

**GENERACIÓN DE UN MODELO BASADO EN INTELIGENCIA ARTIFICIAL PARA
PREDECIR EL VOLUMEN DE ACEITE INCREMENTAL POR INYECCIÓN CÍCLICA
DE VAPOR EN UN CAMPO DE CRUDO PESADO DEL VALLE MEDIO DEL
MAGDALENA**

**BRAYAN GUILLERMO PLATA RODRÍGUEZ
HAROL STIVEN AGUDELO BETANCOURTH**

**PROYECTO INTEGRAL DE GRADO PARA OPTAR AL TÍTULO DE
MAGISTER EN INGENIERÍA DE YACIMIENTOS**

DIRECTORA

**ADRIANGELA CHIQUINQUIRA ROMERO SANCHEZ
INGENIERO DE PETRÓLEO, MSC. EN INGENIERIA DEL GAS Y MSC. EN GESTIÓN
AMBIENTAL PARA LA COMPETITIVIDAD**

CODIRECTOR

**JORGE EDUARDO ROMERO DOMÍNGUEZ
INGENIERO DE PETRÓLEO, MSC. EXTRACCIÓN DE CRUDOS PESADOS**

**FUNDACIÓN UNIVERSIDAD DE AMÉRICA
FACULTAD DE INGENIERÍAS
MAESTRÍA EN INGENIERÍA DE YACIMIENTOS
BOGOTÁ, D.C.**

2025

NOTA DE ACEPTACIÓN

**Firma del director
jurado**

Firma del presidente del

Firma del jurado

Firma del jurado

Bogotá D.C. marzo de 2025

DIRECTIVOS DE LA UNIVERSIDAD

Presidente de la Universidad y Rector del Claustro

Dr. Mario Posada García Peña

Consejero Institucional

Dr. Luis Jaime Posada García Peña

Vicerrectora académica y de Investigación

Dra. María Fernanda Vega de Mendoza

Vicerrector Administrativo y Financiero

Dr. Ramiro Augusto Forero Corzo

Secretario General

Dr. José Luis Marcias Rodríguez

Decano De La Facultad De Ingenierías

Ing. Naliny Patricia Guerra Prieto

Director Del Programa de Ingeniera de Petróleos

Ing. Naliny Patricia Guerra Prieto

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

DEDICATORIA

A nuestras familias, por ser el motor e inspiración durante estos años de estudio y sacrificios para lograr los objetivos. En especial a Ana, Santi, Elena, Luis, Martin y Paula, por ser la motivación y razón de continuar con estos proyectos a pesar de los retos presentados.

A mi compañero de trabajo, Brayan Plata, por su entrega, compromiso y lucha constante durante esta etapa.

AGRADECIMIENTOS

A nuestras familias por la paciencia, amor y acompañamiento durante este proceso.

A nuestros compañeros de maestría y amigos, gracias por su apoyo y compañía.

A la Ing. Adriangela Romero, que por su entrega, disposición, acompañamiento y colaboración brindada fue pieza fundamental para el desarrollo de este trabajo.

Al Ing. Juan David Baquero y Sebastián Nava, quienes, a través de su apoyo y acompañamiento en el proceso, fueron pieza fundamental para darnos soluciones donde creíamos que no había.

A ECOPETROL, por darnos la oportunidad de realizar este proyecto y brindarnos las herramientas para sustentar el trabajo.

A la Universidad de América y los profesores de la maestría por el acompañamiento en el proceso, los conocimientos transmitidos y la orientación para encontrar respuestas y soluciones a los problemas presentados.

A todas las personas mencionadas por demostrarnos que, con amor, entrega y mucha paciencia es posible lograr cada meta propuesta y que además comprobaron que el verdadero amor es el deseo de ayudar al otro para que se supere.

TABLA DE CONTENIDO

	pág.
RESUMEN	13
INTRODUCCIÓN	14
1. OBJETIVOS	15
1.1. Objetivo general	15
1.2. Objetivos específicos	15
2. MARCO TEÓRICO	16
2.1. Yacimientos de crudo pesado	16
2.2. Recobro Mejorado	17
2.3. Recobro térmico	18
2.3.1. Inyección Cíclica de Vapor	18
2.4. Criterios de selección de yacimientos para un proyecto de Inyección Cíclica de vapor	21
2.5. Problemas en la inyección de vapor	22
2.5.1. Baja inyectividad	22
2.5.2. Reducción en la eficiencia del vapor	24
2.6. Código de programación	26
2.7. Lenguajes de programación	26
2.7.1. Python	26
2.7.2. Java	26
2.7.3. R	27
2.7.4. C o C++	27
2.8. Algoritmo	27
2.9. Deep Learning	27
2.10. Análisis de datos	28

2.11. Big data	29
2.12. Machine learning	29
2.12.1. Métodos de Machine learning más utilizados en la industria	30
3. METODOLOGÍA Y DATOS	33
3.1. Fase 1. Descripción del área de estudio y del proceso de Inyección cíclica de Vapor	34
3.1.1. Generalidades geológicas del campo	34
3.1.2. Generalidades petrofísicas del campo	36
3.1.3. Historia de desarrollo y producción	36
3.2. FASE 2. Identificación y parametrización de las variables del proceso	40
3.3. FASE 3. Selección de las variables con mayor impacto en la producción de aceite en el campo de estudio	43
3.4. FASE 4. Construcción del modelo basado en Inteligencia Artificial para estimar la producción de aceite asociada a ICV	44
3.4.1. Regresiones Lineales múltiples (MultiOutputRegressor):	46
3.4.2. Elastic Net: model.ElasticNet	47
3.4.3. Red Neuronal (MLP Regressor): sklearn.neural_network - MLPRegressor	48
3.4.4. Random Forest: RandomForestRegressor – RandomForestClassifier	49
3.4.5. Random Forest Multi-Output	50
3.4.6. Comparación de Modelos Construidos	51
3.5. FASE 5. Validación del modelo con información real del área	52
3.5.1. Regresiones Lineales múltiples	52
3.5.2. Elastic Net: model.ElasticNet	53
3.5.3. Red Neuronal (MLP Regressor): sklearn.neural_network – MLPRegressor	53
3.5.4. Random Forest: RandomForestRegressor – RandomForestClassifier	53

3.5.5. Random Forest Multi-Output	54
4. RESULTADOS Y ANALISIS	55
4.1. Parametrización de las variables del proceso	55
4.2. Selección de las variables con mayor impacto en la producción de aceite en el campo de estudio	65
4.3. Construcción del modelo basado en Inteligencia Artificial para estimar la producción de aceite asociada a ICV	69
4.4. Validación del modelo con información real del área	76
5. CONCLUSIONES	78
RECOMENDACIONES	80
REFERENCIAS	81

LISTA DE FIGURAS

	pág.
Figura 1 <i>Distribución de reservas de petróleo en el mundo</i>	17
Figura 2 <i>Esquema del proceso de Inyección cíclica de vapor</i>	19
Figura 3 <i>Esquema pozo inyector de vapor – pozo productor</i>	20
Figura 4 <i>Respuesta de la producción de un pozo a la inyección cíclica de vapor</i>	21
Figura 5. <i>Metodología para Baja Inyectividad de Vapor</i>	23
Figura 6. <i>Metodología para evitar canalización del vapor entre pozos</i>	24
Figura 7. <i>Metodología para evaluación de las alternativas</i>	25
Figura 8. <i>El proceso de descubrimiento de conocimiento en bases de datos</i>	28
Figura 9. <i>Proceso de Machine Learning</i>	30
Figura 10. <i>Metodología para el desarrollo del proyecto</i>	33
Figura 11. <i>Columna estratigráfica Campo de estudio</i>	35
Figura 12. <i>Historia de producción del campo de estudio</i>	38
Figura 13. <i>Mapa del área de estudio</i>	39
Figura 14. <i>Ciclos de inyección de vapor del campo de estudio</i>	40
Figura 15. <i>Metodología general de entrenamiento del modelo de inteligencia artificial</i>	45
Figura 16. <i>Distribución estadística del SOR y Diagrama de caja de la data inicial sin filtros</i>	56
Figura 17. <i>Comportamiento histórico del pozo ejemplo que presentaba SOR anómalo</i>	57
Figura 18. <i>Distribución estadística del SOR y Diagrama de caja de la data filtrada</i>	58
Figura 19. <i>Efecto de las variables operativas en la eficiencia del proceso de inyección cíclica de vapor</i>	59
Figura 20. <i>Diagramas de caja y bigotes de las variables filtradas utilizadas</i>	60
Figura 21. <i>Cantidad de pozos por ciclo de inyección</i>	62
Figura 22. <i>Histograma de las variables durante el ciclo 5 de inyección cíclica de vapor</i>	63
Figura 23. <i>Comportamiento promedio de las variables en cada ciclo de inyección</i>	64
Figura 24. <i>Mapa de calor con la data inicial</i>	66
Figura 25. <i>Diagramas de dispersión de la data inicial</i>	67

Figura 26. <i>Mapa de calor de las variables depuradas incluyendo las variables calculadas</i>	68
Figura 27. <i>Modelo de predicción Elastic Net</i>	71
Figura 28. <i>Comparación de distintos modelos de predicción</i>	72
Figura 29. <i>Código utilizado para el entrenamiento del modelo Multi-Output</i>	74
Figura 30. <i>Ajuste de las variables N_p por ciclo y PIR por ciclo en el modelo Multi-Output</i>	75
Figura 31. <i>Ranking de variables con mayor impacto en la producción de aceite en el campo de estudio</i>	76

LISTA DE TABLAS

	pág.
Tabla 1. <i>Criterios de Selección de Yacimientos para Inyección Cíclica de Vapor</i>	22
Tabla 2. <i>Principales métodos de Machine learning utilizados en la industria del petróleo</i>	31
Tabla 3. <i>Características y propiedades del área de estudio</i>	36
Tabla 4. <i>Variables exportadas de las bases de datos originales</i>	42
Tabla 5. <i>Continuación - Variables exportadas de las bases de datos originales</i>	42
Tabla 6. <i>Rangos operativos de las variables de estudio</i>	61
Tabla 7. <i>Resultados de los modelos de predicción</i>	73
Tabla 8. <i>Comparación de datos reales y los datos predichos por el modelo para diferentes ciclos</i>	77

RESUMEN

La inyección cíclica de vapor es un método de recobro mejorado altamente utilizado en campos de crudo pesado a nivel mundial. Los campos de crudo pesado del Valle Medio del Magdalena tienen las condiciones propicias para la implementación de este método de recobro sin el cual el crudo no tendría la posibilidad de fluir a superficie y el factor de recobro sería casi nulo.

El campo de estudio implementó la inyección cíclica de vapor casi desde el inicio de su desarrollo y cuenta con más de 20 años de producción asociada al recobro térmico a través de la inyección cíclica de vapor. Este desarrollo ha implicado que en promedio se tengan 18 ciclos de inyección por pozo, lo cual a medida que va incrementando en número de ciclos, va disminuyendo la eficiencia entre el volumen inyectado de vapor y el aceite recuperado durante el respectivo ciclo.

El presente estudio tiene como objetivo desarrollar un modelo basado en inteligencia artificial que permita determinar el volumen a recuperar en futuros ciclos de inyección. Para la construcción de este modelo se determinaron las principales variables que afectaban el recobro de aceite durante un ciclo de inyección y como a partir de estos cálculos se podría determinar si un determinado pozo se encuentra en condiciones de tener un nuevo ciclo de inyección y cuál sería su expectativa de valor, así mismo, generar ranqueo de pozos a inyectar, una vez cumplan ciertas condiciones para realizar un nuevo ciclo de inyección.

Palabras clave: Inyección cíclica de vapor, recuperación térmica, inteligencia artificial, recobro mejorado.

INTRODUCCIÓN

Los campos de crudo pesado de la Cuenca del Valle Medio del Magdalena (VMM), aunque son campos maduros, tienen factores de recobro por debajo del 10%, debido a esto, se deben plantear diferentes alternativas de desarrollo para incrementar este factor de recobro.

Históricamente estos campos han estado sometidos a inyección cíclica de vapor como método de recobro mejorado que ha permitido la recuperación del crudo pesado. Técnicamente estos campos de crudo pesado son aptos para realizar inyección continua de vapor bajo las condiciones actuales, sin embargo, los altos costos en las inversiones, la transición energética, la dificultad para la compra de los altos volúmenes de gas requeridos para la generación de vapor y la fluctuación de los precios, hacen que sea poco probable el desarrollo de un proyecto de recobro térmico de esta envergadura. Es allí, donde se deben revisar alternativas de explotación e incremento del factor de recobro. Una de ellas es mejorar el entendimiento y análisis para proyectar de una forma más precisa el volumen de aceite a recuperar por un futuro ciclo de inyección para un pozo en específico.

Para lograr el entendimiento requerido, se debe realizar un análisis pozo a pozo, el cual, en campos en esta etapa de madurez se hace inviable, ya que son campos que cuentan con más de 100 pozos cada uno y que al integrarse con los otros campos pueden resultar en alrededor de 1000 pozos a analizar.

Debido a esto, se debe recurrir a herramientas tecnológicas como la analítica de datos y la inteligencia artificial que permitan el manejo y análisis de una gran cantidad de data histórica con el fin de mejorar el entendimiento de los parámetros y procesos que influyen en la maximización del volumen de crudo a recuperar por un ciclo de inyección de vapor. Esto cobra mayor relevancia cuando se posee una gran cantidad de data histórica y los pozos cuentan con más de 15 ciclos de inyección lo cual comienza a generar una disminución en la eficiencia del proceso.

1. OBJETIVOS

1.1. Objetivo general

Generar un modelo basado en inteligencia artificial para la predicción del volumen de aceite incremental por inyección cíclica de vapor en un campo de crudo pesado del Valle Medio del Magdalena.

1.2. Objetivos específicos

- Determinar las variables que impactan el rendimiento del proceso de inyección cíclica de vapor a partir de analítica descriptiva y diagnóstica en un campo de crudo pesado del Valle Medio del Magdalena.
- Desarrollar un modelo basado en inteligencia artificial para la predicción de la variación de producción de aceite asociado a la inyección cíclica de vapor entrenándolo con los datos históricos del campo.
- Validar la representatividad del modelo construido comparándolo con los datos históricos de inyección de vapor no usados en la calibración.

2. MARCO TEÓRICO

La necesidad de aumentar las reservas de petróleo del país, el potencial petrolífero en campos de crudos pesados y los bajos factores de recobro logrados, ha generado la necesidad de buscar oportunidades en tecnologías emergentes, para mejorar los procesos de recobro implementados en la industria. A enero de 2022, según la Agencia Nacional de Hidrocarburos (ANH), en Colombia se tiene un factor de recobro del 19%. Sin embargo, con la aplicación de estas tecnologías y con precios del petróleo mayores a 80 USD/Bbl es posible llegar al 40 %.

2.1. Yacimientos de crudo pesado

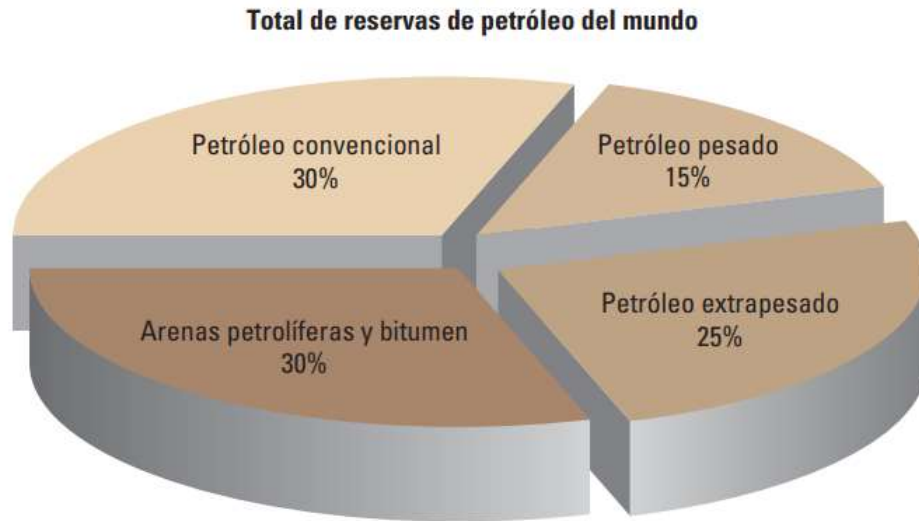
La dificultad en el descubrimiento y adición de reservas a través de nuevos yacimientos de petróleo en el mundo, obliga a la industria a tomar medidas en la inversión de desarrollo de tecnologías que permitan convertir los recursos de petróleo contingentes a reservas. Los yacimientos de crudos pesados, extrapesados, arenas petrolíferas y bitumen conforman aproximadamente un 70 % de los recursos de petróleo totales del mundo como se muestra en la [¡Error! No se encuentra el origen de la referencia.](#) [17]

El crudo pesado se define como petróleo con 22.3° API o menor densidad. Los petróleos de 10° API o menor densidad se conocen como extrapesados, ultra pesados o superpesados porque son más densos que el agua. [17]

Actualmente Colombia cuenta con 7 cuencas sedimentarias que producen petróleo, 3 de ellas producen crudo pesado (API < 22°), como lo es la cuenca de los Llanos Orientales, que cuenta con densidades API entre 12.1 y 18.6°; la cuenca del Valle Medio del Magdalena tiene una producción de crudo pesado de 65 %; en la cuenca del Valle Superior del Magdalena se produce 29.5% de crudo pesado, distribuyéndose 10 departamentos entre estas cuencas [18].

Figura 1

Distribución de reservas de petróleo en el mundo



Nota. En la figura se observa la distribución de reservas mundial según el tipo de crudo. Tomado de: Y. Wu et al, "Feasibility of SAGD as a Follow-Up Process to CSS for a Massive Deep Bitumen Reservoir," SPE Conference Papers, vol. SPE Canada Heavy Oil Technical Conference, pp. 13, 2018. Available: <https://acortar.link/BCHpu6;1.0>. DOI: 10.2118/189750-MS.

2.2. Recobro Mejorado

“El desarrollo de un yacimiento de petróleo obedece a tres etapas, según los métodos de producción que se estén implementando en el campo como lo es el recobro primario, secundario y terciario o recobro mejorado” [19].

Los procesos de recobro mejorado generalmente se dan luego de la implementación de algún proceso de recobro secundario (Inyección de agua o Inyección de gas), sin embargo, esto no implica que el yacimiento siempre tenga esta secuencia en los procesos de recuperación a los que se somete.

“El recobro terciario consiste en la inyección de gases miscibles, químicos y procesos térmicos para desplazar petróleo adicional después de la recuperación secundaria dejará de ser rentable. En los yacimientos de crudos pesados no es factible la inyección de agua, por lo que el uso de procesos térmicos podría ser la única forma de recuperar una cantidad significativa de petróleo” [20]. En este caso de estudio, nos enfocaremos en los procesos térmicos aplicados en un campo maduro de crudo pesado.

2.3. Recobro térmico

Uno de los problemas más relevantes de los yacimientos de crudos pesados son las propiedades físicas del fluido, una de ellas es la viscosidad del petróleo, es por eso que la industria se ha enfocado en desarrollar tecnologías para afectarla directamente a través del incremento de la temperatura, en los procesos de recobro térmico.

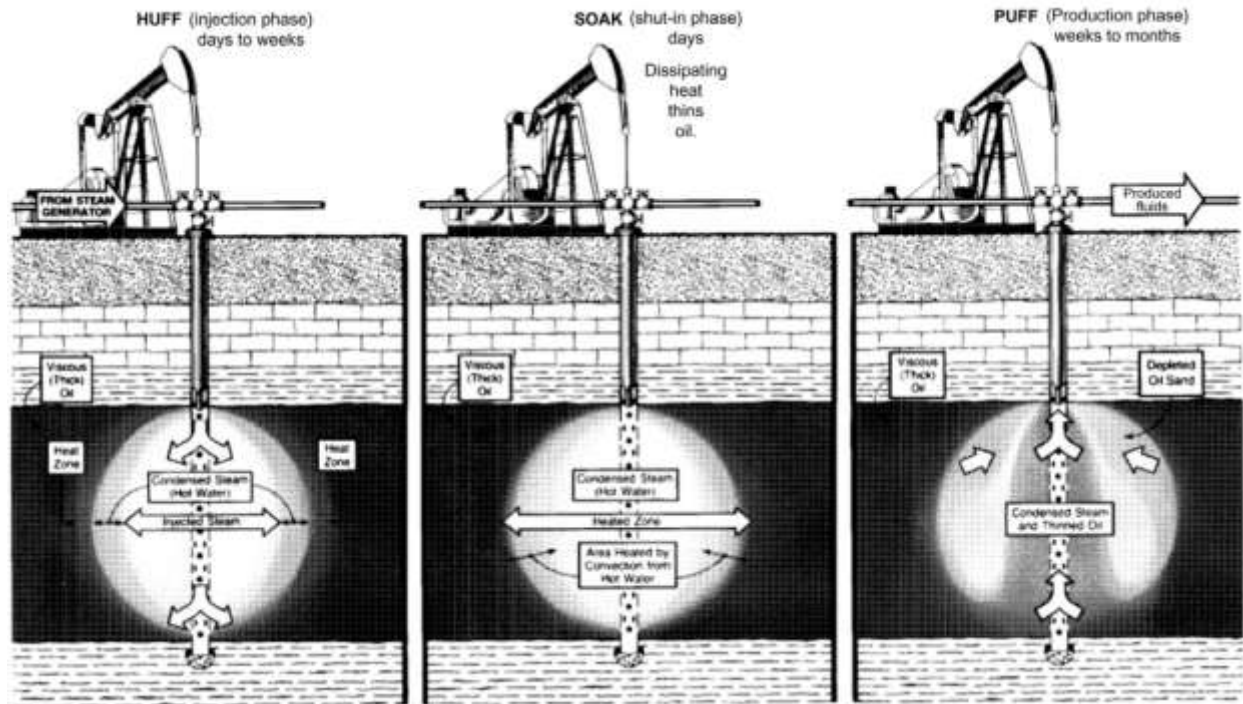
“Los procesos de recuperación térmica consisten en el uso de energía térmica para aumentar la temperatura del yacimiento, reduciendo la viscosidad del petróleo, mejorando la movilidad del fluido al pozo productor; los procesos que más se han implementado para este método son la inyección cíclica de vapor y la combustión in-situ” [20].

