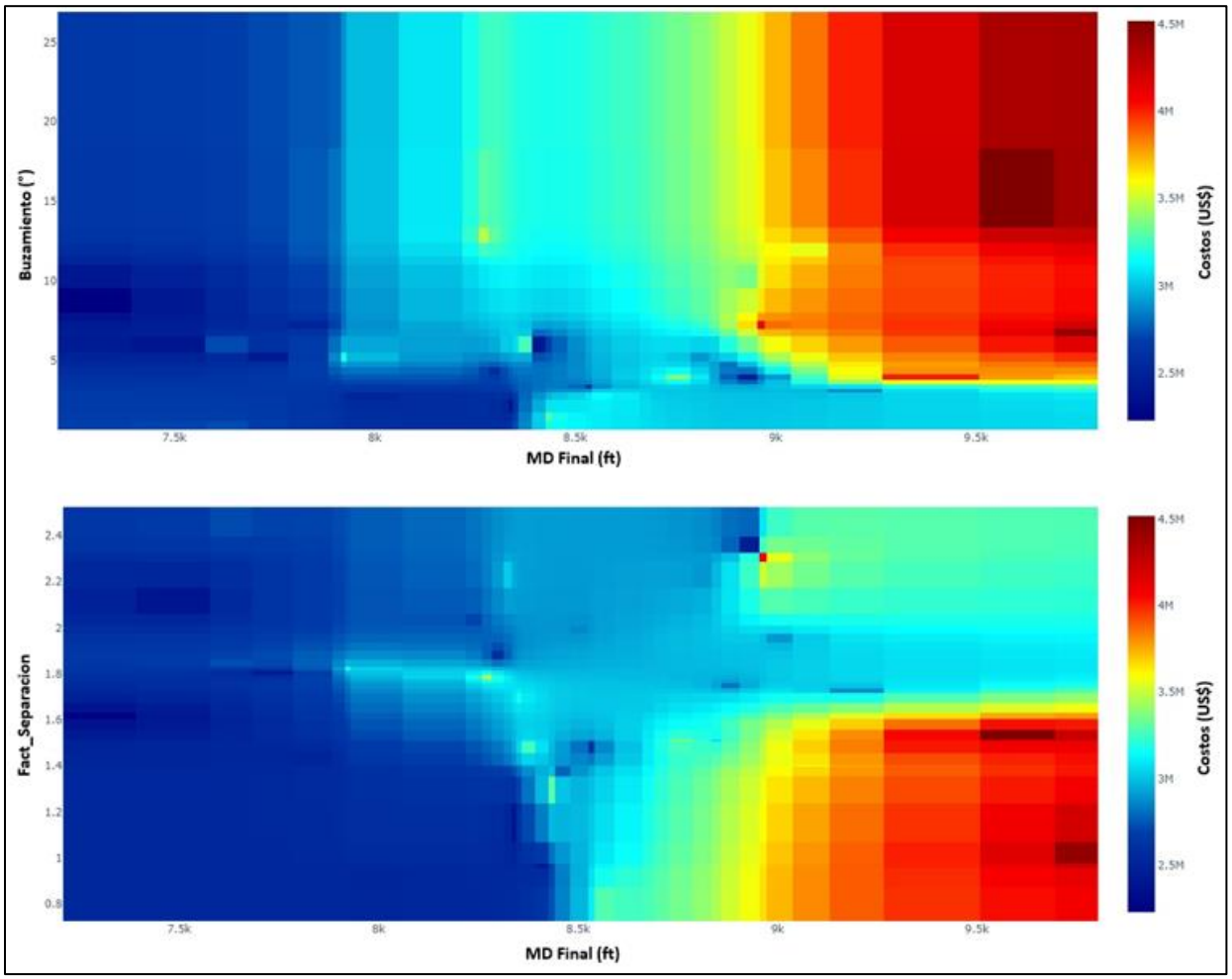
Mapa de calor con las variables más relevantes del modelo DTR para días.



Nota. Esta figura muestra la relación entre las variables MD_Final, Buzamiento y Fact_Separacion.

Figura 34.

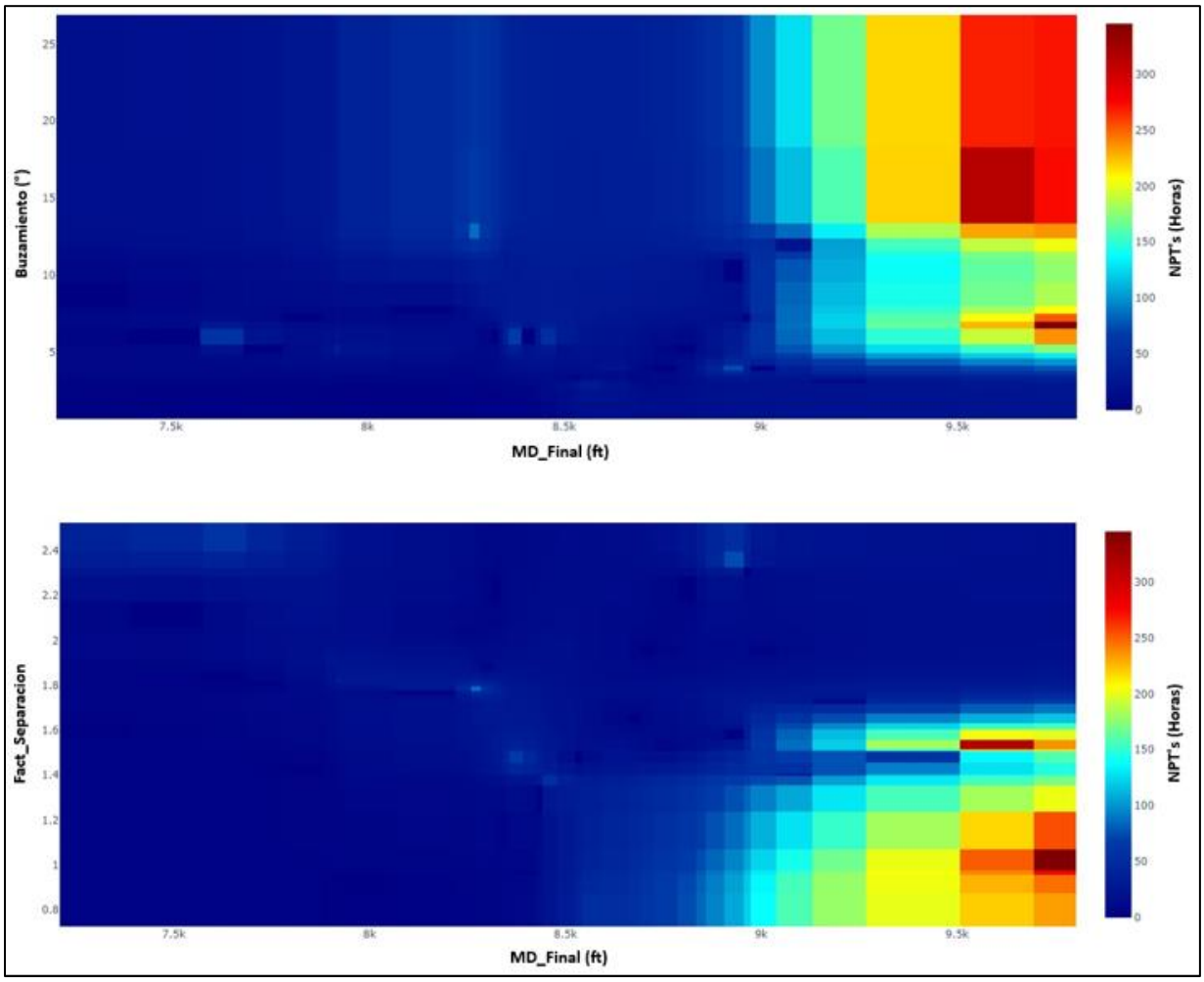
Mapa de calor correlacionando las variables más relevantes del modelo RFR para costos.



