Escuela Superior Politécnica del Litoral

Facultad de Ingeniería en Ciencias de la Tierra

Predicción de la Producción de Pozos de Petróleo del Campo Sacha utilizando Analítica de Datos

INGE -2439

Proyecto Integrador

Previo la obtención del Título de:

Ingeniero en Petróleos

Presentado por:

Roberto Antonio Saltos Ortega

Guayaquil - Ecuador

Año: 2024

Dedicatoria

A Dios, fuente de sabiduría y luz en mi camino, quien ha guiado cada paso de este proceso.

A mis padres, Roberto Saltos Nevarez y Miguela Ortega Merino, y hermana, Sabrina Saltos Ortega, que me han brindado su amor incondicional y su constante apoyo.

A esa persona especial, Odalys Uchubanda, cuya presencia y apoyo incondicional me han acompañado a lo largo de este camino.

Agradecimientos

Quisiera comenzar expresando mi más profundo agradecimiento a Dios. Su guía y sabiduría han sido fundamentales en el desarrollo de este trabajo.

A la Escuela Superior Politécnica del Litoral.

Al MSc Freddy Carrión Maldonado por su guía en este proyecto y enseñanza.

Declaración Expresa

Yo **Saltos Ortega Roberto Antonio** acuerdo y reconozco que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 23 de mayo del 2024.

Saltos Ortega Roberto Antonio

Evaluadores

ANDRES EDUARDO GUZMAN VELASQUEZ

Firmado digitalmente por ANDRES EDUARDO GUZMAN VELASQUEZ
Fecha: 2024.09.09 15:40:49 -05'00'

MSc. Andrés Guzmán Velasquez

Profesor de Materia

Firmado electrónicamente por:
FREDDY PAUL CARRION MALDONADO

MSc. Freddy Carrión Maldonado

Tutor del proyecto

Resumen

La predicción de la producción de pozos petroleros es crucial para optimizar operaciones y reducir costos en la industria petrolera. Esta tesis se centra en desarrollar un modelo predictivo utilizando análisis de series temporales y modelos de regresión para pronosticar la producción de pozos en la arena U inferior del Campo Sacha – BLOQUE 60 ubicado en la provincia de Orellana, Ecuador. El modelo incorpora datos históricos de producción desde 2014 hasta 2024, donde el periodo 2013 - 2021 se utilizó para entrenamiento y el año 2022 para validación de los modelos, con el objetivo de evaluar la producción futura para los años 2023-2024.

El proyecto emplea el modelo SARIMAX (Modelo Autorregresivo Integrado de Media Móvil Estacional con Regresores Exógenos), que es particularmente adecuado para manejar los patrones estacionales y las variables exógenas que influyen en la producción de petróleo. Además, se evaluaron varios modelos de regresión, con XGBOOST logrando el mejor desempeño durante la validación, alcanzando un 98% de eficiencia según la métrica de R^2, mientras que SARIMAX alcanzó un 72%. Sin embargo, en la evaluación del período 2023-2024, el modelo SARIMAX mostró un rendimiento óptimo mayor a Random Forest en un 53%.

Los resultados demuestran la efectividad tanto del modelo SARIMAX como de Random Forest en la predicción de períodos de tiempo discontinuos, considerando las brechas en los datos y proporcionando información útil para ingenieros de campo. Esta propuesta no solo mejoraría la precisión de las predicciones, sino que también ayudaría en la planificación estratégica y la asignación de recursos en el proceso de extracción de petróleo.

Palabras claves: predicción de producción, modelos de regresión, series temporales, extracción de petróleo.

Abstract

Predicting oil well production is crucial to optimize operations and reduce costs in the oil industry. This thesis focuses on developing a predictive model using time series analysis and regression techniques to forecast well production in the "U inferior" sand of the Sacha Field - BLOCK 60 located in Orellana, Ecuador. The model incorporates historical production data from 2014 to 2024 where the period 2014 - 2021 was used for training and the year 2022 for model validation with the objective of evaluating future production for the years 2023-2024.

The project employs the SARIMAX model (Seasonal Moving Average Integrated Autoregressive Model with Exogenous Regressors), which is particularly suited to handle seasonal patterns and exogenous variables influencing oil production. In addition, several regression models were evaluated, with XGBOOST achieving the best performance during validation, reaching 98% efficiency according to the R^2 metric, while SARIMAX reached 72%. However, in the 2023-2024 evaluation, the SARIMAX model showed a 53% optimal performance than Random Forest.

The results demonstrate the effectiveness of both SARIMAX and Random Forest in predicting discontinuous time periods, considering gaps in the data and providing useful information for field engineers and decision makers. This approach not only improves prediction accuracy, but also aids in strategic planning and resource allocation in the oil extraction process.