2.3.1. Inyección Cíclica de Vapor

“La Inyección Cíclica de Vapor también llamada Huff and Puff, es un método de recuperación térmica que implica la inyección cíclica de vapor con el fin de calentar el yacimiento cerca del pozo, mediante un pozo que se utiliza como inyector y productor; Un ciclo consta de 3 etapas: inyección, remojo y producción, se repite para mejorar la tasa de producción de petróleo” [21], como se muestra en la [¡Error! No se encuentra el origen de la referencia..](#)

Figura 2

Esquema del proceso de Inyección cíclica de vapor

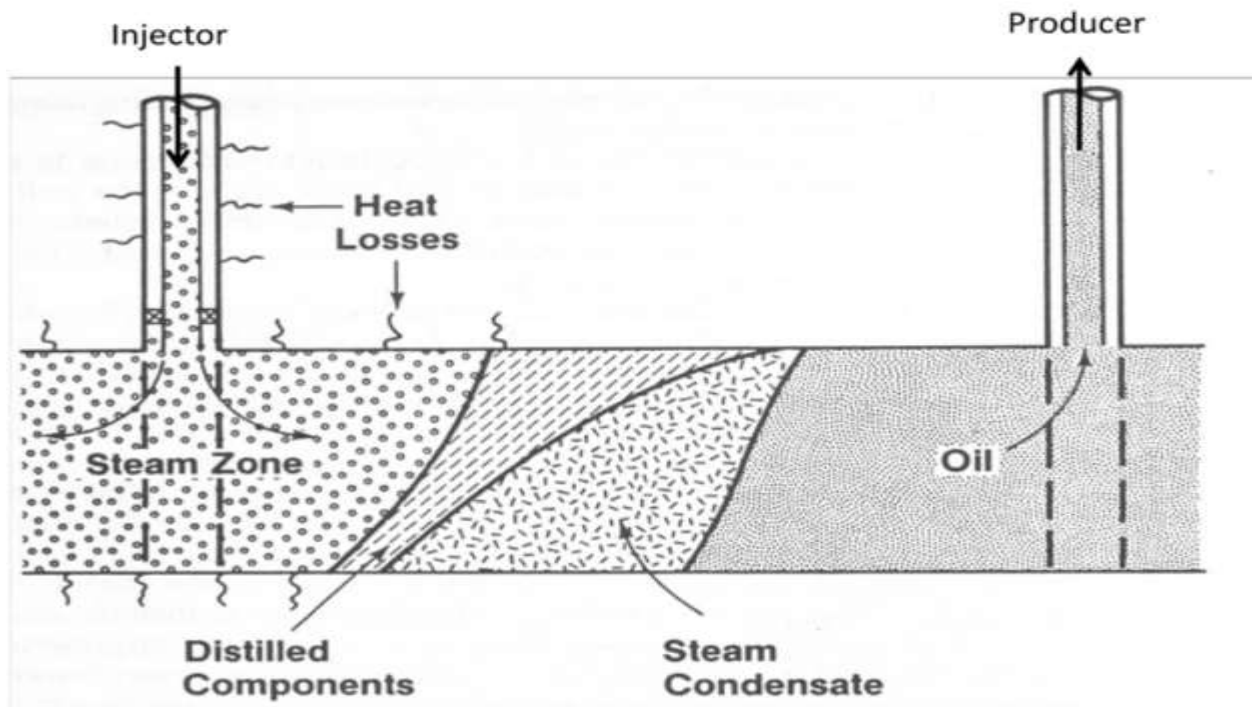


Nota: La figura muestra esquemáticamente el proceso de inyección cíclica de vapor en el yacimiento, siendo el pozo inyector el mismo productor. Tomado de: Johannes Alvarez and Sung-Yun Han, (Jul 01,2013)."Current Overview of Cyclic Steam Injection Process." Journal of Petroleum Science Research.Available: <https://acortar.link/ArTSC0>

En la literatura D. Green y G. Willhite en su libro *Enhanced Oil Recovery*, describen el proceso de la siguiente manera; "Inicialmente se inyecta vapor en un pozo de producción durante un periodo de 2 a 4 semanas. El pozo se cierra dejándolo en remojo antes de volver a producción. La tasa inicial de extracción de petróleo es alta debido a la reducción de la viscosidad del petróleo al aumentar la temperatura del yacimiento y la presión cerca del pozo" [20]. Sin embargo, este método también se implementa en un sistema compuesto por un pozo inyector y un productor, como se muestra en la . No se encuentra el origen de la referencia., siguiendo la misma metodología mencionada anteriormente.

Figura 3

Esquema pozo inyector de vapor – pozo productor

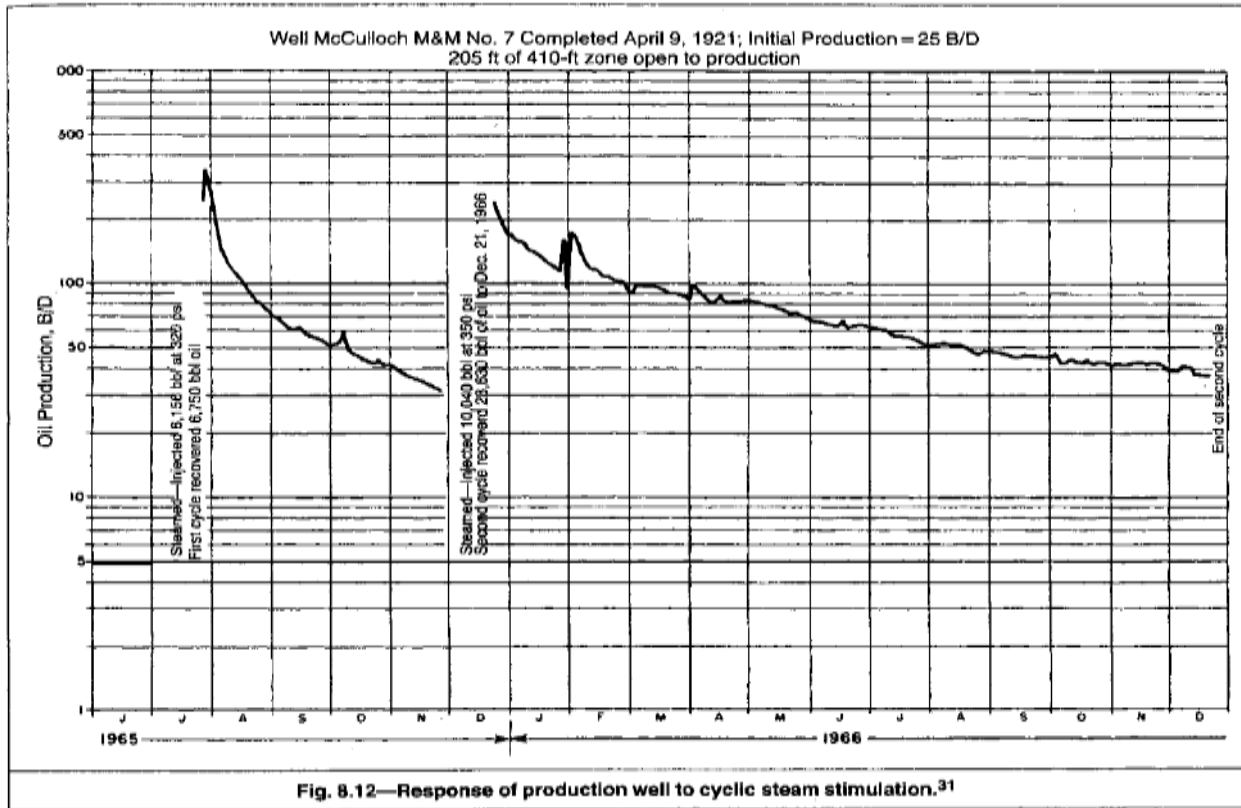


Nota. La figura muestra esquemáticamente el proceso de inyección cíclica de vapor en el yacimiento. Tomado de: D. W. Green and G. P. Willhite, Enhanced Oil Recovery. Texas: Society of Petroleum Engineers, 1998.6.

Las aplicaciones de la inyección cíclica de vapor pueden realizarse en yacimientos de crudos pesados que se encuentren con potencial de energía natural del reservorio o en aquellos que estén depletados y con bajas tasas de producción. El comportamiento típico de los caudales de producción de aceite es de un incremental inicial en la cantidad de fluido producido después del periodo de remojo, luego el comportamiento del caudal disminuye por la pérdida de calor en el sistema, debido al fluido producido y la transferencia de calor con las formaciones adyacentes, resultando en un comportamiento de producción como se observa en la [Figura 3](#), que corresponde al campo Mene Grande en Venezuela, primer campo sometido a inyección cíclica de vapor.

Figura 4

Respuesta de la producción de un pozo a la inyección cíclica de vapor



Nota. En la figura se muestra un comportamiento típico de las tasas de producción de crudo pesado luego de un ciclo de inyección de vapor. Tomado de: D. W. Green and G. P. Willhite, Enhanced Oil Recovery. Texas: Society of Petroleum Engineers, 1998.6

2.4. Criterios de selección de yacimientos para un proyecto de Inyección Cíclica de vapor

En la literatura se han dedicado varios trabajos a establecer criterios para la selección de yacimientos apropiados para la implementación de proyectos de inyección de vapor, teniendo en cuenta “factores geológicos (Barreras de shale, estratificación de permeabilidades, contenido de agua en formaciones adyacentes, espesores de la zona de interés), mecanismo de producción (Sistema de levantamiento) y las características del yacimiento (Presiones, Saturaciones, presencia de gas, temperatura, entre otros)” [22], como se ve en la **Tabla 1**.

2.5. Problemas en la inyección de vapor

Los pozos maduros con más de 15 ciclos de inyección de vapor presentan inconvenientes por la alteración inducida entre la interacción de la roca, el fluido y la energía del yacimiento, se ha evidenciado en varios casos de campos en los que se implementa este tipo de recobro térmico, dejando consigo lecciones documentadas para prever estas situaciones y tener mejores planes de manejo para optimizar el rendimiento de la operación. Unos de estos problemas se mencionan a continuación:

Tabla 1.

Criterios de Selección de Yacimientos para Inyección Cíclica de Vapor

Criterios de Selección Yacimiento para Inyección Cíclica de Vapor		
Espesor de la arena	≥ 30	pie
Profundidad	< 3000	pie
Porosidad	>30	%
Permeabilidad	1000 - 2000	mD
Tiempo de remojo	1-4	días
Tiempo de inyección	14 - 21	días
Número de ciclos	3 - 5	
Saturación de Petróleo	1200	Bbl/acre-pie
Calidad del vapor	80 - 85	%
Gravedad API	<15	°API
Viscosidad del petroleo (Condiciones de yacimiento)	< 4000	cP
Presión de Inyección	< 1400	lpc
Longitud de los ciclos	~ 6	meses
Inyección de vapor / ciclos	7000	Bls
$\frac{kh}{\mu}$	< 200	$\frac{mD - pie}{cP}$

Nota. En la tabla se presentan los criterios de selección de un yacimiento para proyectos de Inyección Cíclica de Vapor. Tomado de: D. Alvarado, C. Bánzer and A. Rincón, Recuperación Térmica De Petróleo. (8th ed.) Caracas: 2002.

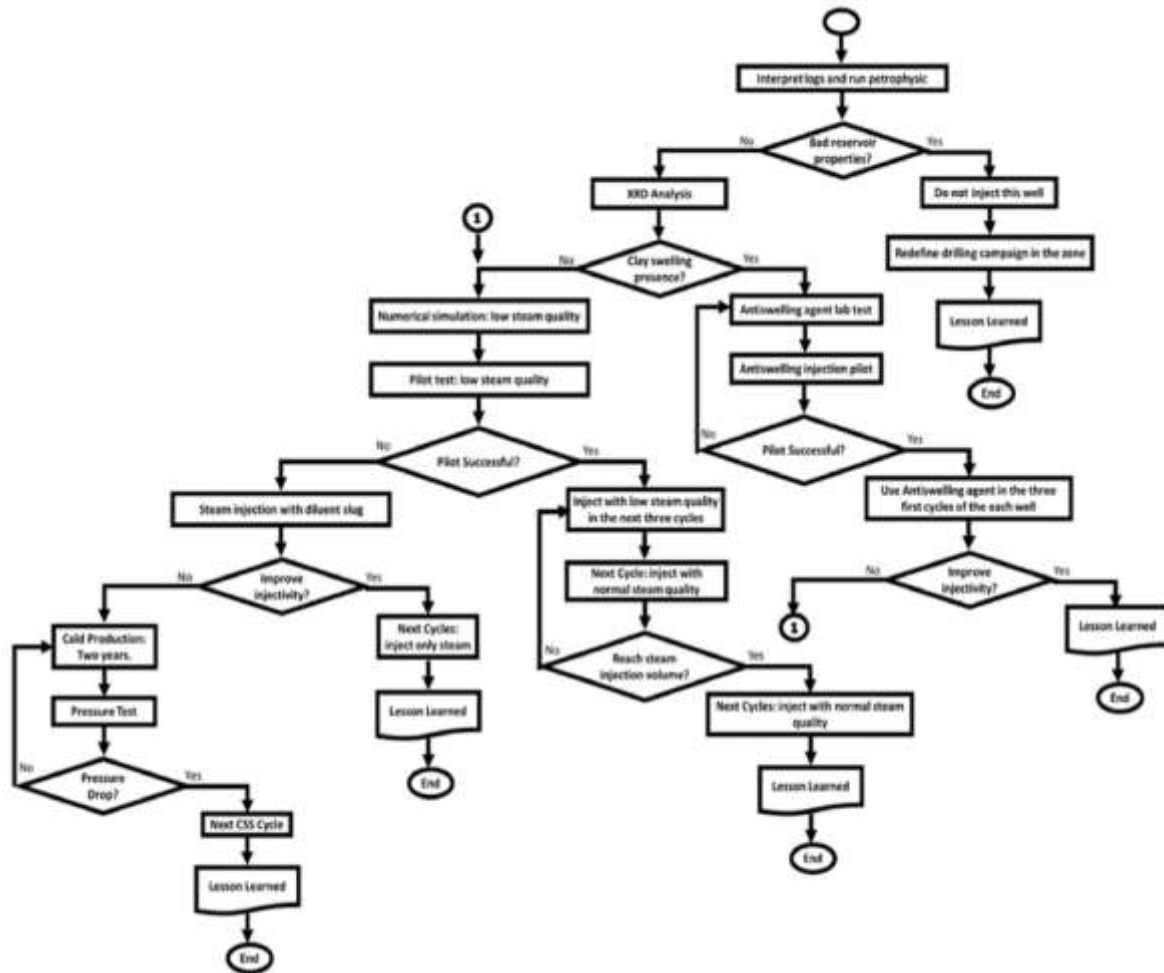
2.5.1. Baja inyectividad

La alteración de las condiciones iniciales del pozo y del yacimiento producto de la producción de hidrocarburos es un factor que en los procesos de Inyección Cíclica de Vapor tiene repercusiones negativas, principalmente en la limitación en la inyectividad de vapor a la formación; Un caso de estudio realizado entre Mansarovar Energy y Ecopetrol identificaron que una de las principales causas de la baja inyectividad son las

malas propiedades del yacimiento, la alta presión, el hinchamiento de la arcilla y el bloqueo de la malla durante la fase de perforación [1]. Por lo que diseñaron una metodología basada en análisis de los equipos de ingeniería de yacimientos de ambas compañías, como se muestra en la **Figura 5**.

Figura 5.

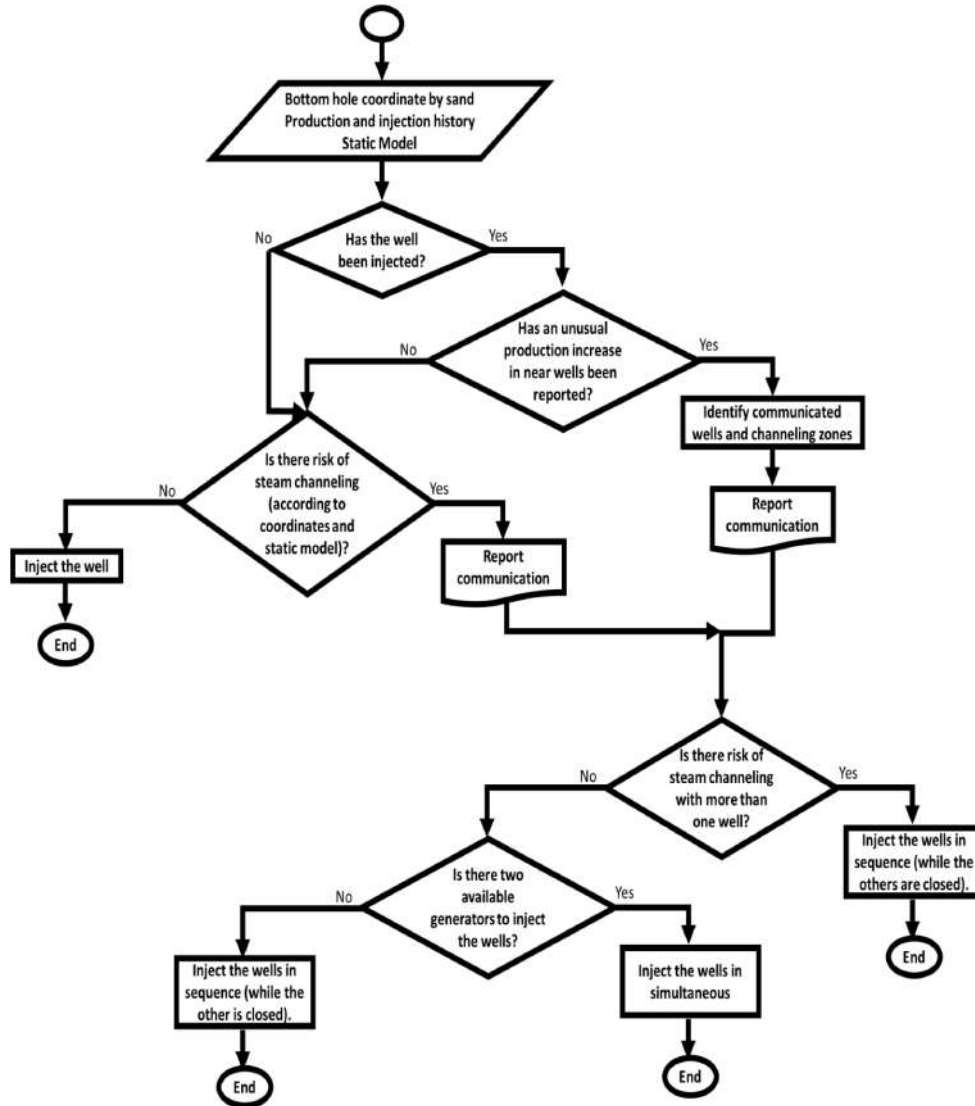
Metodología para Baja Inyectividad de Vapor



Nota. En la figura se presenta un diagrama de flujo para tratar los problemas de baja inyectividad en yacimientos con Inyección de Vapor. Tomado de: E. Trigos, E. Lozano and A. M. Jiménez, "CSS: Strategies to Recovery Optimization," Day 4 Thu, June 14, 2018, 2018. DOI: 10.2118/190791-ms.

Figura 6.

Metodología para evitar canalización del vapor entre pozos



Nota. En la figura se presenta un diagrama de flujo para evitar canalización de vapor entre pozos. Tomado de: E. Trigos, E. Lozano and A. M. Jiménez, "CSS: Strategies to Recovery Optimization," Day 4 Thu, June 14, 2018, 2018. DOI: 10.2118/190791-ms.

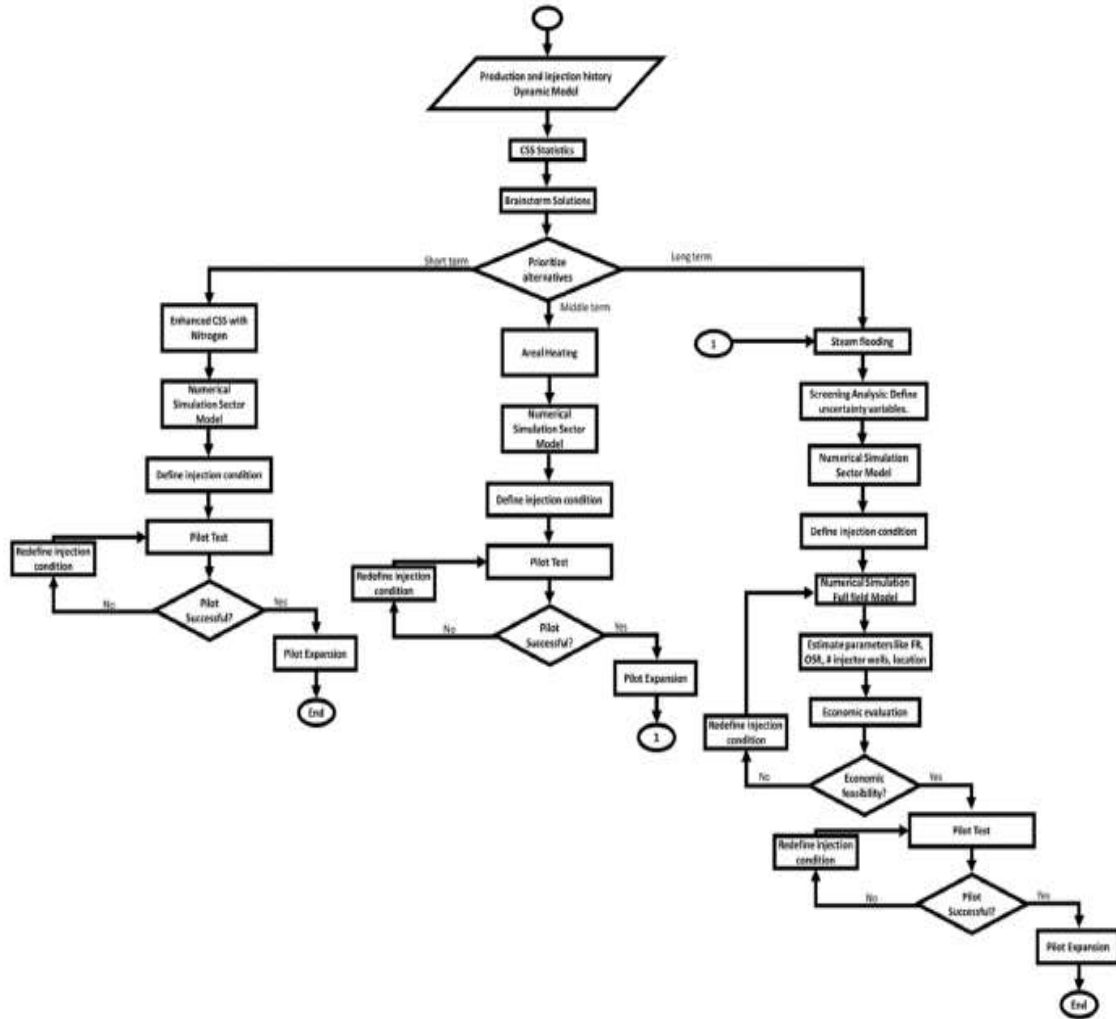
2.5.2. Reducción en la eficiencia del vapor

La reducción en la eficiencia de vapor se evidencia principalmente en la disminución de la Relación Vapor-Aceite (OSR), es decir, se necesita más vapor para recuperar la misma cantidad de aceite incremental, esto se da como consecuencia del aumento en el número de ciclos de inyección. No obstante, existen varias alternativas para contrarrestar este fenómeno y metodologías para evaluar y diseñar alternativas como, la

modificación del esquema de inyección, variando el contenido del fluido a inyectar (Vapor junto con nitrógeno, espuma o gel o inyección continua de Vapor), a continuación, en la **Figura 7** se presenta un diagrama de flujo con la metodología mencionada anteriormente.

Figura 7.

Metodología para evaluación de las alternativas



Nota. En la figura se presenta un diagrama de flujo para evaluar las alternativas para aplicar en pozos con baja eficiencia del vapor. Tomado de: E. Trigos, E. Lozano and A. M. Jiménez, "CSS: Strategies to Recovery Optimization," Day 4 Thu, June 14, 2018, 2018. DOI: 10.2118/190791-ms.

La búsqueda de oportunidades de aplicación de tecnologías modernas a la cadena productiva de la industria petrolera requiere tener bases conceptuales sólidas para entender y comprender la funcionabilidad de estas e identificar los beneficios en su implementación. La incertidumbre es de los factores que más limitan los proyectos, es

por eso, que estas tecnologías pueden tener un papel relevante para soportar planes de desarrollo.