Nota. Esta figura muestra la relación existente entre las variables MD_Final, Buzamiento y Fact_Separacion.

Figura 35.

Mapa de calor correlacionando las variables más relevantes del modelo DTR para NPT's.



Nota. Esta figura muestra la relación existente entre las variables MD_Final, Buzamiento y Fact_Separacion.

En los anteriores mapas de calor (**Figura 33**, **Figura 34** y **Figura 35**), se puede observar una relación directamente proporcional entre las variables Buzamiento y MD_Final. Además, se puede apreciar que a mayor profundidad medida y mayor grado de buzamiento se presentan los valores más altos de días, costos y NPT's dado que a mayor grado de buzamiento las capas estarán más inclinadas y la trayectoria del pozo será más horizontal. Lo anterior puede incurrir en problemas geomecánicos como inestabilidad en

el hueco o restricciones, aumentando los valores de torque y arrastre. Por lo tanto, se requiere mayor cantidad de tiempo para perforar este tipo de pozos.

Asimismo, se puede apreciar que a valores menores de 1,6 en factor de separación se presentan la mayor cantidad de días perforados y a su vez el mayor valor de costos, debido a que la trayectoria direccional que se tiene que llevar a cabo es más compleja para evitar la colisión entre pozos. Por lo cual, se debe realizar un seguimiento estricto al plan direccional y se debe tomar surveys con una frecuencia de cada 30 ft cuando lo normal es tomar surveys cada 90 ft, adicionalmente, se presentan eventos de interferencia magnética generalmente en la primera sección del pozo lo cual ralentiza el trabajo direccional.

Por otra parte, se determinó que no es posible utilizar un solo modelo para la predicción de días, costos y NPT's, dado que para cada variable objetivo existe una configuración especifica en las particiones e hiperparametros.

3.3. Resultados aplicación de los modelos predictivos en la campaña de perforación del 2020.

Para llevar a cabo la implementación de los modelos seleccionados sobre los valores de la campaña 2020, se desarrolló un código al cual se le asignó una variable llamada predicciones_2020 para ingresar nuevos valores en cada modelo y su posterior ejecución. Después, se obtuvieron las predicciones de días, costos y NPT's, los cuales fueron almacenados posteriormente en la **Tabla 15**.

Tabla 15.

Predicciones para tiempos, costos y NPT's 2020.

Pozos 2020	Costos (US$)	Días	NPT's (Horas)
Castilla T1	$ 2.573.024	21,936	14,92
Castilla Norte T1	$ 2.844.474	21,936	74,5
Castilla Norte T2	$ 3.098.044	21,936	14,92
Castilla Norte T3	$ 3.016.109	26,033	14,92
Castilla Norte T4	$ 2.597.816	16,366	14,92

Nota. Esta tabla muestra los resultados obtenidos los modelos predictivos seleccionados.

3.4. Análisis y resultados creación de la interfaz gráfica y evaluación del desempeño del modelo predictivo

Con los resultados obtenidos por los modelos predictivos en la sección 3.3 y los datos recopilados para tiempos, costos y NPT's en la sección 2.1.5., se procedió a crear la base de datos que almacenó la información anteriormente mencionada como se muestra en la **Tabla 16:**

Tabla 16.

Base de datos 2020 final.

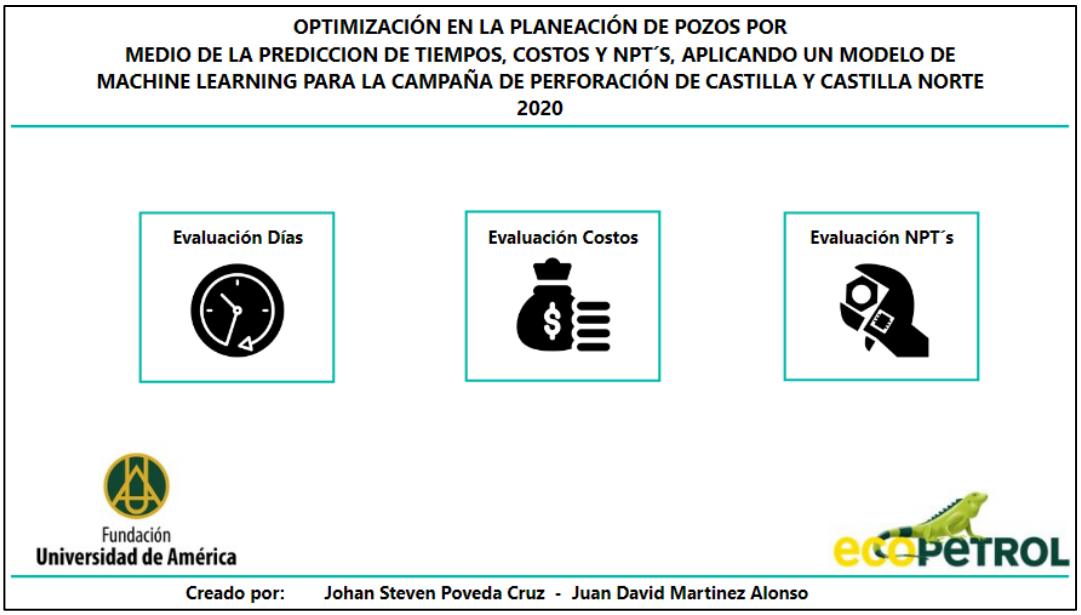
Variables	Unidades
Well name	Adimensional
Días Planeados	Días
Días Ejecutados	Días
Días Pronosticados	Días
Costos Planeados	US $
Costos Ejecutados	US $
Costos Pronosticados	US $
NPT's Ejecutados	Horas
NPT's Pronosticados	Horas

Nota. Esta tabla muestra las variables para posterior evaluación del modelo predictivo seleccionado.

Finalmente, creada la base de datos 2020 final e importada al entorno de Power BI, se llevó a cabo la creación del tablero dinámico el cual, en su página de inicio se encuentra dividida en tres secciones (días, costos y NPT's), lo anterior se muestra en la siguiente figura:

Figura 36.

Página de inicio del tablero dinámico.



Nota. La figura representa las secciones en las que se encuentra dividida la página de inicio del tablero dinámico.

A continuación, se realizó la comparación entre los días planeados vs ejecutados vs pronosticados, con el propósito de evaluar el desempeño del modelo DTR frente a los 5 pozos de la campaña 2020 para el campo Castilla y Castilla Norte.