Keywords: production forecasting, regression models, time series, oil extraction.

Índice general

Abreviaturas

API	American Petroleum Institute
BES	Bombeo Electro-sumergible
BSW	Base Sediments Water
BT	Bagging Tree
DP	Procesamiento de Datos
DT	Decission Tree
GOR	Relación Gas Petróleo
IQR	Rango Intercuartílico
k-NN	k-Nearest Neighbors
LOF	The Local Outlier Factor
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
PIP	Presión de entrada
RF	Random Forest
RNA	Redes Neuronales Artificiales

Simbología

API Gravedad API del Petróleo.

BAPD Barriles de Agua por Día

BFPD Barriles de Fluido Por Día

BPPD Barriles de Petróleo Por Día

CO_2 Dióxido de carbono

Hz Hertz

MMSCFD Millones de Pies Cúbicos Estándar

R^2 Índice de correlación

RMSE Raíz del Error Cuadrático Medio.

Índice de figuras

Índice de tablas

Capítulo 1

1.1 Introducción

La industria petrolera se enfrenta a dos desafíos cruciales: la sobreproducción de aguas de formación y la incertidumbre en la producción de nuevas arenas petroleras. Estos problemas generan importantes costos operativos, ambientales y económicos que impactan negativamente en la rentabilidad y sostenibilidad del sector. (Ayauca et al., 2023)

En este contexto, la aplicación de la analítica de datos y el machine learning se presenta como una herramienta poderosa para abordar estos desafíos de manera efectiva. Estas técnicas permiten procesar y analizar grandes volúmenes de información histórica y actual de los pozos petroleros, identificando patrones y tendencias que no podrían ser detectados a simple vista. Esta información puede ser utilizada para desarrollar modelos predictivos que optimicen la producción de pozos existentes y mejoren la precisión en la estimación de la producción de nuevas arenas petroleras. (Al-Balushi et al., 2019)

En los últimos años, se ha producido un avance significativo en la aplicación del ML en la industria petrolera. Diversos estudios han demostrado la efectividad de los modelos predictivos para caracterizar la geología de los yacimientos petroleros que permite predecir la porosidad, permeabilidad y distribución de las formaciones rocosas que contienen petróleo. (Wan y Chen, 2023), además de optimizar la perforación de pozos con el fin de identificar las ubicaciones más prometedoras para perforar pozos y predecir la profundidad a la que se encuentran los reservorios de petróleo. (Al-Balushi et al., 2019), incluso existen modelos para el estudio de la predicción de fallas en equipos críticos para la producción, permitiendo un mantenimiento preventivo y reduciendo los tiempos de inactividad.

En este trabajo se busca enfocar el análisis de la producción de pozos y el descubrimiento de patrones en la producción de pozos existentes y predecir su producción futura.

1.2 Descripción del Problema

Se abordan dos problemáticas principales de la industria petrolera: la producción excesiva de aguas de formación y la producción de pozos con caudales bajos. La primera problemática se refiere a la cantidad de agua producida por los pozos petroleros y lo difícil que es predecir su caudal a lo largo del tiempo usando métodos convencionales y el sesgo que existe en estos. (Singh & Jain, 2018)

La segunda problemática parte desde la rentabilidad de la producción de pozos con caudales bajos, ya que suelen tener un tiempo de vida menor y una mayor inversión para su explotación debido a que en muchos casos por la declinación brusca de producción se pierden muchos equipos en el cambio de las completaciones, dando lugar a la necesidad de una propuesta que pueda tener un mejor análisis del historial futuro de fluidos de cada pozo evitando el aumento de los costos en las operaciones. Estas dos problemáticas no solo afectan a una sola institución, sino a la industria petrolera en general. (Olson & Anderson, 2018)

1.3 Justificación del Problema

El proyecto cuenta con potencial, ya que permite optimizar las operaciones de completación al momento de elegir el tipo de sistema de levantamiento artificial que se usará a causa de que en el tiempo de vida útil de un pozo sus caudales pueden variar (Guadalupe et al., 2017) y estimar la capacidad de producción de nuevos pozos de petróleo utilizando la información histórica sobre el comportamiento de los pozos que tienen como zona de interés la arena U inferior (Al-Balushi et al., 2019). La importancia de esta propuesta se centra en poder encontrar la relación que existe entre las diversas variables que diariamente se estudian en campo aplicando machine learning que en comparación con estudios como las curvas de declinación presentan un menor uso de recursos siempre y cuando el porcentaje de eficacia

sea lo más cercano al 100% (Dehganhi et al., 2024). Una de las mayores ventajas de este proyecto es que usando la analítica de datos es posible resolver las dos problemáticas antes mencionadas proporcionando una herramienta que pueda predecir el comportamiento de producción de un yacimiento sabiendo que las formaciones no tienen una tendencia lineal y su naturaleza no es fácil de interpretar. (Li R. et al., 2024)

1.4 Objetivos.

1.4.1 Objetivo general

Estimar la cantidad de petróleo que producirá un pozo en la arena U inferior antes de los disparos usando la información histórica de producción para el análisis de los futuros caudales de crudo y riesgo de inversión.