2.6. Código de programación

<<El código de programación es el conjunto de instrucciones que un desarrollador ordena ejecutar a un computador. Este código está estructurado según las reglas correspondiente de cada lenguaje de programación (Sintaxis del lenguaje). La traducción del lenguaje de programación a las instrucciones binarias que entienden las máquinas se realiza mediante compiladores de código o mediante intérpretes de código, según el lenguaje de programación y el entorno elegido>> [23]

2.7. Lenguajes de programación

Como se mencionó anteriormente, la existencia de diferentes lenguajes de programación se da principalmente porque cada uno está diseñado para una necesidad en específico. Actualmente, los lenguajes más destacados son:

2.7.1. Python

Es un lenguaje interpretado y de propósito general que fue desarrollado a principios de los años 90 por Guido Van Rossum en Holanda. Este lenguaje es de los más utilizados actualmente en el desarrollo y construcción de aplicaciones web, desarrollo de software, la ciencia de datos, análisis de datos y Machine Learning (ML). [24]

Los beneficios que hacen de este lenguaje el más usado son:

- Facilidad en la lectura y comprensión de su sintaxis.
- Optimización de la cantidad de líneas de código para escribir un programa.
- Cuenta con una biblioteca estándar con códigos reutilizables.
- Funcionalidad en otros lenguajes de programación como Java, C y C++.
- Capacidad de operar en diferentes sistemas operativos.

2.7.2. Java

Es un lenguaje de programación de propósito general, tipado, multiplataforma y orientado a objetos, adicionalmente es utilizado en la codificación de páginas web, software empresariales y aplicaciones de macrodatos y tecnologías del servidor.

Generalmente se utiliza para el desarrollo de videojuegos, computación en la nube, Macrodatos, Inteligencia artificial (IA) e Internet de los dispositivos. [25]

2.7.3. R

Es un lenguaje de programación ampliamente utilizado en el contexto Data Science e Inteligencia Artificial por tener una comunidad open source muy activa y disponer de gran cantidad de paquetes que disponibilizan, de forma sencilla, algoritmos de predicción, clasificación, optimización, etc. Su mayor aplicación se da en áreas como lo son [26]:

- Investigación científica
- Manipulación de datos
- Análisis estadístico
- Inteligencia Artificial (IA)
- Machine Learning
- Técnicas graficas
- Modelado y predicciones

2.7.4. C o C++

Es un lenguaje de programación compilado, multiparadigma, de tipo imperativo y orientado a objetos desarrollado en 1980, y con amplia aplicación en los sistemas operativos Windows, Mac OS X y Linux, bases de datos, compiladores y navegadores web. [27]

2.8. Algoritmo

Un algoritmo es un conjunto de instrucciones a seguir para resolver un problema concreto. En el contexto de la Inteligencia Artificial, principalmente, utilizamos la palabra algoritmo o pseudocódigo cuando nos referimos a los procesos de entrenamiento de modelos de Aprendizaje Automático o para documentar el diseño del programa.

2.9. Deep Learning

<<El Deep Learning se categoriza como una subfamilia de técnicas que buscan no solamente entrenar un algoritmo para predecir un determinado problema a partir de variables previamente generadas, sino que sea el propio algoritmo el que identifique

los patrones que activarán o no una cierta característica, actuando de la misma manera que las variables en contextos más tradicionales>> [24].

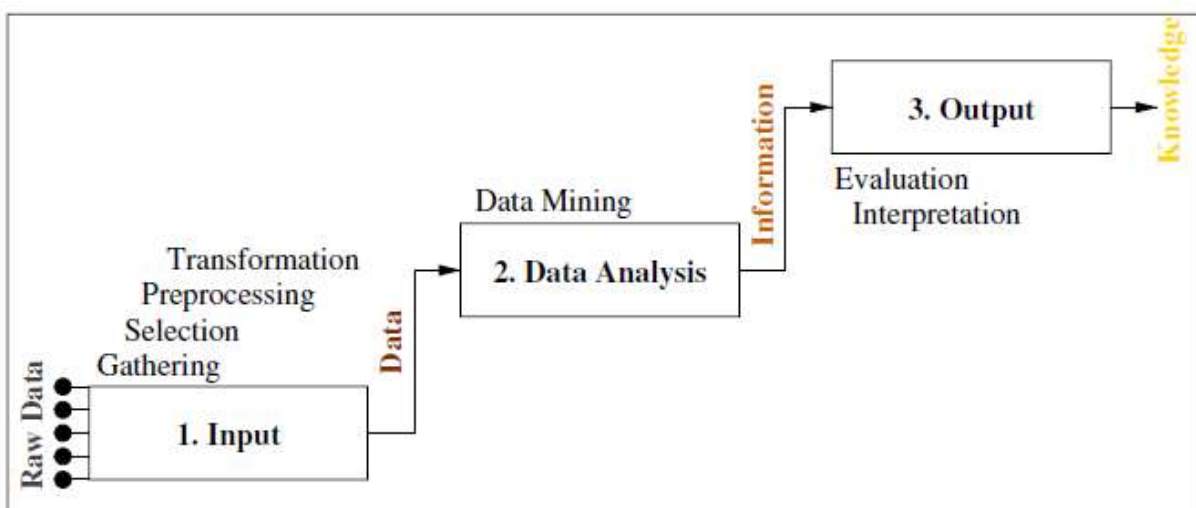
Un factor que impulso el desarrollo del Deep Learning fue el incremento continuo de los datos operativos de las industrias, además de la necesidad de analizarlos e implementarlos en el entrenamiento de algoritmos para mejorar la eficiencia de los procesos.

2.10. Análisis de datos

La obtención de datos es un proceso natural en cada una de las actividades laborales o cotidianas que los seres humanos desarrollamos, se afirma que más del 92% de la información correspondiente a datos se encontraba en medios digitales para el año 2002; la creación de datos suele ser más fácil que la obtención de cosas útiles a partir de estos, es allí donde surge la necesidad de analizar los datos que tiene como objetivo hallar la información más relevante de estos o información oculta a través de diferentes métodos como: muestreo, condensación de datos, enfoques basados en densidad, enfoques basados en cuadrículas, entre otros que a lo largo del desarrollo tecnológico se han logrado automatizar haciendo uso de software. Cabe señalar que el proceso de análisis de datos se lleva a cabo luego de la reunión, selección, preprocesamiento y transformación de datos [28] como lo muestra la **Figura 8**.

Figura 8.

El proceso de descubrimiento de conocimiento en bases de datos



Nota. En la figura se esquematiza el proceso de análisis de datos. Tomado de: Tsai,C. *et al*, 2015). "Big data analytics: a survey." *Journal of Big Data*

2.11. Big data

La obtención de grandes conjuntos de datos ha tenido lugar desde el año 1960, recopilando información a través de diferentes mecanismos, siendo físicos o digitales, esto ha demandado el uso de infraestructura digital en la actualidad que ha dado lugar el nacimiento del término BIG DATA, que hace referencia al conjunto de datos con tres características principales [29]:

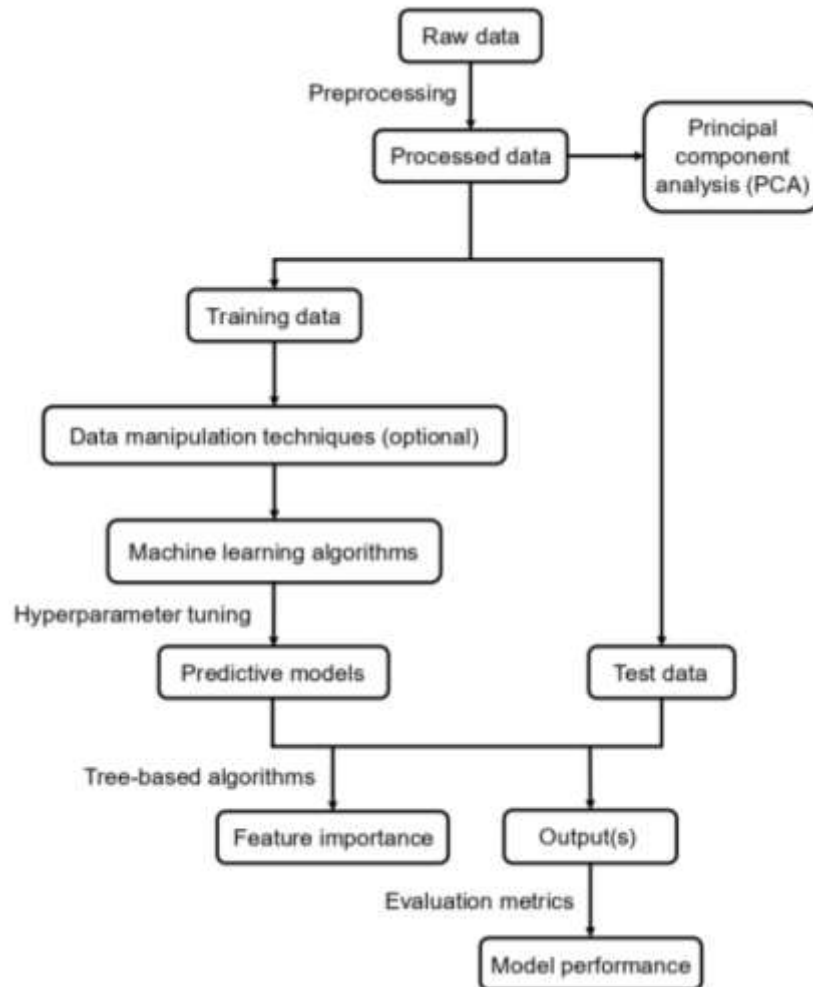
- Volumen: Grandes cantidades de datos con y sin relevancia, obtenidos a través de diferentes mecanismos de lectura o registro, estos pueden llegar a ser decenas de terabytes o petabytes.
- Velocidad: Se obtienen en cortos intervalos de tiempo, lo que demanda grandes espacios de almacenamiento.
- Variedad: Pueden ser datos desde texto, imágenes, sonidos y hasta videos que luego de ser interpretados se almacenan según la necesidad, algunas veces son información irrelevante que requiere procesos posteriores de refinación.

2.12. Machine learning

Corresponde a un subconjunto de la Inteligencia Artificial (IA) que usa algoritmos informáticos para realizar una actividad específica mediante la experiencia y uso de datos, experiencia que adquiere mediante patrones e inferencia de datos para hacer predicciones futuras, con aplicaciones en todos los campos profesionales. Algunos algoritmos de Machine Learning de uso frecuente son: regresión lineal, regresión logística, red neuronal artificial, árbol de decisión, entre otros [30]. El siguiente diagrama de flujo presenta el proceso general del uso de Machine Learning en un proyecto (**Figura 9**).

Figura 9.

Proceso de Machine Learning



Nota. En la figura se muestra el diagrama de flujo para la construcción de un modelo de Machine Learning. Tomado de: Zhong,R., C. Salehi and R. Johnson, (Dec2022)."Machine learning for drilling applications: A review." Journal of Natural Gas Science and Engineering.Available: <https://acortar.link/fUiaCv> DOI: 10.1016/j.jngse.2022.104807.

2.12.1. Métodos de Machine learning más utilizados en la industria

En la industria del petróleo, la capacidad de analizar grandes volúmenes de datos y generar predicciones precisas es crucial para optimizar las operaciones, reducir riesgos y mejorar la eficiencia en la toma de decisiones. Los avances en Machine Learning han permitido a las empresas del sector aprovechar estas tecnologías para resolver problemas complejos como la predicción de la producción de pozos, la caracterización

de reservorios, la optimización de procesos de inyección y la interpretación de datos geofísicos.

Existen diversos métodos de Machine Learning que se adaptan a diferentes necesidades y tipos de datos. Cada uno ofrece ventajas específicas en cuanto a su metodología, precisión y aplicabilidad. A continuación, en la **Tabla 2 y 3** se presenta un cuadro comparativo de los principales métodos utilizados en la industria del petróleo, evaluados según estos criterios para ayudar a seleccionar la técnica adecuada dependiendo del objetivo y los datos disponibles. [31]

Tabla 2.

Principales métodos de Machine learning utilizados en la industria del petróleo

No.	Método	Metodología	Precisión	Aplicabilidad
1	Random Forest	Conjunto de árboles de decisión que mejora la precisión combinando predicciones.	Alta	Aplicado en la predicción de reservas, caracterización de formaciones y estimación de producción, debido a su robustez ante datos ruidosos o complejos.
2	Redes Neuronales Artificiales (ANN)	Modelos supervisados inspirados en el cerebro humano, capaces de aprender patrones no lineales complejos.	Muy Alta	Altamente aplicables en predicciones complejas como la simulación de yacimientos, la predicción de la producción a largo plazo y la optimización de inyección de vapor o gas en pozos de petróleo.
3	Árboles de Decisión	Modelo supervisado que divide los datos en ramas según características importantes.	Moderada a Alta	Útil para problemas de clasificación y regresión, como la predicción de reservas y la optimización de producción de pozos.
4	Regresión Lineal	Modelo supervisado que ajusta una línea recta a los datos para predecir variables continuas.	Moderada a Alta (dependiendo de la calidad de los datos)	Utilizada para modelar relaciones lineales simples, como la predicción de producción de petróleo basada en datos históricos o la estimación de reservas en función de variables físicas y geológicas.

Tabla 2. (Continuación).

No.	Método	Metodología	Precisión	Aplicabilidad
5	Deep Learning	Subtipo de redes neuronales con múltiples capas, adecuado para grandes volúmenes de datos no estructurados.	Muy Alta	Ideal para interpretar datos sísmicos, imágenes geológicas y análisis de big data. Usado en tareas como la clasificación de imágenes sísmicas y la predicción de fallas en equipos de perforación y producción.

Nota. En la tabla se presentan los principales métodos de machine learning utilizados en la industria del petróleo, su precisión y aplicabilidad según las necesidades del estudio.

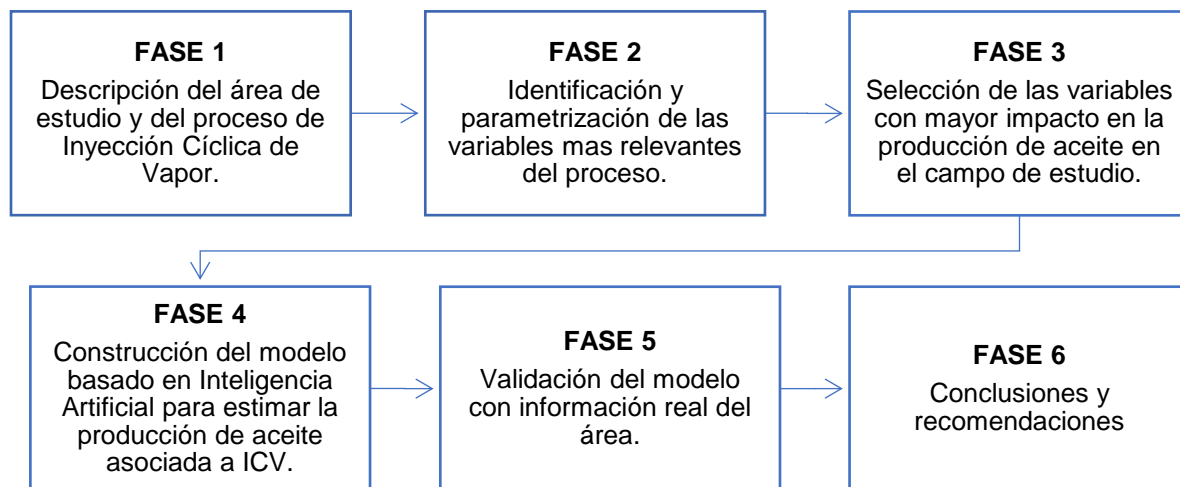
3. METODOLOGÍA Y DATOS

En este capítulo se presenta la metodología utilizada en esta investigación explicativa, enfocada en mejorar la precisión en la predicción del volumen de aceite recuperado mediante inyección cíclica de vapor (ICV). Para ello, se adopta un enfoque cuantitativo orientado a estimar la producción incremental de aceite mediante la construcción de un modelo basado en Inteligencia Artificial. La metodología implementada consta de seis fases para alcanzar los objetivos del proyecto.

1. **Fase 1:** Descripción del área de estudio y del proceso de Inyección Cíclica de Vapor.
2. **Fase 2:** Identificación y parametrización de las variables más relevantes del proceso.
3. **Fase 3:** Selección de las variables con mayor impacto en la producción de aceite en el campo de estudio.
4. **Fase 4:** Construcción del modelo basado en Inteligencia Artificial para estimar la producción de aceite asociada a ICV.
5. **Fase 5:** Validación del modelo con información real del área.
6. **Fase 6:** Conclusiones y recomendaciones

Figura 10.

Metodología para el desarrollo del proyecto



Nota. La figura muestra los diferentes pasos considerados en la metodología de evaluación para la construcción del modelo basado en inteligencia artificial para la predicción del volumen de aceite incremental por inyección cíclica de vapor en un campo de crudo pesado del Valle Medio del Magdalena.

3.1. Fase 1. Descripción del área de estudio y del proceso de Inyección cíclica de Vapor.

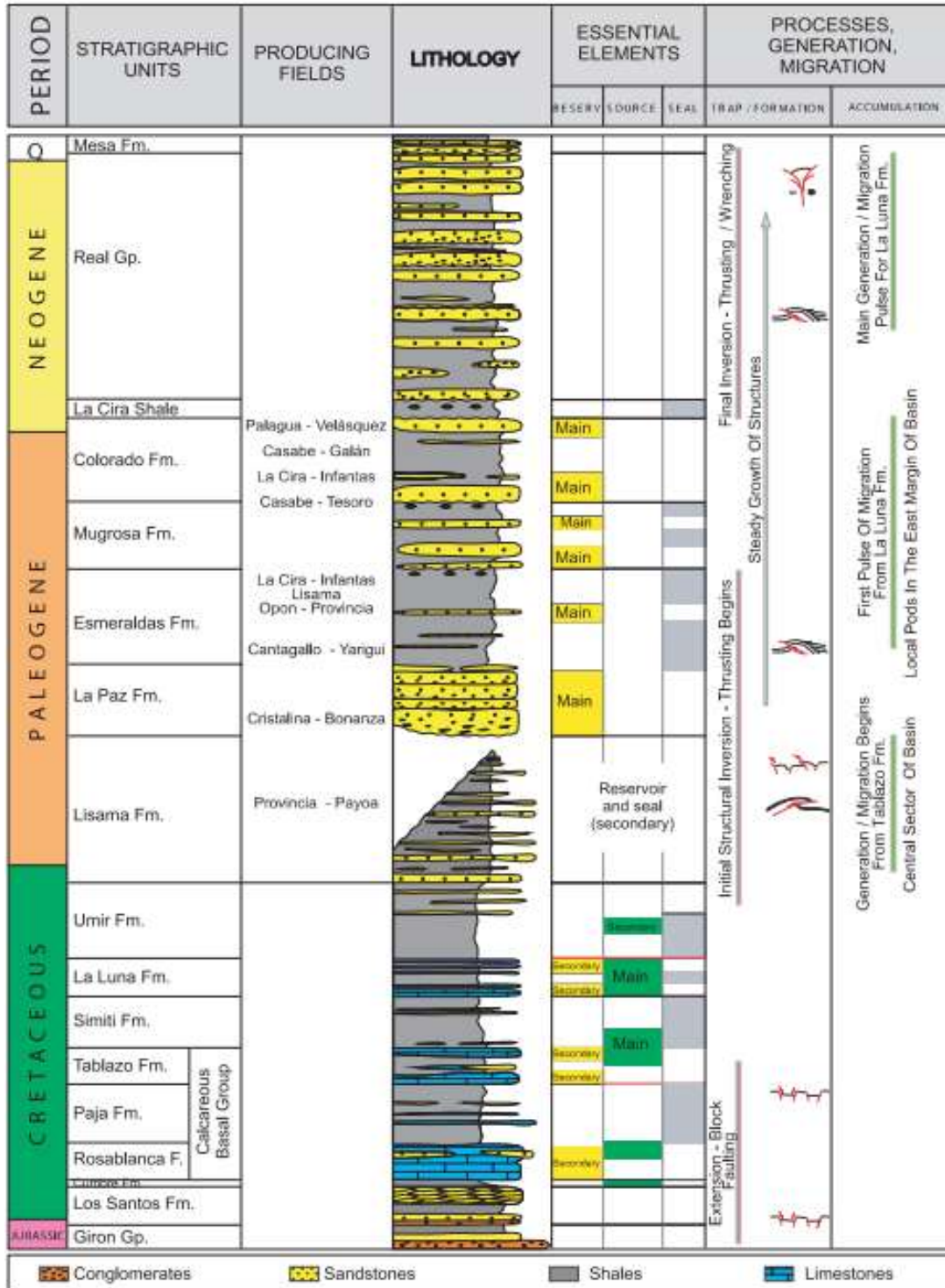
3.1.1. Generalidades geológicas del campo

El primer pozo exploratorio del Campo de estudio se perforó en el año 1982 otorgándose comercialidad en el año de 1987, en el Flanco Oeste de la Cuenca del Valle Medio del Magdalena (VMM), entre las cordilleras Central y Oriental, estratigráficamente el campo tiene formaciones productoras pertenecientes a los paquetes arenosos de las Formaciones Mugrosa y Colorado (Informe del operador), que se originan en el periodo Paleógeno Superior, como se muestra en la **Figura 11**, bajo un ambiente sedimentario de valle fluvial con amplios canales principales y tributarios con desarrollo de facies asociadas, y con intercalaciones de arcillolitas verdosas. Estas formaciones se encuentran en zonas someras lo que ha permitido que el desarrollo del campo se haya enfocado en la inyección cíclica de vapor.

En el desarrollo del campo, el cual inició en el año 2000, se definieron dos zonas reservorio basadas en límites estratigráficos. La zona superior se caracteriza por tener 7 paquetes de conglomerados y areniscas conglomeráticas líticas, estratificadas en bancos gruesos y con intercalaciones de arcillosas grises, por su parte, la zona inferior se compone de 7 paquetes de arenas que se ubican estratigráficamente en equivalencia con la formación mugrosa, lo que hace que se caractericen por ser cuerpos lenticulares de arenisca de grano fino a medio, interestratificados con lodolitas. [32]

Figura 11.

Columna estratigráfica Campo de estudio



Nota. Columna estratigráfica de la Cuenca del Valle Medio del Magdalena. Tomado de: D. Barrero, A. Pardo, C. Vargas, y J. Martínez, *Colombian Sedimentary Basins: Nomenclature, Boundaries and Petroleum Geology, a New Proposal*. ANH and B&M Exploration Ltda., 2007.

3.1.2. Generalidades petrofísicas del campo

En la cuenca del Valle Medio del Magdalena se ha logrado documentar que el 97% de Rocas Reservorio presente en varios de los campos, pertenecen a las areniscas del paleógeno superior, entre ellas se encuentran las formaciones Mugrosa y Colorado, las cuales se caracterizan por tener porosidades del 15 – 20% y permeabilidades de 20 – 600 mD a nivel regional. [32]

En el año 2015 en el campo de estudio se perforaron 4 pozos nuevos, de los que se intervinieron 2 para la toma de núcleos y análisis petrofísicos detallados de la zona superior. De lo anterior, tras la implementación del modelo de porosidad efectiva se determinó que las porosidades presentes en el campo se encontraban entre 20 - 24%, teniendo en cuenta modelos de densidad con matriz de arenisca (Informe del operador). Así mismo, bajo el modelo de permeabilidad basado en registros de Sw y funciones J definidas a partir de datos de presión capilar, se identificó una relación apropiada entre Sw y permeabilidad, obteniendo un rango de trabajo de permeabilidades entre 200 a 2,000 mD y saturaciones de aceite entre 60 y 90%. Las principales características del campo se presentan en la **Tabla 3**:

Tabla 3.

Características y propiedades del área de estudio

Características Área de estudio		
Profundidad	1500 - 2000	ft TVDss
Presión inicial	700 - 1000	psi
T yacimiento	102 - 104	°F
Fluidos del Yacimiento		
API	12 - 14	°API
Viscosidad	2000 - 4000	cP @ T yto
Desarrollo del campo		
Ciclos de Inyección por pozo		12 – 27

Nota. La tabla presenta las principales características del área de estudio en cuanto a sus presiones, tipo de fluido y desarrollo del campo.