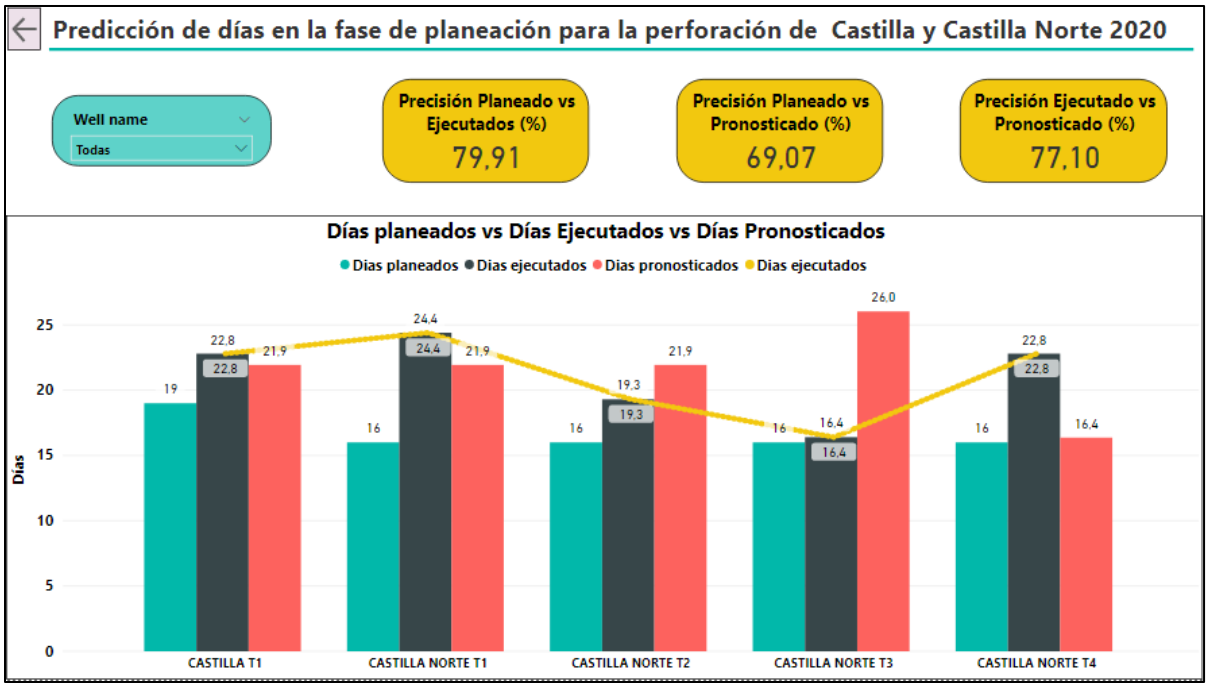
3.4.1. *Análisis y resultados modelo DTR sobre la campaña 2020 para días.*

En la **Figura 37,** se muestra que durante la fase de planeación para perforación se tuvo una precisión promedio del 79,91% de los días planeados con respecto a los ejecutados, esto es lo que actualmente ejecuta la operadora.

En lo anterior se puede evidenciar que la precisión de la planeación actual frente a la ejecución tiene una incertidumbre del 20,09%, esto se debe a un problema que está teniendo la compañía al planear los pozos con un valor general de 16 días mientras que se están ejecutando en un promedio de 21,14 días. Por lo tanto, el presupuesto asignado a una campaña de perforación se puede agotar antes de lo previsto, lo que a su vez genera cambios de alcance y contratiempos tanto logísticos como económicos.

Figura 37.

Tablero dinámico para la variable días.

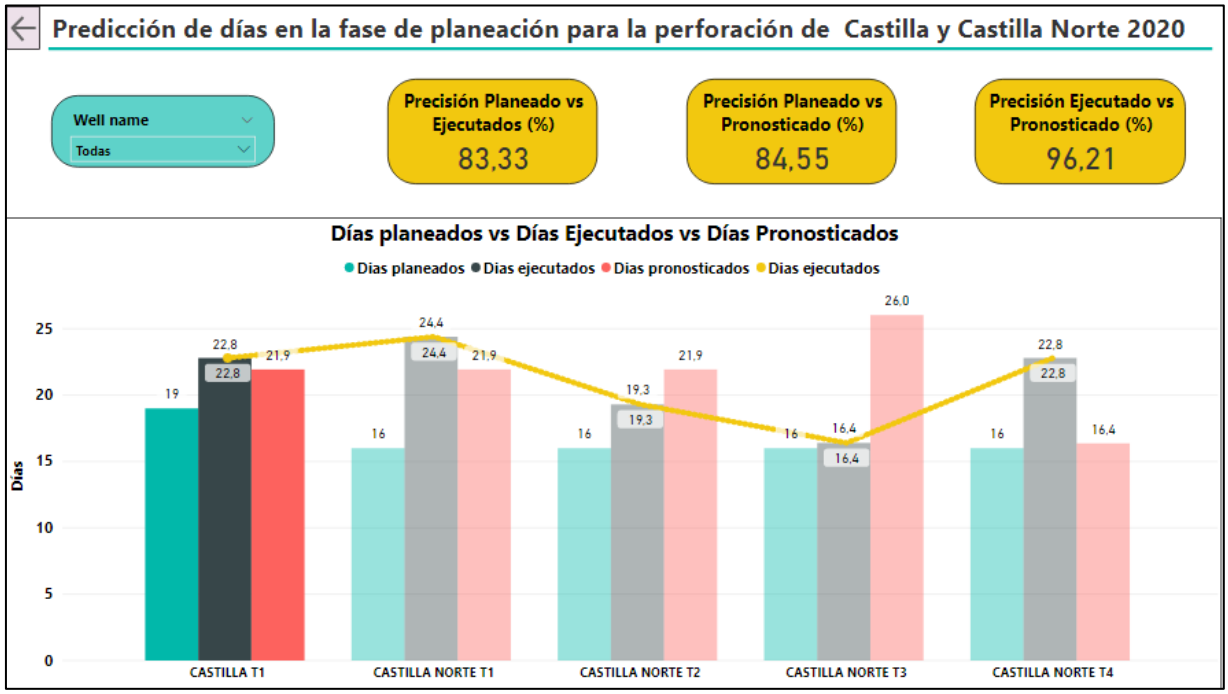


Nota. La figura muestra la comparación entre los días planeados vs ejecutados vs pronosticados.

Por otro parte, cuando se evaluó las predicciones realizadas por el modelo DTR para días pronosticados contra los días ejecutados el porcentaje de precisión promedio es del 77,1%, esto demuestra que el uso de una herramienta predictiva asertiva brinda al personal de planeación otro punto de referencia basado en datos reales como lo son las variables implicadas en la matriz de complejidad y por lo tanto contribuye a la optimización de la planeación.

Figura 38.

Tablero dinámico para la variable días.



Nota. La figura muestra al detalle el pozo seleccionado Castilla T1.

En la **Figura 38** es posible analizar de manera individual los pozos del 2020 con el fin evaluar especificamente el rendimiento del modelo frente a los días planeados, ejecutados y pronosticados. A manera de ejemplo se evaluó el pozo Castilla T1, el cual obtuvo una precision del 83,33% de los dias planeados frente a los dias ejecutados. Por otra parte, se obtuvo una precision de 96,21% de los dias ejecutados en cuanto a los dias pronosticados.