1.4.2 Objetivos específicos

- Clasificar la información obtenida de campo para realizar el preprocesamiento de los datos.
- Implementar un modelo de machine learning utilizando los datos preparados.
- Comparar los resultados obtenidos del modelo de datos con un modelo analítico mediante análisis de curvas de declinación.
- Desarrollar un dashboard que permita el análisis integrado con la generación de predicciones de producción para distintos pozos.

1.5. Marco teórico

1.5.1 Definición de conceptos claves

1.5.1.1. Conceptos de Ingeniería de Producción

Sobreproducción de aguas de formación: Estos fluidos de formación tienen alta salinidad y se encuentran junto al petróleo en los yacimientos. Durante la extracción

del petróleo, estos fluidos también son extraídos y representan un costo considerable para las empresas petroleras debido a su tratamiento y disposición. (IPIEC, 2021)

Crudo: Fluido producido extraído del yacimiento. (SLB Glossary)

BSW: Fracción de agua presente en el crudo. (SLB Glossary)

Sistema de levantamiento artificial: Mecanismo que permite el aumento de presión para aumentar la producción de fluidos de un pozo. (SLB Glossary)

Recuperación primaria: Producción de fluidos de un pozo utilizando la energía del. (SLB Glossary)

1.5.1.2. Analítica de datos

Analítica de datos: Procedimiento que se realiza para recopilar, limpiar, transformar y analizar conjuntos de datos con el fin de obtener información valiosa y útil para la toma de decisiones. (European Knowledge Center for Information Technology, 2023)

Machine Learning (ML): Ciencia que se especializa en el estudio de datos históricos con el fin de poder crear modelos que sean capaces de predecir el futuro para tomar mejores decisiones. (Lopez, 2017)

Algoritmo: Metodología o instrucciones que se debe aplicar a los datos. (Lopez, 2017)

Aprendizaje supervisado: Proporcionar al programa una data que tenga las respuestas correctas para entrenarlo. (Lopez, 2017)

Aprendizaje no supervisado: Entrenar el algoritmo usando una data que no está clasificada para que el programa pueda encontrar las relaciones a partir del aprendizaje supervisado. (Lopez, 2017)

Dashboard: Presentación principal donde se resumen los resultados de un programa o algoritmo.

Sobreajuste: Problema referente a la baja precisión de predicción de un modelo cuando trabaja con aprendizaje no supervisado. (Lopez, 2017) Nota: Especificar el problema.

1.5.1.3. Librerías para analítica de datos

Pandas: Librería de Python enfocada al uso de estadísticas. (Vanderplas, 2016)

Sktlearn: Librería de Python enfocada en el análisis de datos y predicciones. (Vanderplas, 2016)

Matplotlib: Librería de Python enfocada en la creación de visualizaciones o animaciones. (Vanderplas, 2016)

Seaborn: Librería de Python enfocada en la creación de gráficos estadísticos. (Boschetti & Massaron, 2018)

Plotlib: Librería de Python enfocada en la creación de gráficos interactivos.

Streamlit: Librería en Python que permite la creación de entornos de trabajo como aplicaciones o páginas web para el análisis de datos.

1.5.2 Antecedentes

En la formación Montney del Campo Montney en Canadá se llevó a cabo un estudio de minería de datos de 4790 pozos horizontales con los siguientes parámetros geológicos: la ubicación de los pozos, profundidad vertical, longitud lateral y dirección, además de contar con la información de estimulación, etapas de fractura y caudales de fluido. Lo siguiente fue realizar la visualización e interpretación de la data adquirida para observar el comportamiento entre el diseño de estimulación y la producción del primer año. En este caso de estudio se aplicó el método de eliminación de características recursivas para medir la importancia de cada variable en el modelo de predicción teniendo en cuenta 4 aspectos fundamentales del aprendizaje supervisado:

random forest, refuerzo adaptivo, máquina de vectores de soporte y redes neuronales. Los resultados de aplicación demostraron que en el primer año la producción acumulada incrementa con el aumento de la profundidad vertical real del pozo, longitud del pozo, agente de apuntalamiento bombeado por el pozo, fluido inyectado y número de etapas. La validación de esta información se realizó con el coeficiente de Pearson con valores de R entre 0.24 y 0.53 aunque se encontró que no existía relación entre el tipo de terminación del pozo y la producción en el caso de los pozos horizontal con sistema de pozo abierto. El modelo creado se puede usar para mejorar las metodologías de estimulación considerando la masa del apuntalante y los caudales. (Wang & Chen, 2019)