3.1.3. Historia de desarrollo y producción

La historia de producción del campo se retoma a los años 80's, con la perforación de 6 pozos que permitieron identificar la presencia de crudo pesado, delimitar el yacimiento y

dar comercialidad al mismo. Inicialmente, se planteó una fase de desarrollo que contaba con la perforación de 103 pozos durante los años 2000 y 2001.

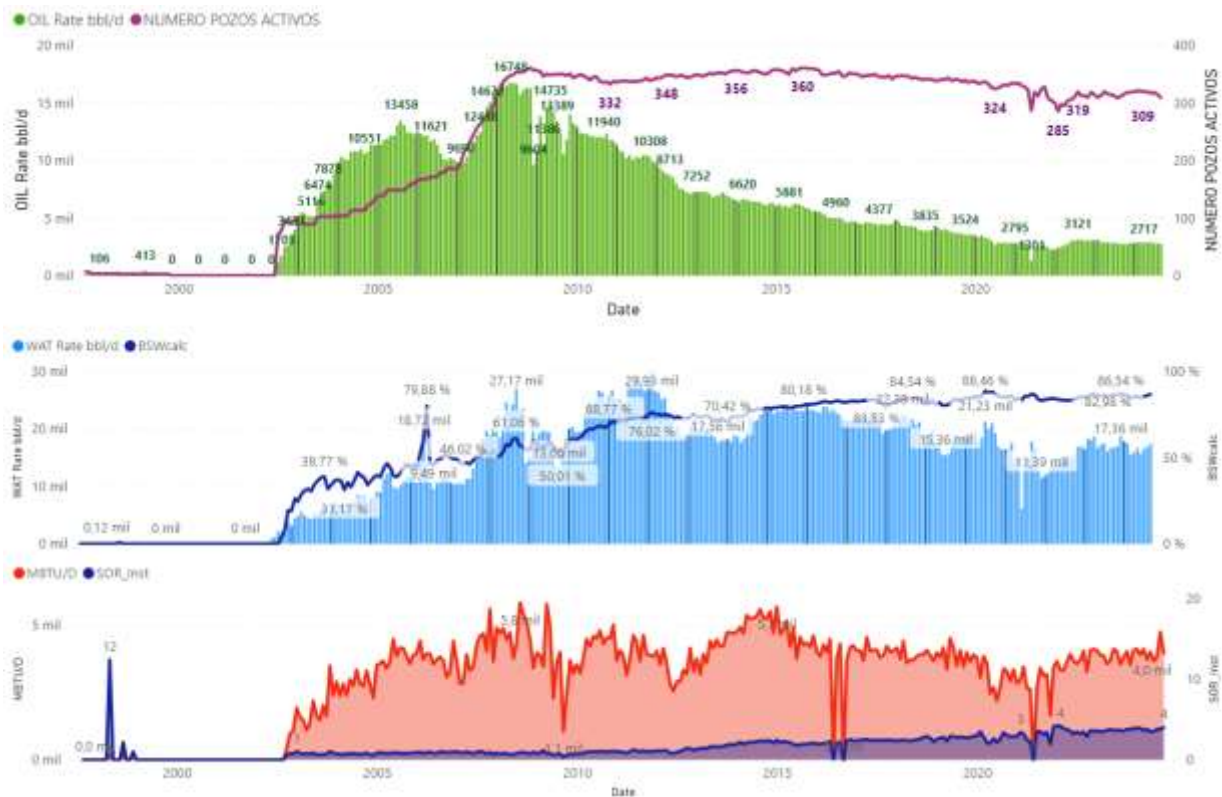
Una vez realizada la perforación en el campo, se inició la producción de crudo, la cual se dio hasta junio de 2002. Debido a la complejidad en la composición y la alta densidad del petróleo producido, fue necesario comenzar la inyección cíclica de vapor seis meses después.

Actualmente, el campo presenta un avanzado estado de maduración en los procesos de inyección de vapor por pozo, lo que ha representado una reducción en la eficiencia del proceso de recobro mejorado a medida que se aumentan los ciclos de inyección. En septiembre de 2024, el campo contaba con más de 400 pozos perforados, de los cuales 309 se encontraban activos. La producción de aceite máxima alcanzada del campo fue de 16.748 BOPD en mayo de 2008, junto con una producción de agua de 20.590 BWPD, un BSW del 53% y una inyección de 4.725 MBTUD en vapor de agua.

Por su parte, la inyección promedio máxima de vapor en el campo se dio en agosto de 2008 con un total de 5.841 MBTUD, lo que represento una producción promedio de 15.704 BOPD de aceite y 24.231 BWPD de agua. El histórico de producción de petróleo, agua e inyección de vapor del campo se muestra en la **Figura 12**.

Figura 12.

Historia de producción del campo de estudio



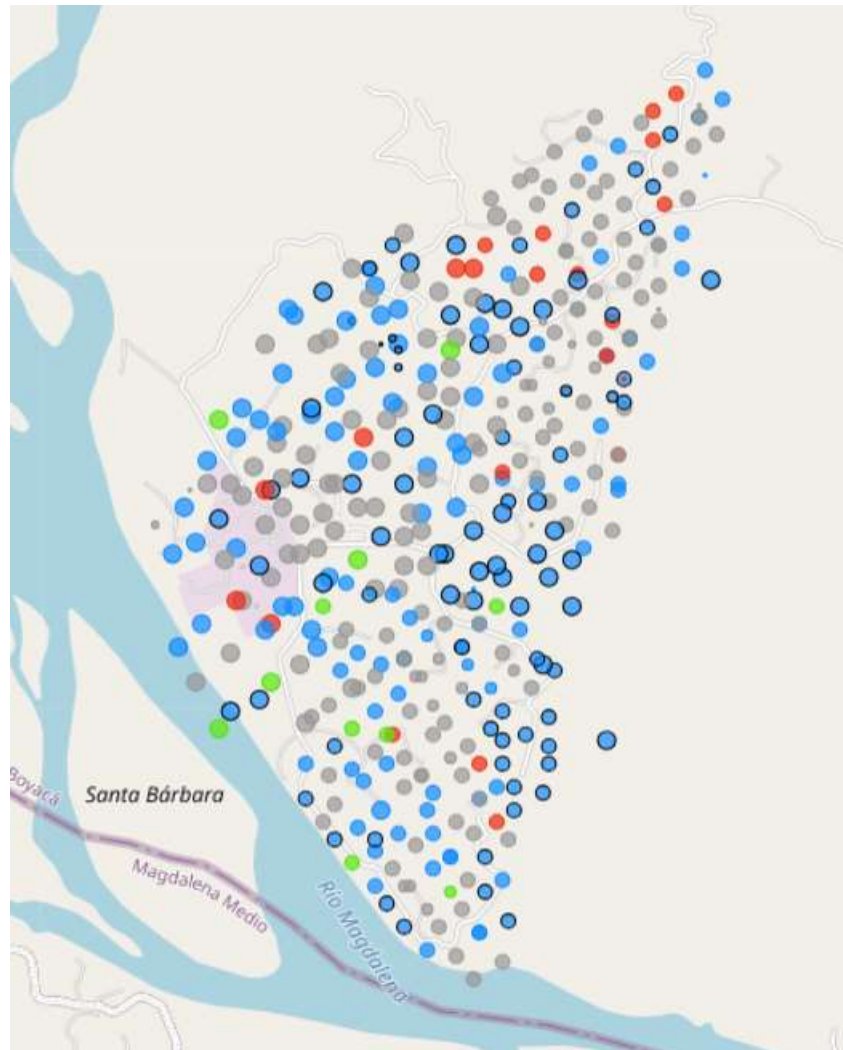
Nota. Producción de aceite, agua e inyección de vapor, tomado de las bases de datos del Operador.

- **Esquema y distribución de pozos en el campo**

Durante la implementación del plan de desarrollo, la necesidad de delimitar y cuantificar los recursos presentes en el campo y optimizar la forma de producción del crudo pesado allí presente, no se implementó un patrón particular de perforación de los pozos, teniendo en cuenta que el método de recobro mejorado pensado para incrementar la productividad del yacimiento podía ser versátil en el uso de la infraestructura de los pozos. Es decir, que cada pozo productor tenía la capacidad de convertirse en un pozo inyector de vapor sin afectar la operación de cada uno de los ciclos. El esquema general de la distribución de pozos en el campo se muestra en la **Figura 13**.

Figura 13.

Mapa del área de estudio



Nota. La figura muestra la ubicación de los pozos del área de estudio, tomado de las bases de datos del Operador

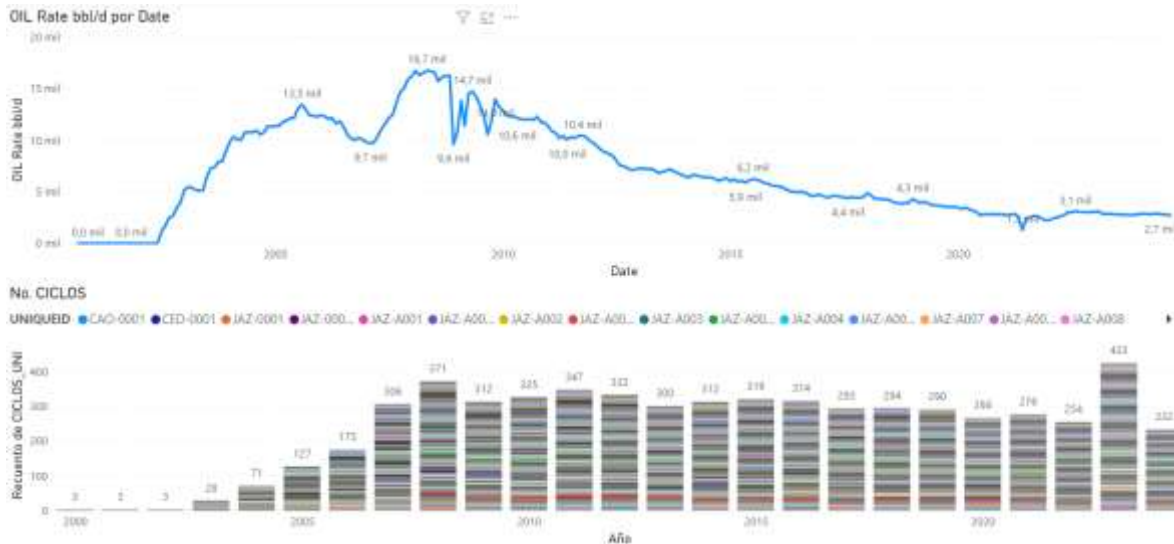
- **Ciclos de Inyección**

Desde los inicios de producción del campo, se identificó que el mecanismo de producción primaria del yacimiento podría verse afectada si no se implementaban procesos externos para incrementar las tasas de producción. Por lo que, la implementación de inyección cíclica de vapor temprana fue fundamental para lograr los factores de recobro que tiene actualmente el campo.

La implementación de ciclos de inyección de vapor en el campo tuvo un crecimiento exponencial entre los años 2003 y 2008, donde se inició con 28 ciclos por año y se

alcanzó un total de 371 ciclos por año, que influyó en alcanzar la tasa máxima de producción de aceite de 16.727 BODP, como se observa en la **Figura 14**. En promedio **Figura 14**.

Ciclos de inyección de vapor del campo de estudio



Nota. La figura muestra el número de ciclos de inyección realizados anualmente en el campo de estudio, tomado de las bases de datos del Operador.

3.2. FASE 2. Identificación y parametrización de las variables del proceso.

En esta fase se identifican y parametrizan las variables más relevantes del proceso de inyección cíclica de vapor (ICV) en el campo de estudio. El proceso comienza con la recopilación y organización de los datos históricos de producción almacenados en el software Oil Field Manager (OFM) y en bases de datos adicionales. Para ello, se utiliza el cuaderno Jupyter Notebook basado en código Python que emplea uso de librerías como *pandas*, *numpy*, *matplotlib.pyplot*, *seaborn*, *missingno* entre otros. Estas librerías incluyen modelos de estadística descriptiva que permiten filtración de datos para obtener mayor calidad de los modelos de aprendizaje.

Esta fase incluye la aplicación de técnicas para comprender los conjuntos de datos y determinar la estadística descriptiva como valores máximos, mínimos, percentiles (25%, 50%, 75%) y desviación estándar, aplicadas a las variables mencionadas anteriormente.

Organiza

El proceso de limpieza de datos (Data Cleaning) incluye la detección y corrección de valores atípicos o nulos, la eliminación de datos duplicados, la estimación de valores faltantes y la verificación de la consistencia y coherencia de los datos y unidades

utilizadas. Lo anterior, busca garantizar la calidad de los datos que serán empleados en el análisis.

Con los datos validados, se realiza un análisis descriptivo de las variables operativas del campo, utilizando herramientas estadísticas y de visualización de datos. Esto permite generar un resumen estadístico de las variables seleccionadas y definir los rangos de datos que se utilizarán en las siguientes fases planteadas.

Dentro del seguimiento que se realiza en campo de la inyección cíclica de vapor se manejan unas variables calculadas las cuales son determinantes para identificar la eficiencia del proceso, algunas de estas son:

- *Steam-Oil Ratio* (SOR): Es la relación entre el volumen de vapor inyectado, expresado en barriles de agua equivalente y el aceite producido en barriles. Esta relación determina cual es la eficiencia del proceso, un SOR más bajo indica un proceso más eficiente, ya que se necesita menos vapor para extraer la misma cantidad de petróleo. Este parámetro es crucial para la evaluación económica y operativa de los proyectos de inyección de vapor. Basados en información histórica de los campos que han utilizado la inyección cíclica de vapor como método de recobro mejorado, un SOR menor a 3 durante un ciclo de inyección es eficiente y genera un VPN positivo al proyecto.

$$SOR = \frac{BWeq\ iny}{oil\ prod\ cum}$$

- *Production-Injection Ratio* (PIR): Es la relación entre el volumen producido total (aceite y agua) expresado en barriles y el de vapor inyectado, expresado en barriles de agua equivalente. Esta relación permite tener un balance de materia de cuanto fluido se está incorporando al yacimiento con respecto al que se está produciendo. Valores menores a 1 indican ineficiencia en el proceso ya que se está inyectando más de lo que se está produciendo y se estaría aumentando la presión en el yacimiento, lo cual, para los procesos de inyección de vapor no es conveniente, puesto que, a mayores presiones, se requiere mayor temperatura para tener el vapor en estado gaseoso y que no se presenten condensaciones tempranas. Así mismo, un valor mayor a 1 indica una eficiencia en cuanto a recobro de fluido, sin embargo, esto no necesariamente indica una eficiencia en cuanto a recuperación de aceite.

$$PIR = \frac{\text{Fluido producido}}{BWEq\ iny}$$

Estas variables calculadas son indispensables para un buen seguimiento del proceso de inyección cíclica de vapor y debido a esto, se deben incluir dentro de la elaboración de los modelos para predecir el aceite incremental asociado a un futuro ciclo de inyección. En las bases de datos iniciales, se identificaron variables asociadas al proceso de ICV, incluidas aquellas calculadas a partir de otros parámetros operativos. Las variables de trabajo definidas inicialmente fueron las siguientes:

Tabla 4.

Variables exportadas de las bases de datos originales

Variable	Nomenclatura
Nombre del Pozo	
Líquidos Acumulados	'LIQ CUM bbl'
Vapor Acumulado	VAP CUM MMBTU'
Caudal inicial	qi'
Inyección de agua equivalente acumulada	BWEiny CUM'
PICO DE PRODUCCIÓN'	
Petróleo Acumulado	OIL CUM bbl'
Agua inyectada equivalente acumulada	WAT INY EQ CUM bbl'

Nota. Variables exportadas de las bases de datos originales

Tabla 5.

Continuación - Variables exportadas de las bases de datos originales

Variable	Nomenclatura
Producción acumulada de petróleo por ciclo	Np por ciclo (total)'
Relación Vapor-Petróleo por ciclo	SOR por ciclo'
DURACIÓN MESES CICLO	
Agua Acumulado	'WAT CUM bbl'
NUMERO DE CICLO	
Relación Vapor-Petróleo Acumulado	'SOR cum'
Relación Producción Líquidos-Inyección por ciclo	'PIR por ciclo'
PICO DE PRODUCCIÓN DE AGUA	
Relación Producción Líquidos-Inyección acumulada	PIR cum'
ESPESOR NETO	
Energía por ciclo	MMBTU Ciclo'

Energía por pie	MMBTU/PIE'
Producción acumulada de agua por ciclo	Wp por ciclo'

Nota. La tabla muestra las variables exportadas de las bases de datos con las cuales se iniciaron los análisis estadísticos.

Aunque el desarrollo del campo de estudio inició en el 2000, los primeros pozos perforados tuvieron data de producción en 1997, fecha en la que se dio comercialidad al campo. De esta forma, de la base de datos de OFM se extrajeron 90318 datos mensuales entre septiembre de 1997 y septiembre del 2024, con un total de 6016 ciclos de inyección distribuidos entre los pozos productores del campo. En promedio 16 ciclos de inyección de vapor por pozo.

Además, se generan diagramas de cajas y bigotes (*box plots*) e histogramas que permiten identificar de forma visual la dispersión, tendencia central y presencia de datos atípicos para cada variable. Este tipo de visualización es crucial para comprender las características de los datos, identificar donde está el dato a revisar y fundamentar la depuración de datos y variables.

La aplicación de esta fase busca aumentar la calidad de los datos cada vez que es aplicada a una base en específico. Es de resaltar que esta metodología puede aplicarse varias veces en función de los resultados que se vayan obteniendo en los modelos desarrollados y en las siguientes fases del proceso.

3.3. FASE 3. Selección de las variables con mayor impacto en la producción de aceite en el campo de estudio

Teniendo en cuenta los resultados de la fase anterior, se aplicarán gráficos como diagramas de dispersión y mapas de calor (*heatmaps*) a las variables y datos obtenidos según sea necesario para explorar correlaciones entre variables y su impacto en la producción acumulada de petróleo.

Para identificar las variables con mayor impacto, se grafica cada una de estas en función del número de ciclos y se aplican mapas de calor. Los mapas de calor muestran la correlación que existe entre dos variables a partir del coeficiente de correlación o dispersión (R^2). El coeficiente de correlación mide la relación lineal entre dos conjuntos de datos, y su valor oscila entre -1 y 1. Las ecuaciones más comunes para calcular estas correlaciones son el Coeficiente de correlación de Pearson, Coeficiente de correlación de Spearman, Coeficiente de correlación de Kendall, Matriz de correlación. Un

coeficiente de 1 indica una correlación perfecta de forma directa, un coeficiente de -1 indica una correlación perfecta de forma inversa y un coeficiente de 0 indica que no existe ninguna correlación entre las variables.

El criterio de selección de las variables a utilizar en los modelos, se determinarán a partir del R^2 el cual debe tener como mínimo un $\pm 0,5$. Adicionalmente, se tendrán en cuenta las variables calculadas como el SOR y el PIR los cuales determinan la eficiencia del proceso de inyección en cada uno de los ciclos.

Además, se realizarán diagramas de dispersión que relacionan las variables con la cantidad de ciclos, junto con histogramas y resúmenes estadísticos para cada variable y ciclo.

Así mismo, con el fin de determinar la importancia y seleccionar las variables para el desarrollo del modelo, se implementa un análisis de importancia de variables a través del algoritmo de machine learning, Random Forest Regressor, ranqueando las variables con mayor impacto sobre el aceite acumulado por ciclo (N_p por ciclo).

Las fases 2 y 3 del del proyecto es la aplicación de la técnica de análisis de datos exploratorio (EDA o “Exploratory Data Analysis”), que incluye entre otros los pasos descritos anteriormente y una vez finalizado se cumple con él primer objetivo específico del proyecto.

3.4. FASE 4. Construcción del modelo basado en Inteligencia Artificial para estimar la producción de aceite asociada a ICV

Para la construcción del modelo de predicción, se planteó el uso de un enfoque basado en inteligencia artificial. Durante el desarrollo del proyecto, se determinó que la generación de múltiples modelos de machine learning sería el enfoque más adecuado para aprovechar los datos históricos del campo de estudio.

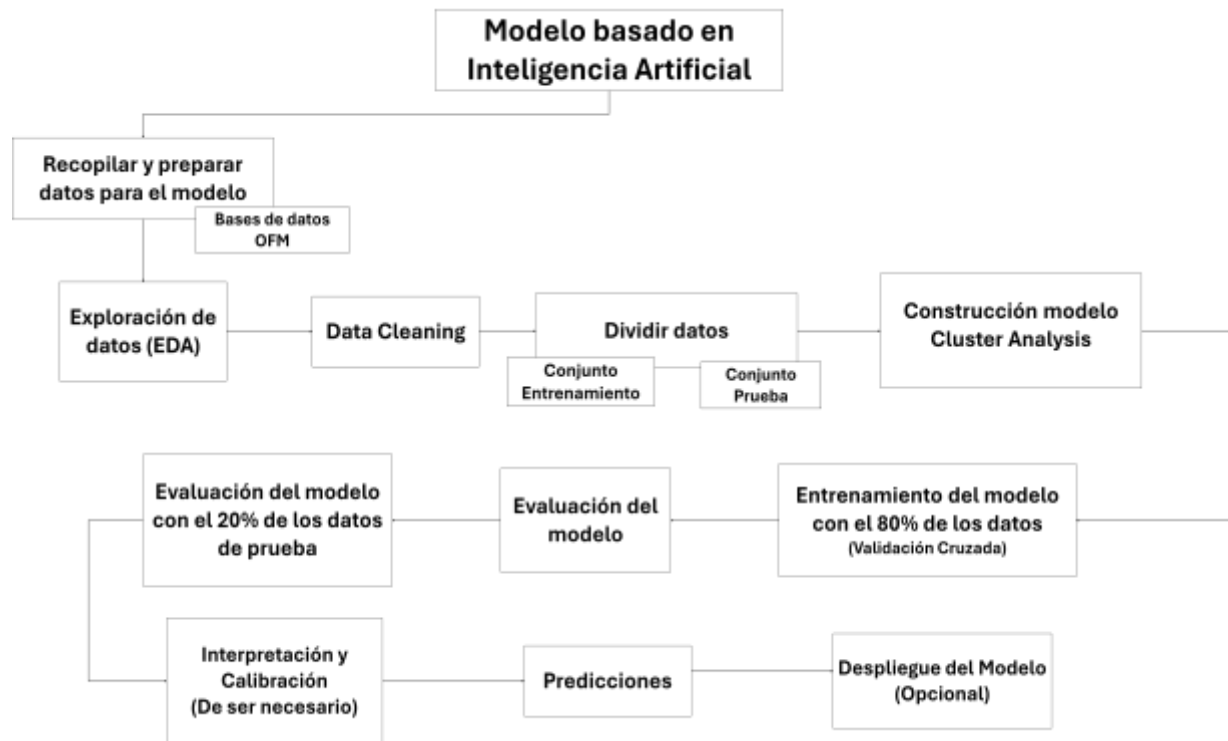
Para ello se incorpora la base de datos procesada en las fases anteriores y se sigue el diagrama de flujo presentado en la figura 15, el cual detalla los pasos necesarios desde la preparación de los datos hasta la evaluación del modelo.

Del diagrama de flujo presente en la figura 15, se destaca que para la preparación de datos se usan las bases de datos organizadas y depuradas en la fase anterior, asegurando su formato, calidad y la normalización de las variables. Adicionalmente, se genera una división de los datos en dos conjuntos:

- Conjunto de entrenamiento: Se compone del 80 % del conjunto de datos históricos del campo.
- Conjunto de prueba: Se compone del 20 % restante, que servirá para validar y evaluar el rendimiento del modelo.

Figura 15.