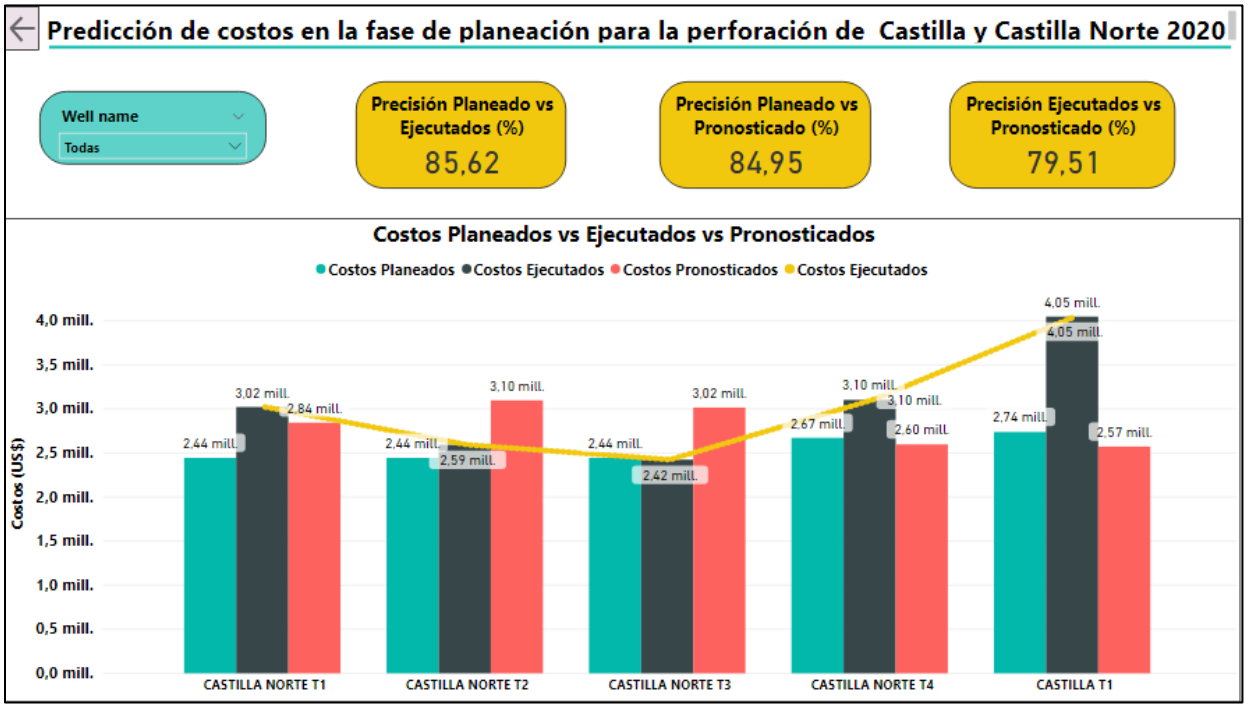
Adicionalmente, es posible observar que los pozos Castilla Norte T3 y Castilla Norte T4 presentan un bajo valor de asertividad en la prediccion por parte del modelo, esto se debe a que se cuenta con una base de datos desbalanceada lo que afecto de manera directa el entrenamiento del modelo dado que no se tuvieron muestras con este tipo de comportamientos.

3.4.2. *Analisis y resultados modelo RFR sobre la camapaña 2020 para costos.*

En la **Figura 39,** se evidenció que durante la fase de planeación para perforación se obtuvo una precisión promedio de 85,62% de los costos planeados con respecto a los ejecutados, esto es lo que actualmente ejecuta la operadora. En la compañía se están planeado los pozos con un costo promedio de $2.548.999 USD mientras que se están ejecutando en un promedio de $3.038.285 USD, es decir que se está planeando por debajo de lo ejecutado, lo cual generaría cambios de alcance para solicitar más recursos económicos.

Figura 39.

Tablero dinámico para la variable costos.



Nota. La figura muestra la comparación entre los costos planeados vs ejecutados vs pronosticados.

A partir de los anteriores datos se obtuvo una precisión promedio de 79,51% entre lo ejecutado versus lo pronosticado, esto fue lo que realizo el modelo predictivo RandomForestRegressor.

Adicionalmente, analizando los pozos a mayor detalle se observó que el pozo Castilla T1, presento un desfase entre el valor ejecutado versus el pronosticado de $1,48 millones USD. Lo anterior fue a causa que durante su ejecución se presentaron eventos de NPT's que aumentaron el costo y por lo tanto afecto de manera directa la predicción realizada por el modelo.

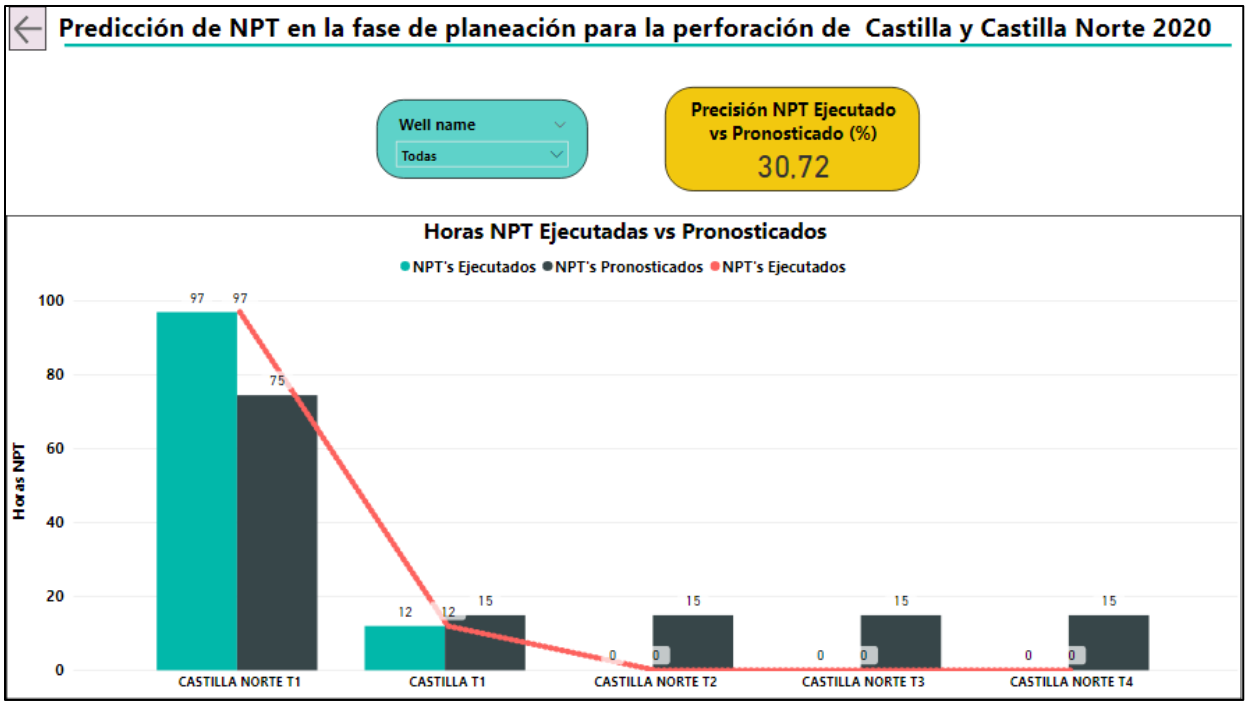
A pesar de las predicciones realizadas por el modelo RFR, esta herramienta resulta ser de utilidad reduciendo la cantidad de horas/hombre para la estimación de costos.