Se desarrolló un modelo que permite tanto predecir la producción como optimizar los parámetros de fracturación, empleando una combinación de técnicas de aprendizaje automático y enfoques de optimización evolutiva. Para alcanzar este fin, el estudio se propuso las metas de predecir las series temporales de producción de petróleo de esquisto mediante el desarrollo y comparación de cuatro modelos diferentes de predicción basados en el aprendizaje automático, introducir un innovador modelo de optimización evolutiva híbrida para estimar el valor presente neto máximo y determinar los parámetros óptimos de fracturación seleccionados. Se emplearon las características geológicas, los parámetros de fracturación hidráulica y la producción de petróleo de esquisto de 10000 pozos generados para desarrollar un modelo de predicción basado en datos. La división fija de los conjuntos de entrenamiento y prueba podría ocasionar problemas significativos de sobreajuste. Para prevenir este problema, se utilizó la validación cruzada de k-fold (k = 8), lo que permitió evaluar tanto el

rendimiento de predicción como la capacidad de generalización del modelo de aprendizaje automático. Este método k-fold alterna entre conjuntos de entrenamiento y prueba, proporcionando así una evaluación completa del modelo. (Dong et al., 2022)
Para evaluar con mayor precisión la exactitud de los cuatro modelos de predicción de producción basados en datos, utilizamos el índice de correlación (R^2) como métrica. El valor de R^2 oscila entre 0 y 1, donde un valor alto señala una fuerte concordancia y un valor bajo indica una concordancia deficiente. (Dong et al., 2022)
Los resultados de los modelos de random forest (RF) y multilayer perceptron MLP presentan valores de R^2 más altos, sin embargo, se observa que RF tiende al sobreajuste, ya que su valor de R^2 en el conjunto de prueba es un 10% menor que en el conjunto de entrenamiento. Esto lleva a la conclusión de que el algoritmo RF se ha ajustado demasiado al conjunto de entrenamiento, lo que afecta negativamente su rendimiento con datos de prueba. En consecuencia, aunque RF produce buenas predicciones en el conjunto de entrenamiento, su desempeño es deficiente en muestras no vistas. Por lo tanto, se concluyó que MLP es el algoritmo más adecuado para establecer el modelo de predicción de producción en este estudio. (Dong et al., 2022)

El proceso inició con una evaluación del código de detección de novedades y el entorno de trabajo de Scikit-learn creando ventanas temporales para las fases de entrenamiento y prueba en el programa. En la primera fase se utilizó una base de datos pública de sensores de pozos y plataformas en operación para aplicar una normalización del módulo StandarScaler con el fin de parametrizar la información. La densidad de desviación de los datos fue detectada por un método de detección de anomalías o valores atípicos no supervisado llamado The Local Outlier Factor (LOF). El modelo

tuvo la caracterización de anomalías en pozos y líneas de producción para el cierre espurio de DHSV, problema que ocurre sin presentar problemas en la superficie y de difícil detección para operadores humanos, por esta razón al realizar operaciones de cierre automáticas es posible la realizar la reapertura con procesos correctivos sin aumentar costos adicionales o disminución de la producción. Los sensores que alimentan la base de datos son el TPT, sensores de presión y temperatura PDG, sensores de presión en superficie y línea de producción, posición del extrangulador o CHOKE. Los resultados óptimos tuvieron un índice cp igual a 0 comprobado principalmente por F1-SCORE y ACCb con una precisión de 0.991 y desviación estándar de 0.014. Por último, se realizó una comparativa bibliográfica con otros métodos de algoritmos de predicción como Random Forest (RF), Decission Tree (DT) y Long Short-Term Memory (LSTM). (Aranha et al., 2024)

Tabla 1.

Comparación del modelo con otros tipos de algoritmos.