Metodología general de entrenamiento del modelo de inteligencia artificial



Nota. El diagrama muestra el flujo de trabajo para la elaboración del modelo de inteligencia artificial. Con los resultados obtenidos después de aplicar lo mencionado anteriormente, se evalúa cada modelo, teniendo en cuenta indicadores como el RMSE (Root Mean Squared Error), el error absoluto medio (MAE) y el coeficiente de determinación (R^2), permitiendo identificar el modelo definitivo para realizar las predicciones y validaciones posteriores. Entre los modelos a evaluar se encuentran las regresiones multivariadas (MLR), Elastic Net, Redes Neuronales (MLP Regressor) y Random forest. A continuación, se describe cada uno de los modelos propuestos, algunas de sus características, hiper parámetros implementados, entre otros:

3.4.1. Regresiones Lineales múltiples (MultiOutputRegressor):

El modelo de *regresión lineal múltiple (MLR)* propuesto se combina con la funcionalidad *Recursive Feature Elimination (RFE)* y bibliotecas de *Statsmodels* para optimizar la selección de variables y mejorar la interpretación estadística de los resultados. El enfoque sigue tres etapas principales.

Inicialmente, como se menciona en el artículo de Aguilar,F, el modelo entrena una regresión lineal múltiple básica; luego, se seleccionan las variables más relevantes mediante RFE con validación cruzada; y finalmente, el modelo se refina y analiza los datos en profundidad con statsmodels. [34]

La Regresión Lineal Múltiple permite modelar la relación entre una variable dependiente y múltiples variables independientes. La ecuación general que la describe es la siguiente:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \epsilon$$

donde β_0 es el intercepto, β_i son los coeficientes de regresión que representan el efecto de cada variable independiente en y , y ϵ es el error residual [34]. Para entrenar el modelo, en el código se implementa el scikit-learn de la siguiente manera:

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
```

La implementación de scikit-learn permitirá minimizar el error de los siguientes indicadores a determinar para evaluar la representatividad del modelo: Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de determinación R2.

Por otra parte, para mejorar la precisión del modelo y evitar el sobreajuste, se aplica la funcionalidad Recursive Feature Elimination (RFE) con validación cruzada teniendo en cuenta 3 particiones de los datos. Con este método se busca seleccionar las variables más relevantes y eliminar progresivamente aquellas que tienen menor impacto en la variable objetivo para la predicción.

Luego de seleccionar las variables, el modelo se reentrena utilizando **statsmodels** [35], permitiendo obtener un análisis más detallado de los coeficientes e intervalos de confianza. Los coeficientes en statsmodels se basan en la siguiente ecuación matricial:

$$\beta = (X^T X)^{-1} X^T y$$

donde X es la matriz de variables independientes y y es el vector de la variable dependiente [35]. Al final, se espera que el statsmodels muestre los valores p de las variables seleccionadas, lo que indica la significancia de cada variable en el modelo. Si un valor p es menor que 0.05, la variable es estadísticamente significativa. También se incluyen métricas como el estadístico F, que evalúa si el modelo en su conjunto es significativo, y criterios de selección de modelos como el Akaike Information Criterion (AIC) y el Bayesian Information Criterion (BIC), que penalizan la complejidad del modelo para evitar sobreajuste [35].

Finalmente, se calculan nuevamente los indicadores para evaluar el desempeño del modelo con las características optimizadas.

3.4.2. Elastic Net: model.ElasticNet

Este modelo combinara la regularización Lasso (L1) y Ridge (L2) para penalizar los coeficientes del modelo, controlando el sobreajuste y conservando interpretabilidad [36]. Adicionalmente, se utiliza validación cruzada (Cross-Validation, CV) y optimización de hiperparámetros con GridSearchCV para entrenar el modelo. La regresión Elastic Net añade términos de regularización L1 y L2 a la regresión lineal tradicional. La función utilizada es:

$$\min_{\beta} \left\{ \sum_{i=0}^n (\gamma_i - X_i^T \beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p x_j^2 \right\}$$

Donde:

- γ_i : Variable dependiente para la observación i
- X_i : Vector de características de la observación i
- β : Vector de coeficientes del modelo
- λ_1 : Parámetro de penalización L1 (Lasso) que controla la selección de características
- λ_2 : Parámetro de penalización L2 (Ridge) que controla la reducción de los coeficientes para evitar sobreajustes.

Adicionalmente el modelo Elastic Net incluye un parámetro α que controla la mezcla entre la penalización L1 y L2, y se define como:

$$\lambda_1 = \alpha \lambda \quad \text{y} \quad \lambda_2 = (1 - \alpha) \lambda$$

donde sí $\alpha=1$, el modelo se comporta como *Lasso*, si $\alpha=0$ se comporta como *Ridge*, y si $0<\alpha<1$, se combinan ambas regularizaciones.

Para la definición de los hiperparámetros, se implementa *GridSearchCV*, evaluando distintas combinaciones de α y λ . Los principales hiperparámetros que se optimizan son:

- **α** : Controla la mezcla entre Lasso y Ridge, *l1_ratio*.
- ***cv***: Número de folds en la validación cruzada.
- **Scoring**: métrica utilizada para evaluar los modelos, R^2 o MSE (Mean Squared Error).

Los hiperparámetros que ajustan el modelo son los siguientes:

- Controlar el balance entre Lasso (L1) y Ridge (L2): 0.1, 0.5, 0.7, 0.9
- Número de valores de alpha explorados internamente en *ElasticNetCV*: 10, 50, 100 (Más valores permiten una mejor búsqueda de regularización)
- Número máximo de iteraciones para convergencia: 100, 500, 1000 (Afecta el tiempo de entrenamiento y precisión)
- Número de folds para validación cruzada dentro de *ElasticNetCV*: 3, 6, 10 (Entre mayor número de folds mejora la estabilidad del modelo, pero aumenta el costo computacional).

3.4.3. Red Neuronal (MLP Regressor): *sklearn.neural_network - MLPRegressor*

En la investigación realizada por X. Glorot y Y. Bengio, mencionan que estos modelos de redes neuronales artificiales feedforward, a diferencia de la regresión lineal tradicional, que asumen una relación lineal entre las variables, las redes neuronales pueden modelar relaciones no lineales entre las características de entrada y la variable de salida [37]. Adicionalmente, mencionan que estos modelos consisten en capas de neuronas interconectadas, donde cada neurona aplica una transformación matemática a la información recibida y la transmite a la siguiente capa [37]. El entrenamiento del modelo se realiza ajustando los pesos y sesgos de la red mediante un algoritmo de optimización basado en descenso de gradiente.

En el modelo a desarrollar, se implementará una red neuronal utilizando la clase *MLPRegressor* de *scikit-learn*, la cual se encapsula en un *Pipeline* para aplicar una normalización robusta con *RobustScaler*. Se propone este tipo de escalado, con el fin de reducir el impacto de valores atípicos en los datos. Posteriormente, se emplea

GridSearchCV para optimizar los hiperparámetros clave del modelo, incluyendo el número de neuronas en las capas ocultas, la función de activación, el método de optimización y el número máximo de iteraciones. La configuración para el modelo es la siguiente:

Configuración de capas ocultas en la red neuronal:

- (50,): Una capa oculta con 50 neuronas
- (100,): Una capa oculta con 100 neuronas
- (50, 10): Dos capas ocultas con 50 y 10 neuronas
- (100, 50, 10): Tres capas ocultas con 100, 50 y 10 neuronas

Función de activación de las neuronas:

- Identity: Función lineal $f(x)=xf(x) = xf(x)=x$
- Tanh: Tangente hiperbólica, ayuda a la convergencia con valores entre -1 y 1.
- Relu: Rectified Linear Unit, ayuda a evitar el problema de gradiente desaparecido.

Algoritmo de optimización para el entrenamiento:

- Lbfgs: Método de optimización basado en aproximaciones de segundo orden (rápido en conjuntos pequeños)
- Adam: Optimizador adaptativo basado en momentos (eficiente en grandes volúmenes de datos)
- Número máximo de iteraciones para el entrenamiento: 300, 600 y 1000 iteraciones.

Optimización de hiperparámetros con GridSearchCV:

- Validación cruzada: 5 folds (cv=5)

3.4.4. Random Forest: RandomForestRegressor – RandomForestClassifier

Este modelo de aprendizaje supervisado basado en métodos de ensamble de árboles de decisión tiene como ventaja principal en su capacidad para reducir la varianza y mejorar la estabilidad de las predicciones mediante la combinación de múltiples árboles de decisión [37]. A diferencia de la implementación de un solo árbol de decisión, que puede ser propenso al sobreajuste, un grupo de múltiples árboles aleatorios promedia las predicciones de varios árboles independientes, lo que resulta en un modelo más robusto. Rodrigo, j. menciona que los modelos Random Forest Regressor obtienen

matemáticamente las predicciones con varias ecuaciones, un ejemplo de ellas se presenta a continuación [38]:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

Donde:

1. T: es el número total de árboles en el bosque,
2. $f_t(x)$: representa la predicción del árbol t para una entrada x
3. \hat{y} : es el valor final predicho por el modelo.

Este proceso de agregación de predicciones se basa en el concepto de **Bootstrap Aggregation (Bagging)**, donde cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento, permitiendo capturar diferentes patrones en los datos sin caer en el sobreajuste [38].

Adicionalmente, para optimizar el rendimiento del modelo, se utilizó **RandomizedSearchCV**, una técnica que permite explorar diferentes combinaciones de hiperparámetros de manera eficiente sin necesidad de evaluar todas las posibles combinaciones, como lo haría una búsqueda en cuadrícula (*Grid Search*). En este caso, se definieron los siguientes hiperparámetros clave:

- **n_estimators**: Número de árboles en el bosque. Se exploraron valores entre 50 y 400, con incrementos de 50.
- **max_depth**: Profundidad máxima de los árboles. Se probaron valores de 10, 20 y sin restricción (None).
- **min_samples_split**: Número mínimo de muestras requeridas para dividir un nodo. Se evaluaron valores de 2, 5 y 10.

Finalmente, la validación cruzada de 3 pliegues (cv=3), permitirá seleccionar la mejor combinación de hiperparámetros, evaluando los resultados con la métrica R^2 .

3.4.5. Random Forest Multi-Output

Tras el análisis de los resultados de los 4 modelos descritos anteriormente, se procede a realizar la profundización en uno de ellos, mejorando sus características y con el fin de que los resultados de las predicciones realizadas sean más precisos.

Partiendo de lo anterior y del modelo Random Forest descrito, se construye un modelo de Random Forest Regressor, optimizado mediante GridSearchCV, para la predicción de dos variables objetivo: la producción acumulada de petróleo por ciclo (Np) y la razón de inyección-producción (PIR).

Como se mencionó en el numeral 3.4.4. este tipo de modelos se basan en la combinación de múltiples árboles de decisión entrenados con subconjuntos aleatorios de los datos, lo que mejora la generalización y robustez del modelo, de igual forma se fundamenta en la misma base matemática. Con la diferencia de que, para manejar múltiples salidas, en este caso se emplea la estrategia MultiOutputRegressor, que permite entrenar un modelo independiente de Random Forest para cada variable objetivo.

Para mejorar el desempeño del modelo, se optimizan los hiperparámetros mediante GridSearchCV, que busca la mejor combinación de parámetros evaluando múltiples configuraciones mediante validación cruzada. Los hiperparámetros ajustados son:

- **n_estimators:** Número de árboles en el bosque (100 y 200).
- **max_depth:** Profundidad máxima de cada árbol (10, 20 y sin restricción).
- **min_samples_split:** Número mínimo de muestras requeridas para dividir un nodo (2 y 5).
- **min_samples_leaf:** Número mínimo de muestras en una hoja (1 y 2)

Una vez encontrados los mejores hiperparámetros, el modelo se entrena utilizando MultiOutputRegressor, lo que permite manejar múltiples variables objetivo de manera independiente.

3.4.6. Comparación de Modelos Construidos

En términos de interpretabilidad, el modelo de Regresiones Lineales Múltiples y Elastic Net destacan por su facilidad de análisis. Ambos modelos permiten entender la contribución de cada variable predictora, aunque Elastic Net es más robusto frente a la multicolinealidad gracias a su combinación de regularización L1 y L2. Sin embargo, estas técnicas tienen limitaciones al modelar relaciones no lineales, lo que las hace menos efectivas en problemas complejos con interacciones no lineales entre variables.

Por otro lado, modelos como Red Neuronal Artificial (MLPRegressor) y Random Forest ofrecen una mayor capacidad de generalización y flexibilidad para capturar patrones no

lineales en los datos. La red neuronal es especialmente poderosa en problemas de alta complejidad, aunque su interpretabilidad es baja y requiere un mayor costo computacional. Random Forest, en contraste, es más interpretable gracias a su capacidad de medir la importancia de las características, pero su desempeño depende de la correcta optimización de hiperparámetros.

Finalmente, el Random Forest Multi-Output representa una alternativa ideal cuando se requiere predecir múltiples variables objetivo simultáneamente. Su capacidad de manejar relaciones no lineales y su robustez ante datos ruidosos lo hacen un modelo versátil. Sin embargo, su principal desventaja es la alta demanda computacional y la dificultad para interpretar sus decisiones, ya que la combinación de múltiples árboles de decisión dificulta la trazabilidad de sus predicciones.

3.5. FASE 5. Validación del modelo con información real del área

Tras la construcción de los 5 modelos mencionados en la fase anterior, cada uno de estos aplicaba diferentes metodologías para mejorar los resultados a generar, no obstante, se calcularon indicadores de evaluación el error medio cuadrado (RMSE), el coeficiente de determinación (R^2) y el error medio absoluto (MAE), para medir la precisión y representatividad del modelo en las predicciones de la producción acumulada de petróleo por ciclo. A continuación, se describe detalladamente la forma en que cada uno de estos modelos fue validado.

3.5.1. Regresiones Lineales múltiples

Para la validación de este modelo, se utilizó la técnica de eliminación recursiva de características (RFE) junto con métricas estadísticas mencionadas anteriormente. El uso de Recursive Feature Elimination (RFE) permitió que el modelo seleccionara las variables más relevantes, eliminando de manera secuencial aquellas que aportaban menos al desempeño del modelo. Se evaluó la combinación óptima de características con base en el coeficiente de determinación R^2 .

Adicionalmente, la validación con statsmodels, se incluyó el análisis de los valores-p asociados a los coeficientes del modelo, considerando que aquellos con valores menores a 0.05 eran estadísticamente significativos. También se calcularon el Akaike

Information Criterion (AIC) y el Bayesian Information Criterion (BIC), donde valores más bajos indicaban un mejor ajuste del modelo.

3.5.2. Elastic Net: *model.ElasticNet*

Para validar este modelo, se utilizó validación cruzada k-fold con $k=5$, lo que permitió dividir los datos en cinco subconjuntos, utilizando cuatro para entrenamiento y uno para validación en cada iteración. Este procedimiento se repitió para cada combinación de hiperparámetros evaluados en GridSearchCV, seleccionando la mejor configuración con base en el mejor R^2 promedio en la validación cruzada.

Los hiperparámetros ajustados fueron el α , que controla la penalización, y el L1_ratio, que define la proporción de regularización entre L1 y L2. Finalmente, el modelo fue evaluado en el conjunto de prueba mediante métricas como R^2 para medir la capacidad explicativa del modelo, RMSE para evaluar la dispersión de los errores de predicción y MAE para determinar el error absoluto medio entre las predicciones y los valores reales.

3.5.3. Red Neuronal (MLP Regressor): *sklearn.neural_network – MLPRegressor*

La validación de la red neuronal artificial se realizó utilizando GridSearchCV en combinación con validación cruzada k-fold con $k=3$. En este proceso, se exploraron diferentes combinaciones de hiperparámetros, incluyendo el número de neuronas por capa, número de capas ocultas, tasa de aprendizaje y funciones de activación.

Para asegurar la estabilidad del modelo, se monitoreó la función de costo durante el entrenamiento y se evitaron problemas como el sobreajuste. Las métricas utilizadas para la evaluación fueron el R^2 ponderado, que mide el ajuste del modelo considerando la varianza de los datos, así como RMSE y MAE.

3.5.4. Random Forest: *RandomForestRegressor – RandomForestClassifier*

En este caso, la validación se llevó a cabo utilizando RandomizedSearchCV, una técnica que permite explorar de manera aleatoria un conjunto amplio de hiperparámetros sin necesidad de evaluar todas las combinaciones posibles, reduciendo así el tiempo de respuesta.

Los hiperparámetros evaluados fueron el número de árboles ($n_estimators$), la profundidad máxima de los árboles (max_depth) y el número mínimo de muestras por

nodo (`min_samples_split`). Para evaluar la robustez del modelo, se utilizó validación cruzada con $k=3$, tomando el mejor conjunto de hiperparámetros con base en el coeficiente de determinación R^2 .

Finalmente, la evaluación en el conjunto de prueba se realizó mediante el R^2 , que mide la capacidad del modelo de generalizar a datos no vistos, junto con métricas de error como RMSE y MAE.

3.5.5. *Random Forest Multi-Output*

Para validar este modelo, se utilizó `GridSearchCV` con validación cruzada $k=3$, explorando diferentes valores para los hiperparámetros, tales como `n_estimators` (cantidad de árboles en el bosque), `max_depth` (profundidad de los árboles) y `min_samples_split` y `min_samples_leaf`.

Dado que este modelo tenía dos salidas simultáneas (Np y PIR), la evaluación se realizó de manera individual y global. Se calcularon métricas de error para cada salida, incluyendo RMSE por variable, R^2 individual y R^2 ponderado, lo que permitió evaluar el desempeño en ambas predicciones de manera balanceada.

4. RESULTADOS Y ANALISIS

En esta sección se presentan los resultados obtenidos en las fases de la metodología, destacando los parámetros y variables más relevantes del proceso para la construcción del modelo de predicción, el entrenamiento y la validación de este. En el desarrollo de la metodología anteriormente descrita, se identificó que en la medida en que se desarrollaban todas las fases, era necesario iterar el proceso, haciendo de esta una metodología adaptativa.

4.1. Parametrización de las variables del proceso

En primer lugar, se recopiló y procesó la información disponible para llevar a cabo un análisis estadístico detallado inicial de los datos provenientes de los pozos en el campo de estudio. El análisis se realizó a través de un código en Python desarrollado en la herramienta Jupyter Notebook logrando realizar el control de calidad de los datos.

Tras el análisis inicial, se detectaron variables con altos niveles de dispersión. Para comprender esta situación, se realizó un análisis detallado de los pozos y los ciclos asociados a los datos anómalos, lo que permitió identificar las siguientes causas principales:

- Pozos en los que se completó la etapa de inyección, pero debido a fallas en su operación, no lograron finalizar la etapa de producción.
- Analizando el comportamiento de los pozos y las variaciones mensuales en la producción de aceite y agua, se identificaron ciclos de inyección que no estaban registrados en la base de datos original o cuya información era incompleta.
- Ciclos de producción prolongados (superiores a doce meses) que, según la información histórica y los comportamientos iniciales de los pozos, corresponden a periodos en los que el vapor inyectado dejó de influir en la producción, resultando en una etapa de producción en frío.
- Dado que la base de datos exportada estaba actualizada hasta septiembre de 2024, se observó que algunos pozos habían sido inyectados recientemente y, hasta ese momento, solo presentaban resultados parciales del ciclo en curso.

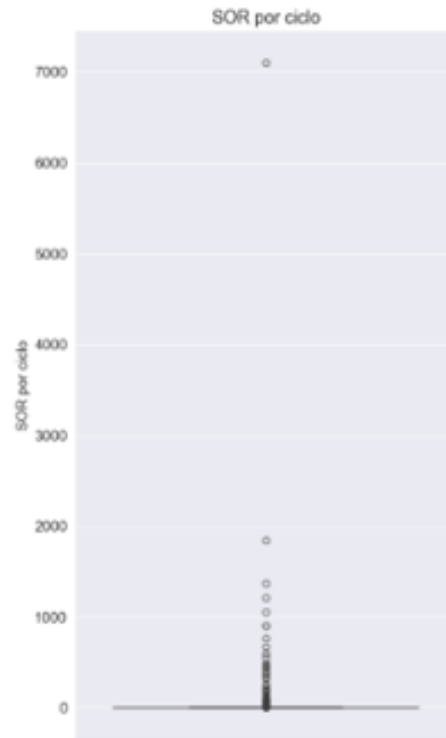
Para la explicación del proceso de depuración de la información, se toma como ejemplo la variable SOR. En la **Figura 16** se muestra la distribución estadística de la variable SOR antes de realizar cualquier filtro, en el cual se identifican datos anómalos que

generan una desviación muy alta. Allí se identifica que, aunque el 75% de los datos se encuentran por debajo de valores de 3.3, la media es de 7. La causa es que existen datos que se encuentra fuera del rango y específicamente el dato máximo que corresponde a 7106.

Figura 16.

Distribución estadística del SOR y Diagrama de caja de la data inicial sin filtros

SOR por ciclo	
mean	7.079
std	103.366
min	0.006
25%	0.852
50%	1.681
75%	3.381
max	7106.472

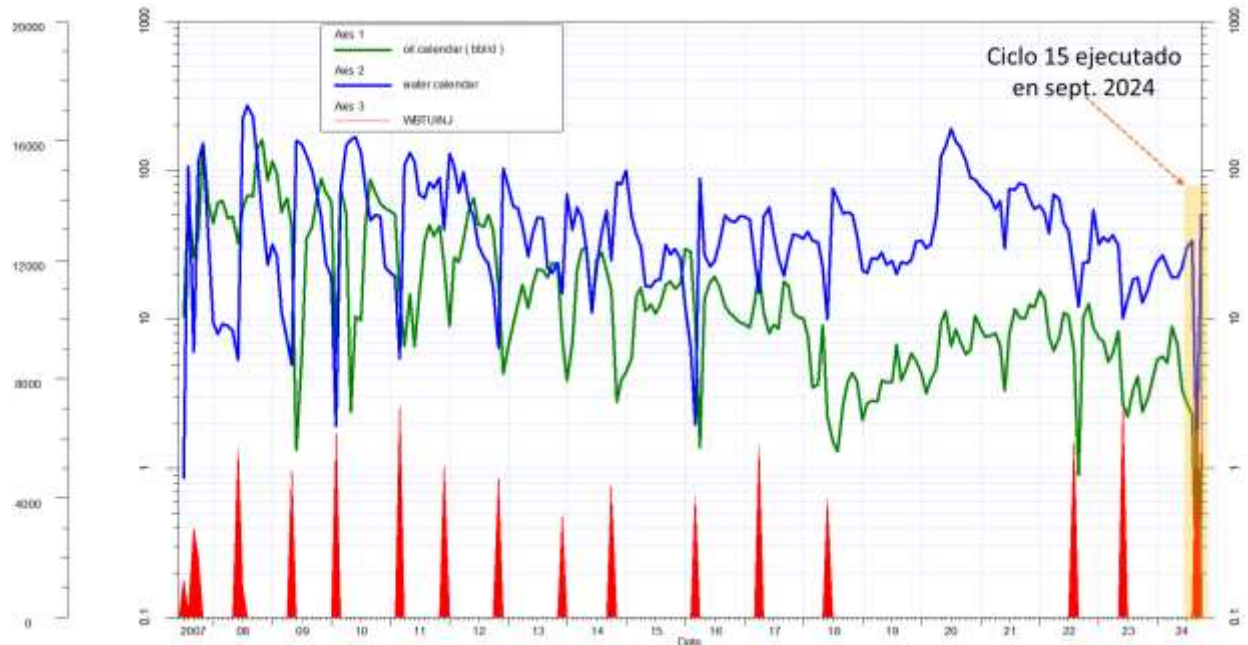


Nota. El diagrama muestra la distribución estadística de la variable SOR en la cual se identifican datos que se encuentran alejados del rango.

A partir de estos resultados preliminares, se identificó cual era el pozo que tenía este valor. En este caso correspondía a un pozo que su último ciclo de inyección había sido en septiembre del 2024 y durante ese mes solo había producido un día, como se muestra en la Figura 17.

Figura 17.

Comportamiento histórico del pozo ejemplo que presentaba SOR anómalo



Nota. La grafica de producción del pozo ejemplo muestra la causa por la cual se tenía un SOR fuera del rango.

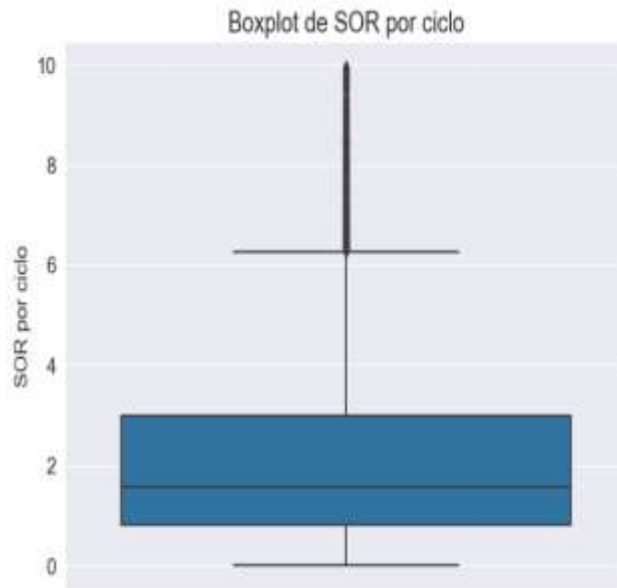
De la misma forma, se identificaron para las múltiples variables los datos que se encontraban por fuera del rango y se realizaron filtros para disminuir la dispersión de estas, teniendo presente las causas por las cuales los datos eran atípicos.

La **Figura 18** muestra la distribución estadística de la variable SOR una vez realizados los diferentes filtros para eliminar la data que presentaba gran incertidumbre y que generaba alta desviación en los cálculos. Se puede evidenciar que aún continúan presentándose variables fuera del rango, sin embargo, la desviación disminuye de manera considerable en comparación con la data inicial.

Figura 18.

Distribución estadística del SOR y Diagrama de caja de la data filtrada

SOR por ciclo	
mean	2.212
std	1.934
min	0.006
25%	0.820
50%	1.572
75%	3.000
max	9.960

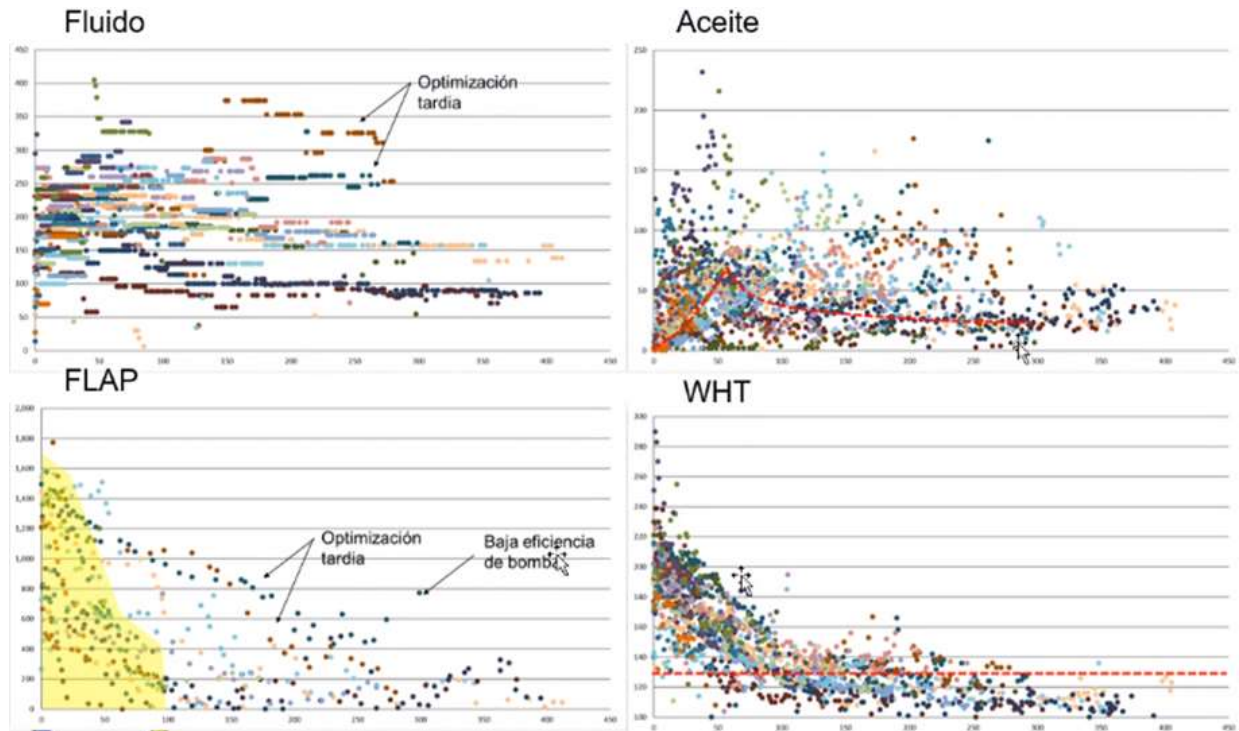


Nota. El diagrama muestra la distribución estadística de la variable SOR luego de filtrar la data que generaba alta dispersión.

Así mismo, en el procesamiento de los datos de producción, se identificaron algunos periodos en donde se realizaron trabajos tardíos de optimización y posibles inconvenientes mecánicos como daño de casing, influjo de los acuíferos por detrás del revestimiento debido al desgaste del cemento como consecuencia de los altos cambios de temperatura el que estaban expuestos por las inyecciones de vapor, que influyeron en el rendimiento de la producción de aceite. Algunos análisis y tendencias de como los distintos parámetros operativos afectan el aceite acumulado por ciclo de inyección se presentan a en la **Figura 19**.

Figura 19.

Efecto de las variables operativas en la eficiencia del proceso de inyección cíclica de vapor



Nota. La figura muestra como las optimizaciones operativas influyen en la eficiencia del proceso de inyección cíclica de vapor.

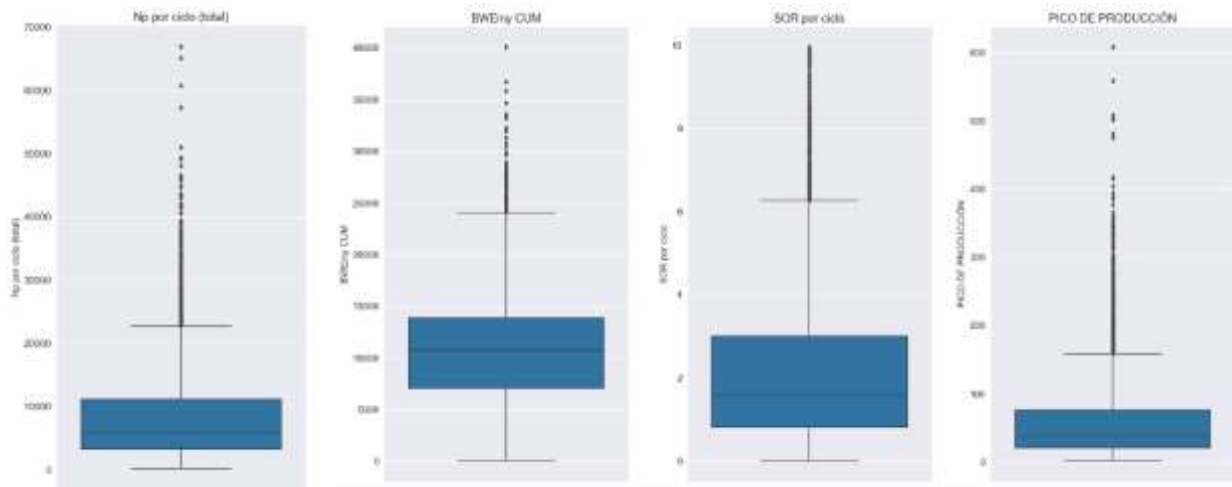
A partir de esta información se identificaron algunos comportamientos anómalos por causas externas al proceso de inyección de vapor y que aumentaban la dispersión de los datos. Cabe resaltar que la información operativa diaria era limitada, por lo cual no se logró realizar un análisis detallado el cual se pudiera incorporar dentro del modelo.

Aunque la producción previa al primer ciclo de inyección es importante para conocer las tendencias y declinaciones de la producción en frío, el objetivo de este estudio es predecir el aceite a recuperar por un nuevo ciclo de inyección, basado en esto, durante la producción en frío, al no estar asociado a un vapor inyectado, muchas variables calculadas no eran posibles de calcular durante estos periodos, por lo cual, también fue necesario filtrar esta producción en frío de la base de datos de trabajo. Así mismo, se generaron unas variables calculadas análogas a los acumulados tanto de inyección como de producción (aceite acumulado, agua producida acumulada, vapor inyectado acumulado), pero que únicamente tienen en cuenta los periodos en los que la producción de aceite y agua estuviera influenciada por la inyección de vapor.

Posteriormente se volvieron a generar los análisis descriptivos y estadísticos. La **Figura 20** muestra el ejemplo de algunas variables utilizadas en el estudio luego de la aplicación de filtros para eliminar datos anómalos. Cabe resaltar que, con estos resultados, se siguen observando alguna dispersión en los datos.

Figura 20.

Diagramas de caja y bigotes de las variables filtradas utilizadas



Nota. La figura muestra rangos estadísticos de algunas de las variables analizadas dentro del estudio luego de la aplicación de los diferentes filtros.

A partir del análisis de los datos históricos y filtros, se construyó un conjunto de estadísticas descriptivas que permitió caracterizar cada variable en términos de sus valores máximos, mínimos, promedio aritmético, percentiles (25%, 50% y 75%) y desviación estándar. A continuación, se presentan los resultados estadísticos en la **Tabla 6**, con los rangos para cada variable en los pozos seleccionados.

Tabla 6.*Rangos operativos de las variables de estudio*

Variable	mean	std	min	Percentil			max
				25%	50%	75%	
OIL CUM bbl	107,970.7	70,072.1	1,322.0	50,136.8	100,278.7	152,978.1	339,818.1
WAT CUM bbl	165,587.1	134,718.4	177.0	57,394.5	137,808.0	244,840.0	1,341,324.0
VAP CUM MMBTU	39,584.9	25,878.1	420.0	18,103.5	36,220.3	57,917.5	143,590.9
WAT INY EQ CUM bbl	106,879.3	69,870.9	1,134.0	48,879.5	97,794.8	156,377.3	387,695.5
CICLO CORREGIDO WINJBTU	16.0	5.2	1.0	13.0	16.0	19.0	27.0
qi	28.6	26.8	0.0	10.9	20.5	38.5	326.7
Np por ciclo (total)	8,539.3	7,595.3	166.6	3,325.7	5,883.0	11,101.4	66,974.9
SOR cum	1.1	0.7	0.0	0.7	1.0	1.3	6.9
PIR cum	2.7	1.5	0.2	2.0	2.4	3.1	34.6
MMBTU Ciclo	4,577.8	1,853.0	55.3	3,293.4	4,460.0	5,556.3	16,789.0
Wp por ciclo	18,169.6	12,173.9	118.2	10,134.6	16,109.1	23,342.1	148,578.4
BWEiny CUM	10,774.2	5,369.8	64.8	7,092.9	10,783.9	13,898.8	40,105.8
Pico de producción	59.6	30.2	10.0	21.2	38.5	76.2	608.2
SOR por ciclo	2.2	1.9	0.0	0.8	1.6	3.0	10.0
DURACIÓN MESES CICLO	9.4	2.7	2.0	7.0	10.0	12.0	12.0

Nota. La tabla muestra los rangos de las principales variables a estudiar dentro del proceso de inyección cíclica de vapor.

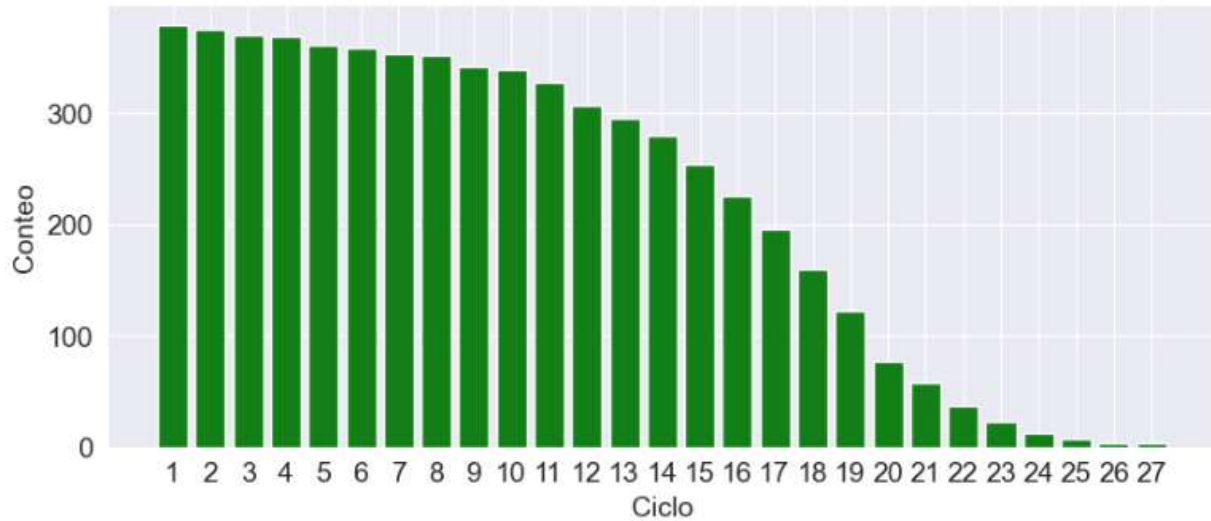
El proceso de depuración de los datos mediante el código desarrollado en Python resultó fundamental para garantizar la calidad de la información analizada. Se eliminaron los registros que generaban una dispersión muy alta y a los cuales se le encontraba una causa externa al comportamiento asociado a la inyección cíclica de vapor.

No obstante, la dispersión de las variables continuaba siendo alta. Basados en estos resultados, se identificó que debido a que a medida que se realizan mayor número de ciclos de inyección de vapor, la eficiencia del proceso comienza a disminuir, por lo cual, una forma de disminuir la dispersión, analizar cada una de las variables dependiendo del ciclo de inyección en el que se encontraba el pozo.

Inicialmente se identificó el número de pozos que se encontraban en cada ciclo de inyección cíclica de vapor luego de la depuración inicial. En la **Figura 21** se observa el número de pozos por cada ciclo, en el cual se puede evidenciar que en los últimos ciclos de inyección la cantidad de pozos es muy baja, lo que puede generar una disminución en la confiabilidad de los rangos estadísticos que se puedan obtener en dichos ciclos.

Figura 21.

Cantidad de pozos por ciclo de inyección



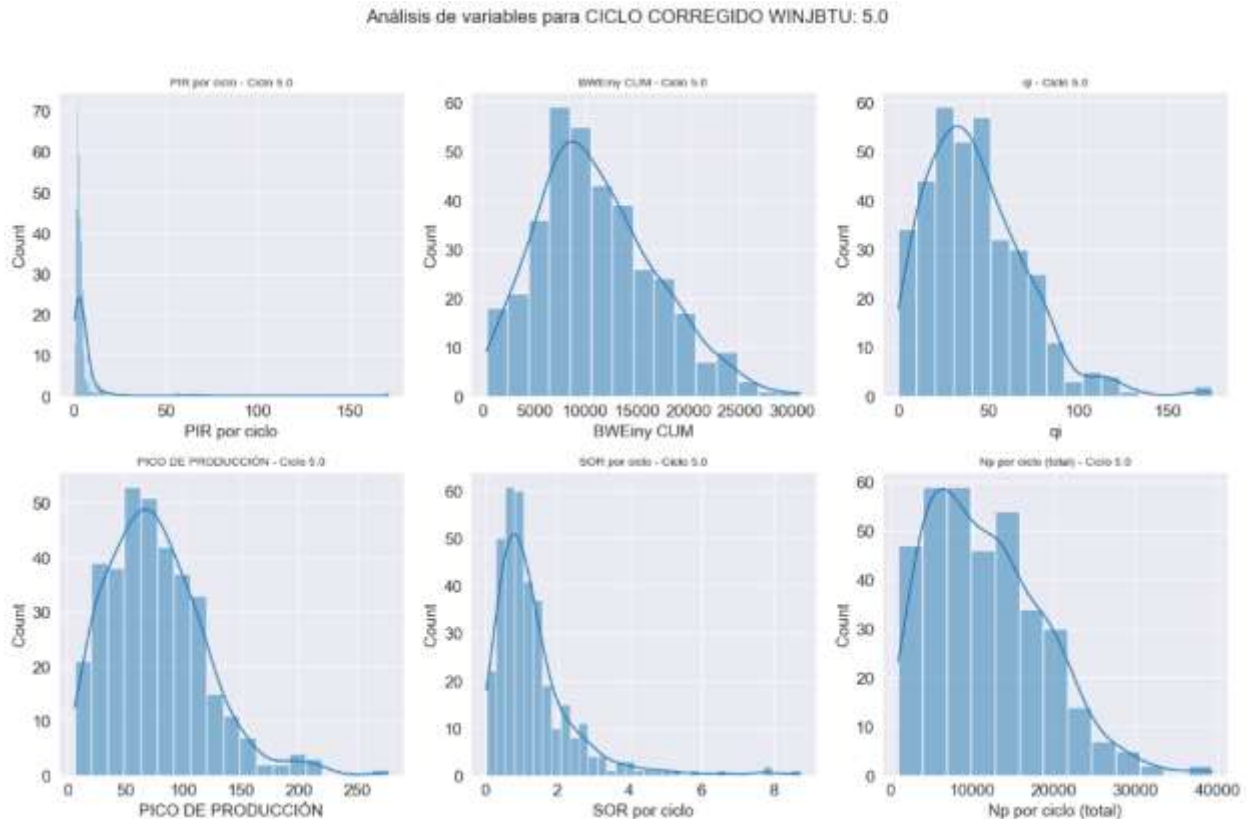
Nota. La gráfica muestra la cantidad de pozos que se encuentran en cada uno de los ciclos de inyección de vapor.

De esta forma se evidenció que desde el ciclo 20, el número de pozos para realizar la estadística del ciclo es menor a 100 pozos y finalmente llegando en los ciclos 26 y 27 a únicamente 3 pozos que cuentan con este ciclo, lo cual puede generar una disminución en la confiabilidad de los resultados provenientes de estos últimos ciclos.

Posteriormente, se realizó todo el análisis descriptivo a las variables en cada uno de los ciclos. La **Figura 22** es un ejemplo del análisis estadístico de las distintas variables para el ciclo de inyección número 5.

Figura 22.

Histograma de las variables durante el ciclo 5 de inyección cíclica de vapor

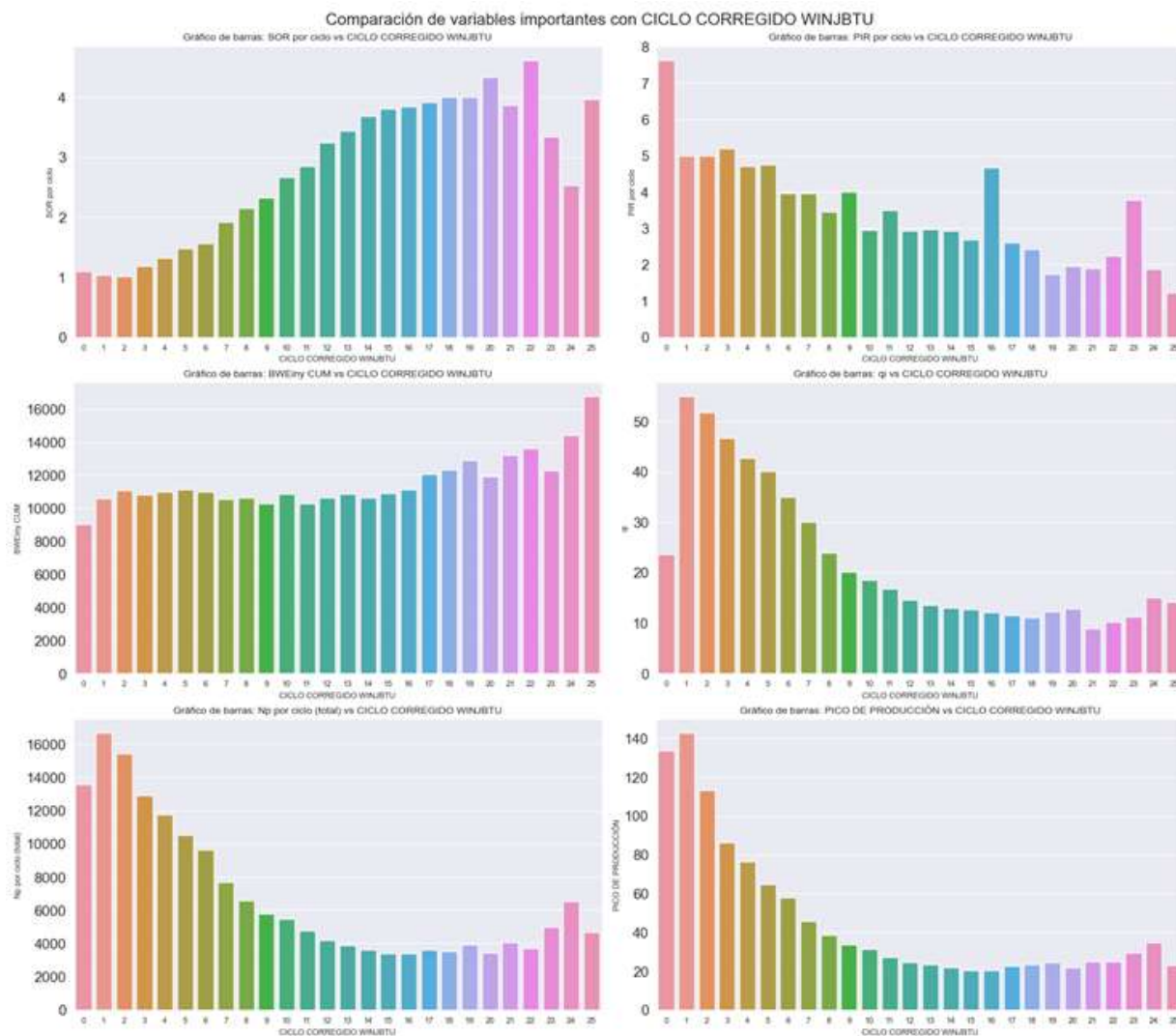


Nota. Las gráficas muestran la distribución probabilística de las variables ejemplo para el ciclo cinco de inyección cíclica de vapor en los pozos del Campo de estudio.

Adicionalmente, se realizó el análisis de cómo era el cambio del comportamiento de las distintas variables a medida que aumentaban los ciclos de inyección. En la **Figura 23** se observa la media de las distintas variables en la cual se puede observar la tendencia a medida que se avanza en el ciclo de inyección. Esta información corrobora la disminución de eficiencia a medida que incrementan los ciclos de inyección, sin embargo, hacia los últimos ciclos, la tendencia cambia su comportamiento, lo cual está relacionado al bajo número de la muestra que tienen estos ciclos y a lo cual se puede deducir que los pozos con mayor número de ciclos han sido históricamente los mejores pozos del campo y debido a esto continúan con eficiencias altas hacia los últimos ciclos de inyección, sin embargo, no es una muestra representativa.

Figura 23.

Comportamiento promedio de las variables en cada ciclo de inyección.



Nota. Las gráficas muestran los datos promedio de las diferentes variables durante cada ciclo de inyección, en el cual se observan las tendencias de disminución de eficiencia a medida que aumentan los ciclos.

Cabe resaltar que a medida que se avanzó en las fases 3 y 4 de la metodología, se encontraban variables fuera de los rangos, y a partir de estos hallazgos se realizaban las correspondientes depuraciones con el fin de mejorar la calidad de los datos teniendo como premisa la respectiva justificación para su filtración, siendo de esta forma un proceso cíclico.

4.2. Selección de las variables con mayor impacto en la producción de aceite en el campo de estudio

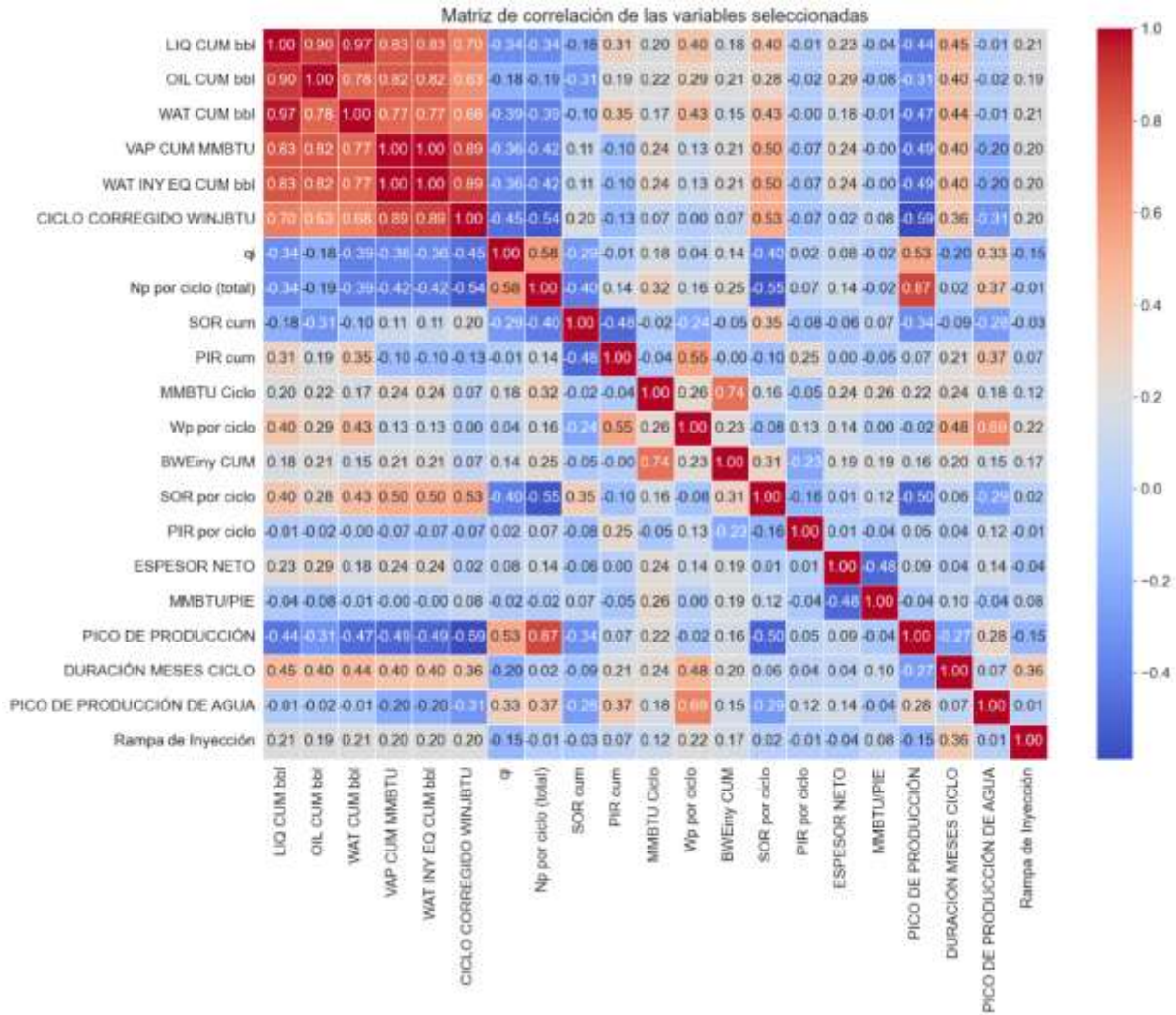
En la implementación de mapas de calor a las variables iniciales exportadas de OFM, se identificaron correlaciones fuertes, correlaciones bajas y correlaciones negativas relevantes de cada variable por ciclo. Teniendo en cuenta que los valores cercanos a +1 indican una correlación positiva fuerte (ambas variables aumentan juntas), los valores cercanos a -1 indican una correlación negativa fuerte (una variable aumenta mientras la otra disminuye) y los valores cercanos a 0 indican que no hay correlación lineal, permitirán definir cuales variables tienen mayor impacto en la producción de aceite.

Con base a lo anterior, los resultados obtenidos del primer mapa de calor realizado a las variables iniciales evidenciaron la alta dispersión de los datos de los pozos, debido a la calidad deficiente de los mismos, la presencia de registros erróneos y la falta de datos. Estos resultados se muestran en la **Figura 24**.

Para profundizar en este análisis, y considerando que la variable objetivo es la producción de petróleo acumulada por ciclo, se desarrollaron diagramas de dispersión. Estos permitieron identificar que la relación entre la mayoría de las variables y la totalidad de los ciclos presentaba valores entre 0.3 y 0.5, lo que indica una correlación baja y poco representativa para la selección de variables (**Figura 25**).

Figura 24.

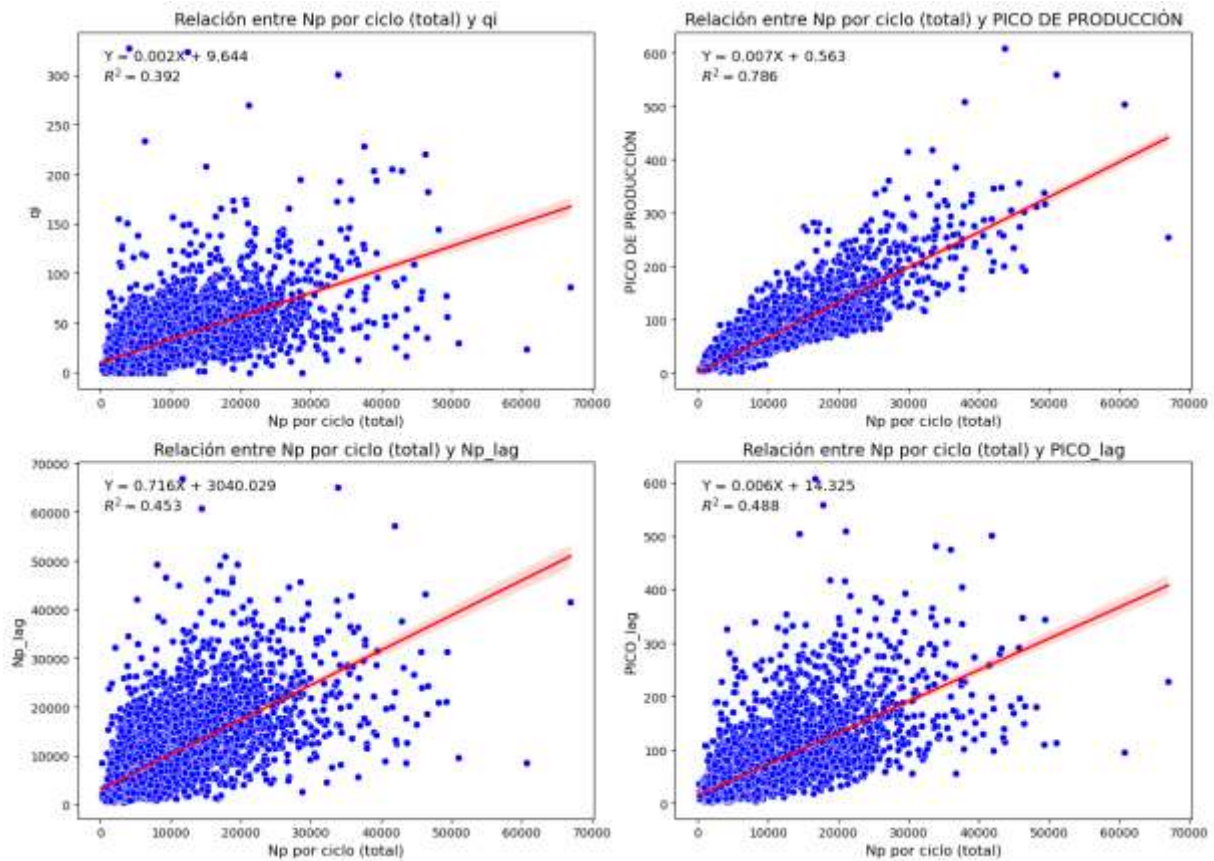
Mapa de calor con la data inicial



Nota: Se evidencia los resultados del mapa de calor inicial, donde las relaciones entre variables eran bajas y poco representativas.

Figura 25.

Diagramas de dispersión de la data inicial

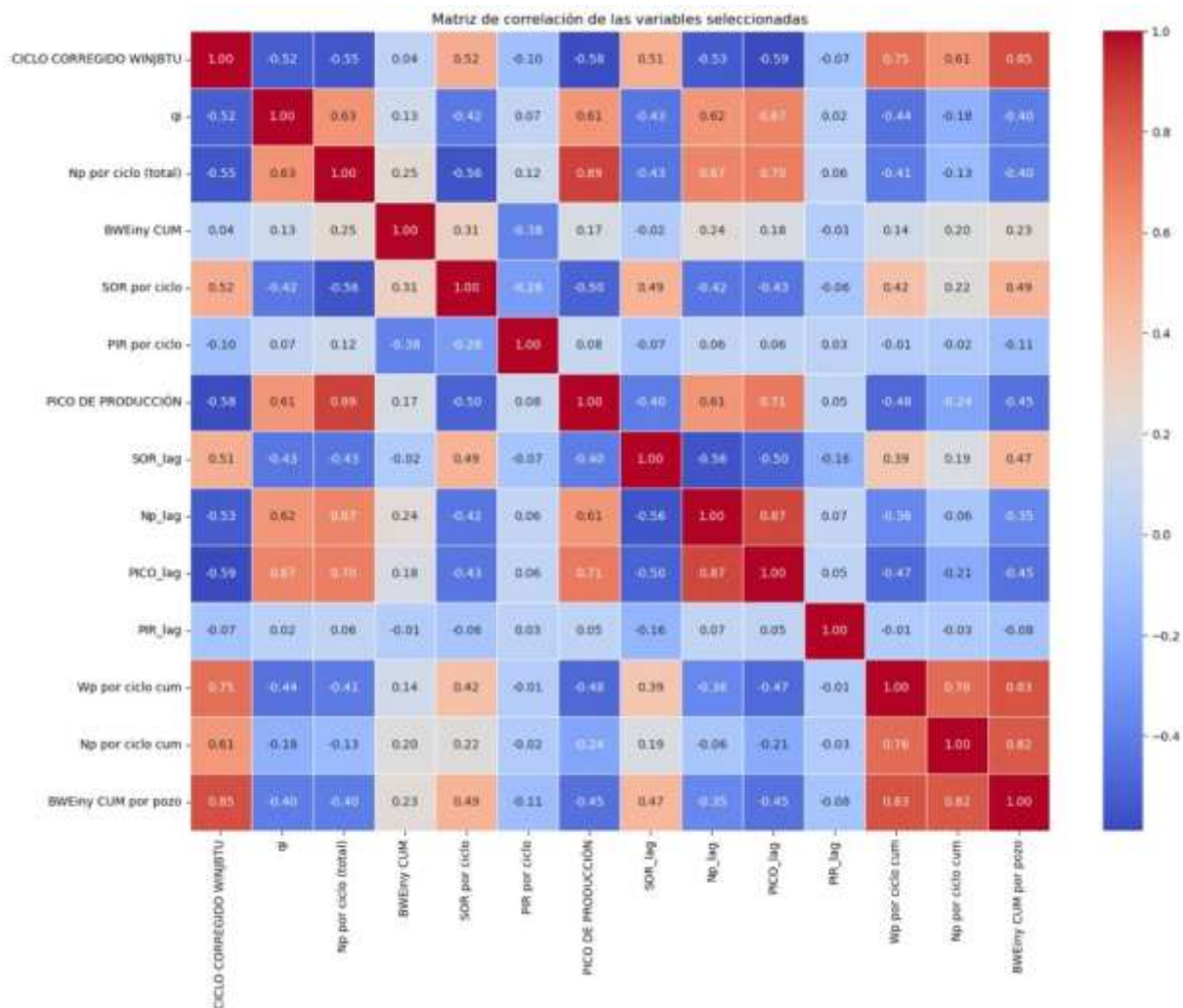


Nota: Se evidencia alta dispersión entre la variable objetivo (Np) y las demás variables.

Una vez realizados los filtros de la data que se encontraba fuera del rango explicada en la **Sección 4.1** y la creación de nuevas variables calculadas asociadas únicamente a los tiempos en los que se tenía efecto de la inyección de vapor en la producción de aceite, se incorporaron nuevas variables las cuales hacen referencia al ciclo anterior y se identificaron con la nomenclatura "*_lag*". Estas variables sirven como data de entrada para predecir el acumulado de producción de aceite del siguiente ciclo. El uso de estas variables dentro del modelo se explicará en mayor detalle en la **Sección 4.3** del presente documento. El desarrollo de estos procesos para filtrar la data anómala permitió capturar la tendencia de cada variable con respecto a la producción de aceite y mejorar su correlación dentro del análisis. El mapa de calor obtenido con esas variables se presenta en la **Figura 26**.

Figura 26.

Mapa de calor de las variables depuradas incluyendo las variables calculadas



Nota: El mapa de calor muestra la correlación que hay entre las variables del proceso, luego de ser depuradas y procesadas.

Teniendo en cuenta que el objetivo del estudio es la producción incremental asociada a la inyección cíclica de vapor “Np por ciclo (total)”, del mapa de calor anterior, se puede analizar lo siguiente:

- **Correlaciones Positivas relevantes**

El caudal inicial, "q_i" (0.63), muestra una correlación positiva moderada con el “Np por ciclo (total)”, lo que sugiere que los pozos con un alto caudal inicial tienden a mantener un buen desempeño a lo largo del ciclo.

En cuanto al "PICO DE PRODUCCIÓN" (0.89), se presenta una alta relación con el "Np por ciclo (total)", indicando que los ciclos con mayores picos de producción tienden a generar mayores volúmenes acumulados de petróleo.

Por parte de la producción acumulada en el ciclo anterior y el Pico de producción obtenido en el ciclo anterior, "Np_lag" y "PICO_lag" (0.67 y 0.70), se presenta una fuerte correlación con el "Np por ciclo (total)", lo que sugiere que el desempeño de un ciclo anterior influye directamente en el siguiente.

- **Correlaciones Negativas Significantes**

Para las variables "CICLOS CORREGIDO WINJBTU" (-0.55) y "SOR por ciclo" (-0.56), se identificó que, con el avance de los ciclos, la producción de petróleo tiende a disminuir. Estos resultados tienen coherencia con la teoría de la inyección cíclica de vapor, la cual indica que a medida que aumentan los ciclos de inyección se presenta una disminución de la eficiencia del proceso, ya que disminuye la saturación de aceite en el área cercana al pozo y adicionalmente se presenta una declinación natural del yacimiento.

- **Correlaciones asociadas al ciclo de inyección**

Teniendo en cuenta que todos los análisis tienen en común que a medida que aumentan los ciclos de inyección la eficiencia durante dicho ciclo tiende a disminuir con respecto al ciclo anterior. Una forma en la cual se podría aumentar el efecto de los ciclos de inyección (CICLOS CORREGIDO WINJBTU) con respecto al acumulado de producción durante el ciclo (Np por ciclo (total)), fue incorporando los acumulados totales de producción de aceite, agua e inyección de vapor, los cuales están directamente relacionados con el ciclo de inyección, en cuanto, a medida que el ciclo de inyección aumenta, los acumulados totales de producción e inyección también lo hacen.

A partir de los resultados obtenidos de la fase 2 y la fase 3 del presente estudio se continúa con la fase 4 que consiste en la construcción del modelo predictivo.

4.3. Construcción del modelo basado en Inteligencia Artificial para estimar la producción de aceite asociada a ICV

Una vez seleccionadas las variables con mayor impacto en el acumulado de producción de aceite y las bases de datos procesadas, se realizaron algunos modelos para identificar el grado de ajuste que presentaban. Inicialmente se construyeron los siguientes modelos:

- Regresión lineal múltiple
- Elastic net
- Red neuronal (MLP Regressor)
- Random Forest

La **Figura 27** muestra como ejemplo el código utilizado para la construcción del modelo de Elastic Net.

Figura 27.

Modelo de predicción Elastic Net