3.4.3. *Analisis y resultados modelo DTR sobre la camapaña 2020 para NPT's.*

En la **Figura 40,** se muestra que durante la fase de perforación se tuvo una precisión promedio de 30,72% de las horas de NPT's ejecutadas con respecto a las pronosticadas. Lo anterior se debe a que se tuvo una base de datos desbalanceada dado que los pozos que se tuvieron para entrenar y probar el modelo predictivo DecisionTreeRegressor en su mayoría presentaron eventos de NPT's asociados a problemas en hueco abierto.

Figura 40.

Tablero dinámico para variable NPT's.



Nota. La figura muestra la comparación entre los costos planeados vs ejecutados vs pronosticados.

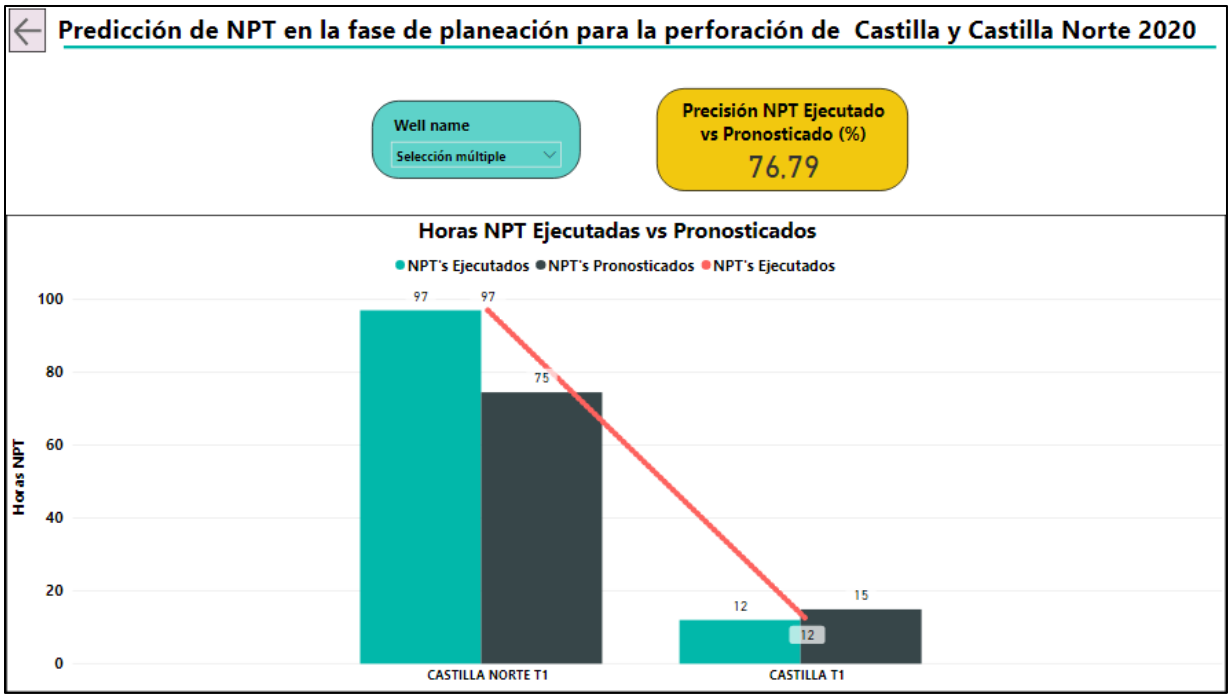
En la anterior figura, se puede observar que los pozos del 2020 Castilla Norte T2, T3 y T4 no presentaron tiempos no productivos relacionados con problemas en hueco abierto. Por lo tanto, se tomó la decisión de realizar un análisis exclusivamente con los pozos Castilla Norte T1 y Castilla T1.

Durante el análisis realizado para estos dos pozos en específico como se muestra en la **Figura 41**, se identificó que el pozo Castilla T1 obtuvo una precisión del 83,78% del valor ejecutado con respecto al valor pronosticado, mientras que el pozo Castilla Norte T1, obtuvo una precisión del 69,8% entre lo ejecutado vs el pronosticado.

En conclusión, el modelo DTR presento una precisión promedio del 76,79% asociado a las predicciones de tiempos no productivos.

Figura 41.

Tablero dinámico para variable NPT’s.



Nota. La figura muestra la comparación entre los costos planeados vs ejecutados vs pronosticados para los pozos con horas de NPT operacional.

Finalmente, al haber realizado la evaluación de los modelos predictivos sobre la campaña de perforación del campo Castilla y Castilla Norte 2020, se determinó que los modelos implementados para llevar a cabo las predicciones para días, costos y NPT’s obtuvieron una precisión promedio de 77,8% teniendo en cuenta que para la predicción de NPT’s solo se evaluaron dos pozos, sin embargo si se hubieran tenido en cuenta todos los pozos para la predicción de NPT’s, el desempeño en general de los modelos seria del 62,45%. Lo anterior se logró exclusivamente a partir de las variables implicadas de la matriz de complejidad relacionadas con perforación.

Hay que resaltar la importancia que tiene este tipo de tecnología en nuestra actualidad, ya que al implementar una herramienta predictiva que considera variables diseño de ingeniería del pozo como variables geomecánicas y geológicas brindaran un soporte confiable y asertivo durante la toma de decisiones.

Este tipo de herramientas puede llegar a ser un elemento indispensable que reducirá el tiempo invertido por un ingeniero durante la fase de planeación de pozos, contando adicionalmente con la predicción de tiempos no productivos asociados a problemas en hueco abierto, lo cual ayudará a realizar una estimación de días y costos mucho más acertada.

Con el propósito de llevar este proyecto más allá de la teoría y fuera de las delimitaciones de este trabajo de grado, se procedió a crear un aplicativo web mediante el uso de la librería *streamlit* y *Visual Studio Code* el cual cuenta con los modelos predictivos seleccionados con la finalidad de implementar tecnologías como Machine Learning para que de esta forma personas sin conocimientos previos en el área de programación puedan hacer uso de esta tecnología.

Lo anterior se puede apreciar en el **Anexo C** y **Anexo D** donde se muestra la interfaz de la herramienta digital.

CONCLUSIONES

Mediante la aplicación de Machine Learning, se logró optimizar el tiempo invertido para la planeación de pozos en la fase de perforación para la estimación de tiempos y costos, que actualmente tiene una duración promedio de 21 horas/hombre por semana (ejecutado por la compañía), lo cual se logró reducir a 15 minutos aplicando los modelos predictivos planteados en este trabajo de grado.

Implementando los modelos predictivos se logró optimizar la inversión por parte de la compañía reduciendo en $ 1.050.000 COP semanal, los costos asociados de un ingeniero de planeación para la estimación de tiempos y costos en el área de perforación.

A partir de las variables implicadas en la matriz de complejidad relacionadas con perforación y el resultado final de la misma, se determinó que es posible realizar predicciones asertivas para las variables objetivo días y costos con una precisión (R2) de 77.1% y 79.51% respectivamente, pese a contar con una baja densidad de datos.

El modelo DesicionTreeRegressor presentó un buen desempeño para la predicción de la variable días con 77.1% de precisión (R2), comparando lo ejecutado por la compañía versus lo pronosticado por el modelo.

El modelo RandomForestRegressor presentó un buen desempeño para la predicción de la variable costos con 79.51% de precisión (R2), comparando lo ejecutado por la compañía versus lo pronosticado por el modelo.