```
print("\n=== 3) ELASTIC NET ===")

# 3.1: Definir el modelo base para la búsqueda
cv_model_en = ElasticNetCV()

# 3.2: Hiperparámetros a explorar
param_grid_en = {
    'l1_ratio': [0.1, 0.5, 0.7, 0.9],
    'n_alphas': [10, 50, 100],
    'max_iter': [100, 500, 1000],
    'cv': [3, 6, 10] # Cross-validation interna del ElasticNetCV
}

# 3.3: GridSearch
grid_search_en = GridSearchCV(
    cv_model_en,
    param_grid_en,
    cv=5, # CV externo para la búsqueda
    error_score='raise',
    scoring='r2'
)
grid_search_en.fit(X_train, y_train)

best_params_en = grid_search_en.best_params_
best_model_en = grid_search_en.best_estimator_

print("Best Hyperparameters (ElasticNet):", best_params_en)
print("Best Model (ElasticNet):", best_model_en)

# 3.4: Entrenar la mejor combinación final
# (Ya lo hace 'fit', pero si quieres instanciar manualmente:)
elastic_final = ElasticNet(
    l1_ratio=best_model_en.l1_ratio_,
    alpha=best_model_en.alpha_,
    max_iter=best_model_en.max_iter
)
elastic_final.fit(X_train, y_train)

# 3.5: Métricas en test
y_pred_en = elastic_final.predict(X_test)

r2_en = r2_score(y_test, y_pred_en)
mae_en = mean_absolute_error(y_test, y_pred_en)
rmse_en = sqrt(mean_squared_error(y_test, y_pred_en))

print("\n>>> MÉTRICAS ELASTIC NET (TEST) <<<")
print("R²:", r2_en)
print("MAE:", mae_en)
print("RMSE:", rmse_en)

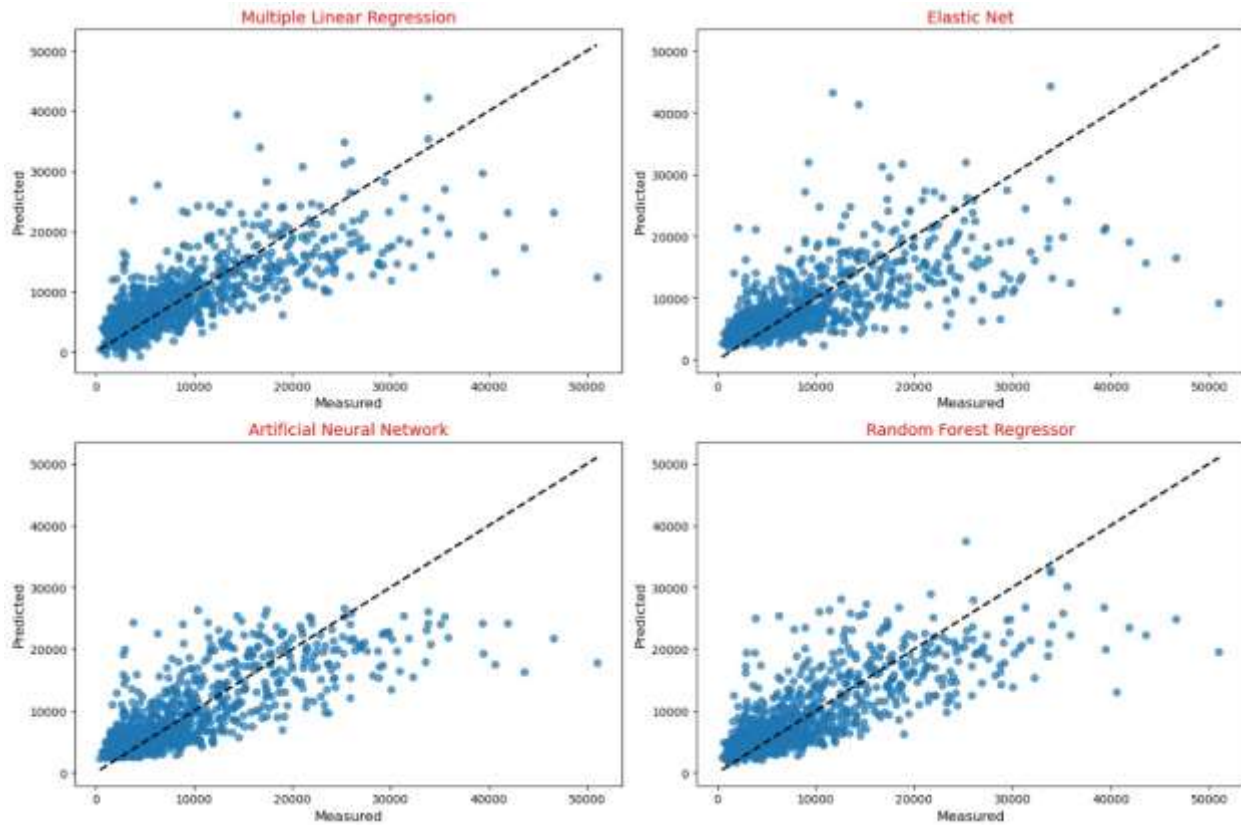
# Guardar predicción
prediction['Elastic Net'] = y_pred_en
```

Nota. La figura muestra el código utilizado para la construcción del modelo Elastic Net en la herramienta Jupyter.

La **Figura 28** muestra los resultados de las predicciones obtenidas en cada modelo mencionado, a partir de estas gráficas se puede validar el grado de dispersión de cada modelo de forma visual.

Figura 28.

Comparación de distintos modelos de predicción



Nota. La figura muestra los resultados obtenidos para los distintos modelos de predicción en los cuales se puede identificar visualmente que tan alta es la predicción que están realizando.

Para cada uno de estos modelos se calculó el R^2 (coeficiente de determinación), el MAE (Error Absoluto Medio) y el RMSE (Raíz del Error Cuadrático Medio). A partir de estos resultados se puede deducir que el modelo que más estaba explicando el proceso de inyección cíclica de vapor era la red neuronal con un R^2 de 0.62 y el que menos lo predecía era el Elastic Net que únicamente llegaba a 0.47, sin embargo, ninguno de estos modelos era lo suficientemente robusto para tener una predicción confiable del proceso.

Tabla 7.

Resultados de los modelos de predicción

<i>Modelo</i>	<i>R2</i>	<i>MAE</i>	<i>RMSE</i>
LinearReg (RFE)	0.58	3185.02	4834.84
ElasticNet	0.47	3532.97	5478.58
NeuralNet	0.63	3048.67	4586.8
RandomForest	0.62	3059.73	4620.8

Nota. La tabla muestra el resumen del coeficiente de determinación, error absoluto medio y la raíz del error cuadrático medio para cada uno de los modelos de predicción.

Luego de la revisión de diferentes modelos, se utilizó el modelo Multi-output, el cual se importa a través de sklearn. Para este modelo se utilizaron como datos de entrada el ciclo de inyección que se quería predecir, el vapor a inyectar durante el ciclo a predecir y adicionalmente se incluyeron las siguientes variables del ciclo anterior:

- Caudal inicial
- Agua producida acumulada hasta el ciclo anterior
- Aceite acumulado hasta el ciclo anterior
- SOR del ciclo anterior
- Pico de producción del ciclo anterior

La **Figura 29** muestra un segmento del código utilizado para la construcción y entrenamiento del modelo Multi-output.

Figura 29.

Código utilizado para el entrenamiento del modelo Multi-Output