El modelo DesicionTreeRegressor presentó un desempeño bajo para la predicción de NPT’s asociados a problemas en hueco abierto con un valor del 30.72% de precisión (R2), comparando lo ejecutado por la compañía versus lo pronosticado por el modelo.

A partir de los resultados obtenidos por la validación cruzada se concluyó que los modelos predictivos no son generalizables frente a nuevos datos, por lo cual se requiere de una mayor cantidad de datos para entrenar y probar los modelos con la finalidad de conseguir predicciones más asertivas.

Mediante la evaluación realizada para los modelos predictivos sobre la campaña de perforación del 2019, se determinó que no fue factible seleccionar un modelo en

específico para la predicción de días, costos y NPT's, dado que el modelo DTR presento una predicción asertiva para días y horas de NPT's mientras que el modelo RFR presento predicciones más asertivas para costos.

Se determino que no todas las variables implicadas en la matriz de complejidad relacionadas con perforación presentaron la misma relevancia, dado que las variables MD_final, Buzamiento y Fact_Separacion, fueron las variables que aportaron en gran medida durante la obtención de las predicciones para días, costos y NPT's.

Realizando un análisis exploratorio sobre el conjunto de datos del 2019, se consiguió que los datos con los cuales fueron entrenados y probados los modelos predictivos, no presentaran valores atípicos que afectaran la precisión de estos.

Los pozos que presentaron valores de días y costos ejecutados por encima de lo planeado, disminuirán el presupuesto asignado para la campaña de perforación lo que a su vez generan cambios de alcance para solicitar más recursos económicos.

La compañía está generalizando los valores para días y costos en la fase de planeación con un promedio de 16,6 días y $2.548.999 USD respectivamente para cada pozo de la campaña de perforación de Castilla y Castilla Norte 2020, sin embargo, se evidencio que los pozos se están ejecutando en un promedio de 21,14 días y con un valor de $3.038.285 USD, mientras que los modelos predictivos realizaron en promedio una estimación para días de 21,64 y para costos de $2.825.893 USD, logrando obtener valores más acertados con respecto a lo ejecutado, convirtiéndose en una herramienta de gran utilidad en la fase planeación de pozos.

Los pozos donde la predicción obtenida por los modelos no fue lo suficientemente asertiva para días y costos, se debe a que *no* se tuvieron en cuenta varias clases de NPT´s operacionales como fallas de herramientas, problemas en taladro, además no se tuvieron en cuenta los NPT's no operacionales que aumentaron el porcentaje de error en un promedio de 37.55% sobre las predicciones realizadas por los modelos.

Las predicciones obtenidas para la cantidad de horas de NPT's, no fue lo esperado debido que para esta variable en específico no se consideraron más aspectos técnicos

como el tipo de lodo de perforación, formaciones litológicas, diámetro de la broca, entre otros.

RECOMENDACIONES

Se recomienda la adición de los pozos perforados durante el 2020 a los datos de entrenamiento y validación, con el fin de continuar alimentando y a su vez conseguir modelos más robustos para la obtención de predicciones más asertivas para el Campo Castilla y Castilla Norte.

Las variables que no aportaron en gran medida durante la toma de decisiones para la ejecución de los modelos predictivos, se pueden prescindir de ellas y en su lugar, se recomienda la integración de otras variables como las implicadas en la matriz de riesgos, con el fin de lograr un mejor asertividad de los modelos predictivos.

La metodología implementada puede ser aplicada en los distintos campos de Ecopetrol, siempre y cuando se cuente con una buena densidad de datos con el propósito de obtener predicciones más asertivas.

Se debe considerar que en una campaña de perforación los pozos no deben planearse con los mismos días y costos, debido a que cada pozo presenta características únicas y de acuerdo a eso se debería realizar su predicción individual.

Se recomienda la implementación del modelo RandomForestRegressor para la predicción de días, dado que este tipo de modelo no presenta predicciones constantes como las que se obtuvieron con la ejecución del modelo DecisionTreeRegressor.

Se recomienda evaluar la implementación de este tipo de modelos predictivos en otras áreas de la compañía como lo son completamiento y producción.

Se puede plantear la creación de una línea base para la fase de planeación de las variables días y costos basándose en las predicciones de los modelos predictivos.

Se recomienda llevar los modelos predictivos a herramientas digitales (Páginas Web) que permitan un fácil manejo y acceso a tecnologías como Machine Learning, dado que no todas las personas cuentan con conocimientos de programación.

BIBLIOGRAFÍA

[1] L. Hollanda, R.Castello-Branco, C.Lins, R.Morais. (2016). La geopolítica de petróleo y gas: El papel de América Latina. [En línea]. Disponible: https://www.kas.de/documents/252038/253252/7_dokument_dok_pdf_43642_4.pdf/1efc8f05-aa3d-f32d-bda7-12eaa80b1283?version=1.0&t=1539651484171.

[2] Cleverdata. "¿Qué es Machine Learning?". [En línea]. https://cleverdata.io/que-es-machine-learning-big-data//. [Acceso: agosto 9, 2020].

[3] Jessamyn Sneed, "Predicting ESP Lifespan With Machine Learning", *OnePetro*, pp 863, jul, 2017, doi: https://doi.org/10.15530/URTEC-2017-2669988 [Acceso: septiembre 9, 2020].

[4] Xinxin Hou et al., "Lost Circulation Prediction in South China Sea using Machine Learning and Big Data Technology", *OnePetro*, pp 1, May, 2020, doi: https://doi.org/10.4043/30653-MS [Acceso: septiembre 9, 2020].

[5] Agencia Nacional de Hidrocarburo. "Estudios integrados y modelamientos". [En línea]. http://www.anh.gov.co/Informacion-Geologica-y-Geofisica/Estudios-Integrados-y-Modelamientos/Presentaciones%20y%20Poster%20Tcnicos/Campos.pdf. [Acceso septiembre 9, 2020]

[6] F. Quiroz Rincón, C. A. Rivera Fraile., *Selección del reacondicionamiento de los pozos candidatos a procesos de optimización de producción en el campo Castilla mediante el análisis de las propiedades petrofísicas y el historial de producción del campo*, tesis pre. Facultad de ingenierías, Fundaciones Universidad de América, Bogotá, Colombia, 2020.

[7] Ecopetrol S.A. Información Interna.

[8] IBM. "¿Qué es la inteligencia artificial?". [En línea]. https://www.ibm.com/ar-es/analytics/journey-to-ai?p1=Search&p4=43700056616721039&p5=e&cm_mmc=Search_Google-_-1S_1S-_-LA_ISA-_-artificial%20intelligence%20ibm_e&cm_mmca7=71700000071219660&cm_mmca8=kwd-

335413671090&cm_mmca9=Cj0KCQjw2or8BRCNARIsAC_ppyb7RVe4WiskEGJ_OV6BrFTIENcd_p2xsgK4OXO9SJ6kr7XsGTd3m84aAnIaEALw_wcB&cm_mmca10=458383629984&cm_mmca11=e&gclid=Cj0KCQjw2or8BRCNARIsAC_ppyb7RVe4WiskEGJ_OV6BrFTIENcd_p2xsgK4OXO9SJ6kr7XsGTd3m84aAnIaEALw_wcB&gclsrc=aw.ds. [Acceso: octubre 11, 2020].

[9] Escuela profesional de nuevas tecnologías. "Historia y evolución de la inteligencia artificial". [En línea]. https://www.cice.es/noticia/historia-evolucion-la-inteligencia-artificial/#:~:text=1956.&text=Los%20cient%C3%ADficos%20Marvin%20L.,por%20primera%20vez%20el%20t%C3%A9rmino. [Acceso: octubre 11,2020].

[10] IBM. "¿Qué es el aprendizaje automático?". [En línea]. https://www.ibm.com/cloud/learn/machine-learning#:~:text=Machine%20learning%20is%20a%20branch,being%20programmed%20to%20do%20so.&text=The%20better%20the%20algorithm%2C%20the,as%20it%20processes%20more%20data. [Acceso: octubre 11, 2020].

[11] IBM. "Enfoques de machine Learning". [En línea]. https://www.ibm.com/es-es/analytics/machine-learning. [Acceso: octubre 11, 2020].

[12] AprendeIA. "Aprendizaje supervisado en Machine Learning". [En línea]. https://aprendeia.com/todo-sobre-aprendizaje-supervisado-en-machine-learning/. [Acceso: octubre 11, 2020].

[13] ScikitLearn. "Arboles de decisión". [En línea]. https://scikit-learn.org/stable/modules/tree.html#regression. [Acceso: octubre 11, 2020].

[14] S. S. Gosavi., *Machine Learning Methods for Fault Classification*. Tesis master. Universidad de Stuttgart, Stuttgart, Alemania, 2014.

[15] DeepAI. "¿Qué es un bosque aleatorio?". [En línea]. https://deepai.org/machine-learning-glossary-and-terms/random-forest#:~:text=What%20is%20a%20Random%20Forest,decision%20models%20to%20improve%20accuracy. [Acceso: octubre 11, 2020].

[16] RandomForest. (08, mayo, 2013). “Definicion Random Forest”. [En linea]. http://randomforest2013.blogspot.com/2013/05/randomforest-definicion-random-forests.html. [Acceso: octubre 11, 2020].

[17] K. J. Butt., A study of feature selection algorithms for accuracy estimation. Tesis master. Universidad Politécnica de Cataluña, Barcelona, España, 2012.

[18] JacobSoft. “Support Vector Regression SVR”. [En línea]. https://www.jacobsoft.com.mx/es_mx/support-vector-regression/. [Acceso: octubre 11, 2020].

[19] Halliburton Landmark. “OpenWells® Operations Reporting Software”. [En línea]. https://www.landmark.solutions/OpenWells. [Acceso: octubre 11, 2020].

[20] Microsoft Power BI. “POWER BI”. [En línea]. https://powerbi.microsoft.com/en-us/. [Acceso: octubre 11, 2020].

[21] Jupyter. “Jupyter Notebook”. [En línea]. https://jupyter.org/. [Acceso: octubre 11, 2020]

[22] IONOS. “Jupyter Notebook: documentos web para análisis de datos y código en vivo”. [En línea]. https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/jupyter-notebook/. [Acceso: octubre 11, 2020]

[23] AI Powered Decisions. “Que es Python”. [En línea]. https://luca-d3.com/es/data-speaks/diccionario-tecnologico/python-lenguaje. [Acceso: octubre 12, 2020].

[24] Journal of Machine Learning Research. “Scikit-Learn: Machine Learning in Python”. [En línea]. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. [Acceso: octubre 12, 2020].

[25] Numpy. “Some information about the NumPy project and community”. [En línea]. https://numpy.org/about/. [Acceso: octubre 12, 2020].

[26] AprendeIA. “Introducción a la librería de Python”. [En línea]. https://aprendeia.com/introduccion-a-numpy-python-

1/#:~:text=NumPy%20es%20un%20paquete%20de,garantizan%20c%C3%A1lculos%20eficientes%20con%20matrices.. [Acceso: octubre 12, 2020].

[27] Learnpython. “Pandas Basics”. [En línea]. https://www.learnpython.org/es/Pandas%20Basics#:~:text=Pandas%20es%20una%20herramienta%20de,observaciones%20y%20columnas%20de%20variables. [Acceso: octubre 12, 2020].

[28] R. Poveda. (27, diciembre, 2020). “¿Que es Matplotlib?”. Raerpo. [En línea]. http://raerpo2.blogspot.com/2011/12/que-es-matplotlib.html. [Acceso: octubre 12, 2020].

[29] InteractiveChaos. “Seaborn”. [En línea]. https://www.interactivechaos.com/manual/tutorial-de-seaborn/presentacion#:~:text=seaborn%20es%20una%20librer%C3%ADa%20de,la%20creaci%C3%B3n%20de%20atractivas%20gr%C3%A1ficas.&text=La%20documentaci%C3%B3n%20oficial%20est%C3%A1%20disponible,habitualmente%20con%20el%20alias%20sns. [Acceso: octubre 12, 2020].

[30] D. R. Ayala T., H. F Torres Q., *Análisis técnico económico del tiempo no productivo (NPT) en las operaciones de perforación del campo oso,* tesis pre, facultad de ingeniería en geología y petróleos, Quito, Ecuador, 2016.

[31] ScikitLearn, “Metrics and scoring: quantifying the quality of predictions”, [En linea]. Disponible: https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score [Acceso: octubre 12, 2020].

[32] Schlumberger Oilfield Glossary. “dip”. [En línea]. Disponible: https://www.glossary.oilfield.slb.com/en/Terms/d/dip.aspx. [Acceso: diciembre 2, 2020].

[33] Alberta Government. “Alberta petroleum royalty guidelines : principles and procedures”. [En línea]. Disponible: https://open.alberta.ca/publications/alberta-petroleum-royalty-guidelines-principles-and-procedures#detailed. [Acceso: diciembre 2, 2020].

[34] Schlumberger Oilfield Glossary. “unconfined compressive strength (UCS)”. [En línea]. Disponible: https://www.glossary.oilfield.slb.com/en/Terms/u/unconfined_compressive_strength.aspx . [Acceso: diciembre 2, 2020].

[35] Cambridge Dictionary. “real cost”. [En línea]. Disponible: https://dictionary.cambridge.org/es/diccionario/ingles/real-cost. [Acceso: diciembre 2, 2020]

[36] (13, enero, 2019). “Directional Difficulty Index (DDI)”. Perforador 2.0. [En línea]. https://perforador20.wordpress.com/2019/01/13/directional-difficulty-index-ddi/. [Aceso: diciembre 2, 2020].

[37] Schlumberger Oilfield Glossary. “fault”. [En línea]. Disponible: https://www.glossary.oilfield.slb.com/Terms/f/fault.aspx. [Acceso: diciembre 2, 2020].

[38] Cambridge Dictionary. “optimization”. [En línea]. Disponible: https://dictionary.cambridge.org/es/diccionario/ingles/optimization. [Acceso: diciembre 2, 2020]

[39] Schlumberger Oilfield Glossary. “overburden”. [En línea]. Disponible: https://www.glossary.oilfield.slb.com/en/Terms/o/overburden.aspx#:~:text=1.%20n.%20%5BGeology%5D,of%20interest%20in%20the%20subsurface.&text=Pressure%20versus%20depth%20plot. [Acceso: diciembre 2, 2020].

[40] PetroWiki “Reservoir pressure and temperature”. [En linea]- Disponible: https://petrowiki.org/Reservoir_pressure_and_temperature#Reservoir_temperature. [Acceso: diciembre 2, 2020].