```
#####
# SECCIÓN 3: ENTRENAMIENTO MODELO MULTI-OUTPUT
#####

rf = RandomForestRegressor(random_state=42)
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

grid_search = GridSearchCV(
    estimator=rf,
    param_grid=param_grid,
    cv=3,
    scoring='r2',
    n_jobs=-1
)

model_multi = MultiOutputRegressor(grid_search)
model_multi.fit(X_train, y_train)

# Evaluación
y_pred = model_multi.predict(X_test)

rmse_vals = np.sqrt(mean_squared_error(y_test, y_pred, multioutput='raw_values'))
print("RMSE [Np, PIR]:", rmse_vals)

r2_weighted = r2_score(y_test, y_pred, multioutput='variance_weighted')
print(f"R² ponderado: {r2_weighted:.2f}")

# R² individual por cada variable
r2_indiv = r2_score(y_test, y_pred, multioutput='raw_values')
print("R² individual [Np, PIR]:", r2_indiv)

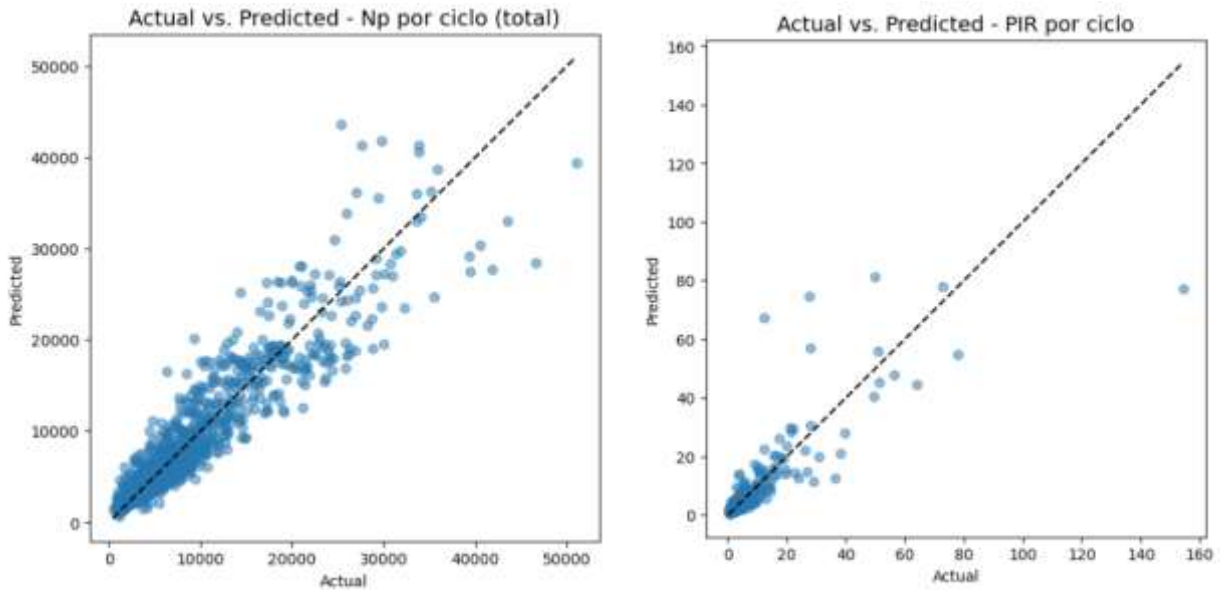
# Ver parámetros óptimos por cada salida
for i, est in enumerate(model_multi.estimators_):
    print(f"Mejores params de '{target_cols[i]}':", est.best_params_)
```

Nota. La figura muestra un segmento del código utilizado para la construcción del modelo Multi-Output en la herramienta Jupyter.

Las variables que se predijeron fueron el aceite acumulado del ciclo actual y el PIR del ciclo actual. En la **Figura 30** se presenta de manera gráfica el ajuste de las dos variables, se puede determinar que se presenta mayor ajuste en el Np por ciclo teniendo en cuenta que ha sido la variable que durante todo el proyecto se ha buscado calcular. El R^2 del modelo para predecir el Np por ciclo de del 0.74, mientras que el PIR tiene un R^2 de 0.72.

Figura 30.

Ajuste de las variables Np por ciclo y PIR por ciclo en el modelo Multi-Output

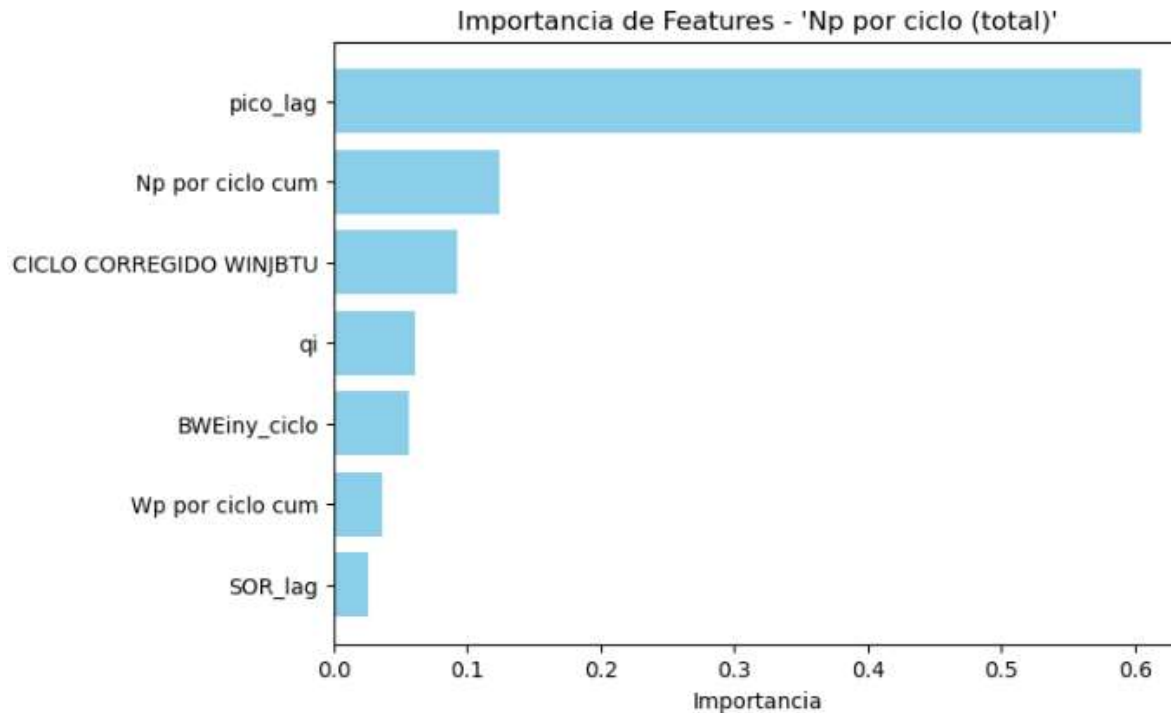


Nota. La figura muestra el comparativo entre la data predica y la real para las variables Np por ciclo y PIR por ciclo.

Finalmente se realizó el comparativo de cuáles eran las variables que mayor preponderancia tenían en el modelo, donde se identificó que el pico de producción tiene un 60% de peso en los resultados que arroja el modelo, seguido por el aceite acumulado total del ciclo anterior. Las variables con mayor impacto en la producción de aceite acumulado del ciclo se observan en la **Figura 31**.

Figura 31.

Ranking de variables con mayor impacto en la producción de aceite en el campo de estudio.



Nota: La figura muestra el grado de influencia que tiene cada una de las variables en la predicción del modelo

4.4. Validación del modelo con información real del área

Los resultados de la comparación entre los valores predichos por el modelo y los valores reales obtenidos de la producción histórica del campo mostraron un ajuste muy cercano. En particular, el análisis mostró un **Coefficiente de Determinación (R^2)** de **0.74**, lo que evidencia que el modelo puede explicar el 74% de la variabilidad en el volumen de aceite incremental.

Para validar el modelo de una forma rigurosa y simular condiciones reales, se agregó un conjunto de datos de datos independiente y no visto durante el entrenamiento. Con este nuevo set se validó la efectividad de predicción del modelo y su capacidad de generalización, evitando el sesgo o sobreajuste. La **Tabla 8** muestra la comparación entre los datos reales y los pronosticados para un pozo ejemplo del Campo de estudio en el que se pueden observar en su mayoría diferencias menores al 10%, lo que lo hace un modelo aceptable.

Tabla 8.

Comparación de datos reales y los datos predichos por el modelo para diferentes ciclos

Pozo X1			
Ciclo	Np_pred	Np real	desviación
7	10548.21	11084.22	5%
8	10637.58	9615.79	-11%
9	6559.654	6377.07	-3%
10	6054.492	6452.48	6%
11	7091.722	7270.42	2%

Nota: la tabla muestra la comparación entre los datos reales y los datos estimados para un pozo ejemplo que no fue utilizado dentro de la construcción del modelo.

Adicionalmente, se realizó un análisis de sensibilidad sobre el modelo, sometiéndolo a perturbaciones controladas en los datos de entrada (valores atípicos, datos faltantes, etc.) con el objetivo de evaluar la estabilidad del modelo y comprobar su capacidad predictiva ante escenarios en los que la calidad de la información histórica pueda verse comprometida.

El modelo demostró tener un comportamiento estable, ya que pudo sostener un rendimiento adecuado a pesar de estas perturbaciones controladas en la manipulación de sus variables de entrada. Esto debido a las relaciones generadas entre las variables que se usaron de entrada y salida para el entrenamiento del modelo.

5. CONCLUSIONES

Dado que un resultado confiable en las predicciones de los modelos depende de la data de entrada y la calidad de la misma es directamente proporcional con la representatividad de los modelos a implementar, la aplicación de la técnica de análisis de datos exploratorio (EDA) tomó aproximadamente el 70% del tiempo del proyecto, llegando a un punto en el cual, aunque se encontraban datos anómalos, no era posible justificar su retiro o ajuste ya que no se contaba con el detalle para determinar de la causa de dicho valor.

Mediante el uso de herramientas de analítica diagnóstica y el desarrollo de modelos de analítica de datos, se logró determinar qué variables como los caudales iniciales del ciclo y el pico de producción por ciclo son altamente correlacionables con el acumulado de producción del ciclo. Así mismo, el SOR tiene una relación inversa con el acumulado de producción y a partir de los resultados se observa que las declinaciones tienen una tendencia similar entre pozos por lo que no se vuelve una variable relevante para la elaboración del modelo de predicción.

El comportamiento de los ciclos previos de un pozo influye significativamente en el rendimiento del siguiente, lo que refuerza la importancia de ejecutar los trabajos operativos correctamente en los ciclos anteriores para pronosticar la producción futura de un pozo.

Teniendo en cuenta que el ciclo de inyección tiene una influencia directa en la eficiencia del proceso, sin embargo, dentro del mapa de calor solo tenía una correlación de -0.52, se utilizaron las variables creadas durante el desarrollo del proyecto de aceite y agua acumulado durante el efecto de la inyección cíclica para aumentar el efecto de esta variable dentro del modelo.

Durante el desarrollo de los modelos se identificó la complejidad de manejar una variable única y al tiempo manipular todos los datos para predecir una variable con ciertos datos esperados. Por lo que se concluye que la mejor metodología es compartimentalizar los datos y establecer rangos semejantes para la construcción del modelo, haciendo que la construcción de estos modelos inicie con resultados de R^2 para la predicción muy bajos, entre 0.3 y 0.4, y con la compartimentalización aumenta hasta 0.74. En el caso del presente estudio, la división de datos se realizó por el número de ciclo de inyección.

Los resultados obtenidos de los 4 modelos desarrollados inicialmente (Regresión lineal múltiple, Elastic net, Red neuronal (MLP Regressor) y Random Forest) alcanzaron un R^2 por debajo de 0.63 lo cual indica que no van a generar predicciones muy acertadas en el cálculo del aceite incremental para el futuro ciclo de inyección.

A partir de los resultados obtenidos luego de la construcción del modelo Multi-Output, se puede concluir que este modelo presenta un ajuste moderado, asociado principalmente a la naturaleza y ruido que presentan los datos y se es necesario tener un mayor detalle e información de ingeniería y operaciones para ajustar o inferir la data, sin embargo, bajo el ajuste actual, el modelo logró predecir el aceite incremental de un futuro ciclo de inyección con una desviación por debajo del 12%, lo cual en términos de reservas puede ser categorizado como reservas probadas y probables.

Existen variables determinantes que pueden mejorar las predicciones de los procesos como la temperatura, presión y sumergencia diaria, sin embargo, en la actualidad estos datos no se están llevando de una forma constante dentro de las bases de datos oficiales, por lo cual, no es fácil rastrear todos estos datos en archivos distintos a los de la base de datos oficial que permitan el uso de estos en los modelos de aprendizaje.

RECOMENDACIONES

Tener una base de datos robusta y confiable es fundamental para un análisis y construcción de un modelo representativo. Durante la revisión de la información se detectaron ciclos faltantes que pueden desplazar los demás ciclos de los pozos categorizándolos erróneamente. Se recomienda revisar los archivos históricos que permitan identificar estos ciclos y que complementen la información.

El uso del modelo contempla que no se presentan cambios por variables operativas, se recomienda considerar el uso de modelos predictivos avanzados que permitan predecir el comportamiento de estas variables bajo diferentes escenarios operativos, con el fin de tomar decisiones más informadas y mejorar el desempeño del proceso a largo plazo.

Teniendo en cuenta que la predicción del acumulado de producción del ciclo tiene como variables de entrada los resultados del ciclo anterior, es importante validar que el ciclo anterior sea representativo y no tenga algún tema operativo que esté alterando los resultados de dicho ciclo.

Una vez se logre una actualización en la cantidad y calidad de la data, se recomienda la aplicación de metodologías de agrupación para la organización de data conforme a las características propias de cada pozo (espesores, locación, propiedades petrofísicas, arena completada) que permitan generar modelos predictivos conforme a las agrupaciones realizadas.

Se recomienda en futuros estudios de analítica de datos, involucrar características geológicas del campo de estudio, ya que factores como el espesor y continuidad de la arena, las saturaciones iniciales de aceite y agua, la cercanía al acuífero, afectan la eficiencia de la inyección cíclica de vapor y las variaciones de acumulados entre un ciclo y otro, por lo cual al segmentar los datos según estas características puede generar modelos predictivos con mayor precisión.

REFERENCIAS

- [1] E. Trigos, E. Lozano and A. M. Jimenez, "CSS: Strategies to Recovery Optimization," *Day 4 Thu, June 14, 2018*, 2018. DOI: 10.2118/190791-ms.
- [2] R. A. Pérez *et al*, "Optimizing Production Performance, Energy Efficiency and Carbon Intensity with Preformed Foams in Cyclic Steam Stimulation in a Mature Heavy Oil Field: Pilot Results and Development Plans," *Day 1 Mon, April 25, 2022*, 2022. DOI: 10.2118/209399-ms.
- [3] C. Alberto *et al*, "SPE-199135-MS New Alternative to Optimize Cyclic Steam Injection: Field Pilot," *America Bogota User On*, vol. 21, 2022-07.
- [4] R. Perez *et al*, "Improving CSS Performance with Preformed Foam: Teca - Cocorna Field Case," *Day 1 Mon, July 27, 2020*, 2020. DOI: 10.2118/199104-ms.
- [5] G. A. Espinosa *et al*, "Assessment of the effect of Cyclic Steam Stimulation (CSS) operational variables on well productivity including geomechanical modeling.
- [6] A. Vakhin *et al*, "Improvement of CSS Method for Extra-Heavy Oil Recovery in Shallow Reservoirs by Simultaneous Injection of in-Situ Upgrading Catalysts and Solvent: Laboratory Study, Simulation and Field Application," *Day 1 Mon, March 21, 2022*, 2022. DOI: 10.2118/200082-ms.
- [7] R. Perez *et al*, "Experimental performance of steam-based hybrid technologies to improve energy efficiency in a colombian heavy oil reservoir," in 2020, Available: <https://www.onepetro.org/conference-paper/SPE-201564-MS>. DOI: 10.2118/201564-MS.
- [8] E. M. Trigos, M. E. Lozano and A. M. Jimenez, "Cyclic Steam Stimulation Enhanced with Nitrogen," *IOR 2019 – 20th European Symposium on Improved Oil Recovery*, pp. 14, 2018. DOI: 10.3997/2214-4609.201900155.
- [9] J. Liu *et al*, "The Application of Complex Displacement in Cyclic Steam Stimulation CSS & Steam Flooding SF Development in Liaohe Oilfield: A Field Performance Study," *Day 1 Wed, March 16, 2022*, pp. 16, 2022. DOI: 10.2118/208940-ms.
- [10] S. Yang *et al*, "Utilization of Time-Lapse Seismic Data to Semi Quantify Residual Oil Saturation by Karhunen-Loeve Transform and Neural Artificial Network during CSS," *SPE Conference Papers*, vol. SPE Europec featured at 78th EAGE Conference and Exhibition, 2016. Available:

- <https://www.geofacets.com?mapId=10.2118/180077-MS&#amp;cld=PrimoMarcFeed&#amp;vld=1.0>. DOI: 10.2118/180077-MS.
- [11] P. Sarma *et al*, "Cyclic steam injection modeling and optimization for candidate selection, steam volume optimization, and SOR minimization, powered by unique, fast, modeling and data assimilation algorithms," in 2017, Available: <https://www.onepetro.org/conference-paper/SPE-185747-MS>. DOI: 10.2118/185747-MS.
- [12] Ersahin,A. and T. Ertekin, (Dec 01,2019)."Artificial Neural Network Modeling of Cyclic Steam Injection Process in Naturally Fractured Reservoirs." *SPE Reservoir Evaluation & Engineering*.Available: <https://www.onepetro.org/journal-paper/SPE-195307-PA>. DOI: 10.2118/195307-PA.
- [13] O. Izgec *et al*, "A Simulation Augmented Machine Learning Approach for Cyclic Steam Stimulation Development Targeting Lower Carbon/Higher Return," *Day 3 Thu, April 28, 2022*. DOI: 10.2118/209336-ms.
- [14] J. J. Trivedi *et al*, "Real-time steam allocation workflow using machine learning for digital heavy oil reservoirs," in 2019, Available: <https://www.onepetro.org/conference-paper/SPE-195312-MS>. DOI: 10.2118/195312-MS.
- [15] Y. Wu *et al*, "Feasibility of SAGD as a Follow-Up Process to CSS for a Massive Deep Bitumen Reservoir," *SPE Conference Papers*, vol. SPE Canada Heavy Oil Technical Conference, pp. 13, 2018. Available: <https://www.geofacets.com?mapId=10.2118/189750-MS&#amp;cld=PrimoMarcFeed&#amp;vld=1.0>. DOI: 10.2118/189750-MS.
- [16] Alzate-Espinosa,G. A. *et al*, (Nov 4,2021)."Assessment of the impact of stress path and strain-dependent permeability on reservoir productivity in CSS." *American Rock Mechanics Association*. DOI: ARMA-IGS-21-077.
- [17] Alboudwarej,H. *et al*, 1980)."Empieza a adquirir importancia el petroleo pesado y las arenas bituminosas The new oil sources." *Energeticos: Boletin Informativo Del Sector Energetico*.

- [18] *Perfiles-Crudo* (). Available: <https://www.eiticolombia.gov.co/es/informes-eiti/informe-2077/perfiles-hidrocarburos/perfiles-crudo/>.
- [19] S. Thomas, "Enhanced Oil Recovery - An Overview," *Oil & Gas Science and Technology*, vol. 63, no. 1, pp. 9-19, Jan. 2008. [Online]. Available: <https://api.istex.fr/ark:/67375/80W-08T91436-Z/fulltext.pdf>. doi: 10.2516/ogst:2007060.
- [20] D. W. Green and G. P. Willhite, *Enhanced Oil Recovery*. Texas: Society of Petroleum Engineers, 1998.6.
- [21] Johannes Alvarez and Sung-Yun Han, (Jul 01,2013)."Current Overview of Cyclic Steam Injection Process." *Journal of Petroleum Science Research*.Available: <https://www.airitilibrary.com/Publication/alDetailedMesh?DocID=P20150604011-201307-201508180022-201508180022-116-127>.
- [22] D. Alvarado, C. Bánzer and A. Rincón, *Recuperación Térmica De Petróleo*. (8th ed.) Caracas: 2002.
- [23] *Glosario de big data* (). Available: <https://piperlab.es/glosario-de-big-data/codigo/#:~:text=En%20el%20contexto%20de%20la,un%20lenguaje%20de%20programaci%C3%B3n%20espec%C3%ADfico>.
- [24] *Glosario de Big data* (). Available: piperlab.es/glosario-de-big-data/python/.
- [25] *¿Qué es Java?* (). . Available: <https://aws.amazon.com/es/what-is/java/>.
- [26] *¿Qué es R?* (). . Available: <https://datademia.es/blog/que-es-r>.
- [27] Escuela de Ingeniería Industrial "El lenguaje C++". Available: www2.eii.uva.es/fund_inf/cpp/temas/1_introduccion/introduccion.html.
- [28] Tsai,C. *et al*, 2015)."Big data analytics: a survey." *Journal of Big Data*.
- [29] OIC. *¿Qué es big data?* (). . Available: <https://www.oracle.com/co/big-data/what-is-big-data/>.
- [30] Zhong,R., C. Salehi and R. Johnson, (Dec2022)."Machine learning for drilling applications: A review." *Journal of Natural Gas Science and Engineering*.Available: <https://dx.doi.org/10.1016/j.jngse.2022.104807>. DOI: 10.1016/j.jngse.2022.104807.
- [31] *ANALÍTICA DE NEGOCIOS COMO ESTRATEGIA DE TRANSFORMACIÓN EMPRESARIAL* (). Available: <https://blog.formaciongerencial.com/analitica-de-negocios-como-estrategia-de-transformacion-empresarial/>.

- [32] D. Barrero, A. Pardo, C. Vargas, y J. Martínez, *Colombian Sedimentary Basins: Nomenclature, Boundaries and Petroleum Geology, a New Proposal*. ANH and B&M Exploration Ltda., 2007.
- [33] L. V. Sosa Jerez y L. C. Zamora Alvarado, *Estructura De Redes Neuronales (Mlp) Y Su Aplicación Como Aproximador Universal*". Universidad Distrital Francisco José de Caldas, 2022.
- [34] Aguilar, F. (2017). "Multiple Regression and Recursive Feature Elimination (RFE)." Medium. [En línea] Disponible: <https://medium.com/@feraguilari/multiple-regression-and-recursive-feature-elimination-rfe-34af0c6ae51b>.
- [35] Seabold, S., & Perktold, J. (2010). "Statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference, 57-61. [En línea] Disponible: <https://doi.org/10.25080/Majora-92bf1922-011>
- [36] M. Blondel y J. Van den Bossche, "Scikit-learn", *Scikit-learn*. [En línea]. Disponible en: <https://scikit-learn.org/stable/>.
- [37] X. Glorot y Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", International Conference on Artificial Intelligence and Statistics, pp. 249–256, 2010.
- [38] J. A. Rodrigo (2020), "Random Forest con Python", Ciencia de datos. [En línea]. Disponible en: https://cienciadedatos.net/documentos/py08_random_forest_python.