[41] Schlumberger Oilfield Glossary. “profundidad vertical verdadera”. [En línea]. Disponible: https://www.glossary.oilfield.slb.com/es/terms/t/true_vertical_depth#:~:text=La%20distancia%20vertical%20existente%20entre,profundidad%20utilizadas%20por%20los%20perforadores. [Acceso: diciembre 2, 2020].

TÉRMINOS TÉCNICOS

ÁNGULO DE ATAQUE: Es el ángulo entre el pozo y el plano de la capa geológica.

BUZAMIENTO: Es la magnitud de la inclinación de un plano con respecto a la horizontal [32].

CLASIFICACION LAHEE: Es una asignación previa a la clasificación que se asigna a cada pozo en función de las complejidades geológicas y la existencia conocida de acumulación de hidrocarburos en el área donde se perforara el pozo. [33]

COSTO EJECUTADO: Refleja el costo total en que se ha incurrido realmente y que se ha registrado durante la ejecución del trabajo para una actividad o componente. [35]

DATA ANALYZER: Software que permite realizar consultas y análisis de datos de operaciones de pozos.

DDI: Es implementado para evaluar la dificultad relativa que puede ser encontrada al perforar un pozo direccional, los parámetros que intervienen en la ecuación son la MD, TVD, el desplazamiento y la tortuosidad. [36]

DENSIDAD: Unidad de masa sobre volumen, generalmente denotada por g/cm^3

FALLA GEOLOGICA: Una rotura o superficie plana en una roca quebradiza a través de la cual se observa un desplazamiento. Dependiendo de la dirección relativa de desplazamiento entre las rocas, o bloques de falla, a cada lado de la falla, su movimiento se describe como normal, inverso o cizallamiento. [37]

OPTIMIZACIÓN: La optimización es la acción de desarrollar una actividad lo más eficientemente posible, es decir, con la menor cantidad de recursos y en el menor tiempo posible. [38]

OVERBURDEN: Roca que recubre un área o punto de interés en el subsuelo. [39]

PREDICCIÓN: Es una declaración sobre lo que se cree que sucederá en el futuro.

PRESION YACIMIENTO: Es una medida de la presión del fluido en un yacimiento.

PROFUNDIDAD MEDIDA (MD): Medida final del pozo en pies [Ft].

RESISTENCIA A LA COMPRESIÓN NO CONFINADA (UCS): Es la tensión de compresión axial máxima que puede soportar una muestra cilíndrica derecha de material en condiciones no confinadas: la tensión de confinamiento es cero. [34]

TEMPERATURA YACIMIENTO: La temperatura del yacimiento se rige principalmente por la proximidad del reservorio al manto terrestre, y por las capacidades relativas de intercambio de calor y conductividades térmicas de las formaciones que forman la secuencia litoestática que incluye el reservorio. [40]

TIEMPO EJECUTADO:

TVD: Es la distancia vertical entre un punto en el pozo (usualmente la profundidad actual o final) y un punto en la superficie. [41]

VS: Es la distancia horizontal desde la línea central de un pozo hasta cierto punto en la trayectoria de un pozo, medida a lo largo de un azimut predefinido en un plano horizontal.

ANEXOS

ANEXO A

Cantidad	Campo	Pozo
1	Castilla	Castilla J1
2	Castilla	Castilla J2
3	Castilla	Castilla J3
4	Castilla	Castilla J4
5	Castilla	Castilla J5
6	Castilla	Castilla J6
7	Castilla	Castilla J7
8	Castilla	Castilla J8
9	Castilla	Castilla J9
10	Castilla	Castilla J10
11	Castilla	Castilla J11
12	Castilla	Castilla J12
13	Castilla	Castilla J13
14	Castilla	Castilla J14
15	Castilla	Castilla J15
16	Castilla	Castilla J16
17	Castilla	Castilla J17
18	Castilla	Castilla J18
19	Castilla	Castilla J19
20	Castilla	Castilla J20
21	Castilla	Castilla J21
22	Castilla Norte	Castilla Norte J1
23	Castilla Norte	Castilla Norte J2
24	Castilla Norte	Castilla Norte J3
25	Castilla Norte	Castilla Norte J4
26	Castilla Norte	Castilla Norte J5
27	Castilla Norte	Castilla Norte J6
28	Castilla Norte	Castilla Norte J7
29	Castilla Norte	Castilla Norte J8
30	Castilla Norte	Castilla Norte J9
31	Castilla Norte	Castilla Norte J10
32	Castilla Norte	Castilla Norte J11
33	Castilla Norte	Castilla Norte J12
34	Castilla Norte	Castilla Norte J13
35	Castilla Norte	Castilla Norte J14
36	Castilla Norte	Castilla Norte J15
37	Castilla Norte	Castilla Norte J16
38	Castilla Norte	Castilla Norte J17
39	Castilla Norte	Castilla Norte J18
40	Castilla Norte	Castilla Norte J19
41	Castilla Norte	Castilla Norte J20
42	Castilla Norte	Castilla Norte J21
43	Castilla Norte	Castilla Norte J22
44	Castilla Norte	Castilla Norte J23
45	Castilla Norte	Castilla Norte J24
46	Castilla Norte	Castilla Norte J25
47	Castilla Norte	Castilla Norte J26
48	Castilla Norte	Castilla Norte J27
49	Castilla Norte	Castilla Norte J28
50	Castilla Norte	Castilla Norte J29
51	Castilla Norte	Castilla Norte J30
52	Castilla Norte	Castilla Norte J31
53	Castilla Norte	Castilla Norte J32
54	Castilla Norte	Castilla Norte J33
55	Castilla Norte	Castilla Norte J34
56	Castilla Norte	Castilla Norte J35
57	Castilla Norte	Castilla Norte J36

ANEXO B

Cantidad	Campo	Pozo
1	Castilla	Castilla Norte JJ1
2	Castilla Norte	Castilla Norte JJ1
3	Castilla Norte	Castilla Norte JJ2
4	Castilla Norte	Castilla Norte JJ3
5	Castilla Norte	Castilla Norte JJ4

Anexo C


Variables Matriz de Complejidad
Factor de separación
0.00
0.00
10.00
Gradiente de presión del yacimiento
0.00
0.00
15.00
DDI
0.00
0.00
10.00
Máxima densidad en el Overburden
0.00
0.00
15.00
VS/TVD
0.00
0.00
1.00
Dureza de la roca
0
0
15000
Fallas
0
0
5
Profundidad Final del pozo (MD)
0
0
20000
Buzamiento
0.00
0.00
50.00
Resultado final Matriz
0.00
0.00
50.00
Ángulo de ataque
0.00
0.00
90.00

Anexo D

Variables Matriz de Complejidad

	Input	Unidades
Factor de separación	0	Adimensional
DDI	0	Adimensional
VS/TVD	0	Adimensional
Fallas	0	Adimensional
Buzamiento	0	Grados
Ángulo de ataque	0	Grados
Gradiente de presión del yacimiento	0	ppg
Máxima densidad en el Overburden	0	ppg
Dureza de la roca	0	Psi
Profundidad Final del pozo (MD)	0	Ft
Resultado final Matriz	0	Adimensional

La predicción para días es : `14.2` Días

La predicción para Costos es : US$ `2560662.2358802306`

La predicción para NPTs es : `2.0` Horas

Predicción Exitosa