Algoritmos	**Accuracy**	**F1- SCORE**	**ACCb**
RF	0.8708	-	-
DT	0.6000	0.4900	-
LSTM (g=0.6)	0.9992	0.9360	-
LOF (cp=0)	0.991	0.9969	0.9910

Nota. Datos tomados del caso de estudio de Aranha et al. (2024)

En el campo Sacha se perforaron nuevos pozos en el año 2021 en la arenisca T inferior en donde se implementó un modelo de predicción de producción de crudo. La base de datos contenía información petrofísica y de fluidos para desarrollar el modelo a partir de un pozo de referencia, utilizando un software comercial para analizar el

comportamiento de los pozos. Paralelamente, se desarrollaron dos algoritmos mediante el lenguaje de programación Python y técnicas de Machine Learning con redes neuronales artificiales (RNA) y procesos de clustering, uno basado en los datos de presión de entrada (PIP) de la bomba electrosumergible (BES), y otro que además incorpora la salinidad del agua de formación del reservorio. La diferencia de la información real y el software comercial fue del 2% sin embargo las simulaciones en Python presentaron errores del 10% y 0.5%, respectivamente. Para la predicción de gas, el valor real fue de 0.07 MMSCFD, mientras que la simulación con el software comercial arrojó 0.41 MMSCFD. El modelo con Python ofreció aproximaciones más precisas, con valores de 0.11 MMSCFD y 0.10 MMSCFD. El aumento de variables en la base de datos, incluyendo la PIP de la BES y la salinidad del agua de formación permitieron reducir el porcentaje de error aumentando la precisión en la predicción de producción de fluidos y gas. (Altamirano – Cárdenas et al., 2023)

La arenisca Weber del campo Rangely en Colorado, Estados Unidos, se encontraba bajo un proceso de CO2-EOR en el año 2012. Este caso de estudio tuvo una recuperación secundaria con inyección de agua desde 1976 hasta 1986, cuando se iniciaron las operaciones con CO2. Este cambio provocó un aumento alrededor del 10% en la producción de petróleos de los primeros 5 años y a partir del sexto año se presenció un nuevo declive en el caudal. La predicción del factor de recobro para ambos métodos de recuperación se realizó a través del método convencional de curvas de declinación al igual que en otros 15 campos en el país. (Jahediesfanjani, 2017)

Figura 1.

Caudales, declinación y coeficiente de determinación por análisis de curvas de declinación.

Case study number in appendix C1	Oil field	State	Stratigraphic unit containing the reservoir	Waterflood			CO_2-EOR		
				q_i (bbl/day)	D_i (/year)	R^2	q_i (bbl/day)	D_i (/year)	R^2
1	Sable*	TX	San Andres Limestone	2,950	299.8	0.99	2,390	182.4	0.99
2	Rangely	CO	Weber Sandstone	89,500	215.9	0.98	88,000	169.9	0.92
3	Lost Soldier	WY	Tensleep Formation	18,000	427.8	0.95	23,000	314.3	0.96
4	Lost Soldier	WY	Madison Formation	6,900	497.9	0.98	8,500	329.9	0.96
5	Wasson	TX	San Andres Limestone	421,000	300.9	0.98	120,000	43.1	0.95
6	Wasson	TX	Clear Fork Group	26,000	279.3	0.99	20,000	65.0	0.93
7	Dollarhide	TX	Thirtyone Formation	23,000	405.9	0.89	11,000	87.6	0.85
8	Dollarhide	TX	Clear Fork Group	7,500	426.8	0.95	9,000	213.3	0.97
9	Salt Creek	TX	"Canyon-age reservoir"	49,000	154.3	0.94	116,000	295.1	0.96
10	Seminole	TX	San Andres Limestone	106,000	232.3	0.87	97,000	129.4	0.96
11	Twofreds	TX	Ramsey Member	6,000	1,042.5	0.91	2,400	176.1	0.96
12	Vacuum	NM	San Andres Limestone	61,000	271.6	0.96	44,000	132.7	0.97
13	Cedar Lake	TX	San Andres Limestone	14,000	217.6	0.99	14,500	94.8	0.96
14	North Hobbs	NM	San Andres Limestone	1,200	387.3	0.97	1,900	402.6	0.96
15	Yates	TX	San Andres Limestone	256,000	250.2	0.91	244,000	221.7	0.98
Average				**72,500**	**360.7**	**0.951**	**125,400**	**190.5**	**0.952**

Nota. Datos tomados del Estudio Geológico de Reston, Virginia, Estados Unidos. (2017)

Tabla 2.

Información de casos de estudio.

Casos de estudio	**Referencias**	**Descripción**
Caso Montney	Wang, S., & Chen, S. (2019). Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. Journal Of Petroleum Science & Engineering, 174, 682-695. https://doi.org/10.1016/j.petrol.2018.11.076	Mejorar la fractura analizando los parámetros de caudales y masa del apuntalante.
Modelo Híbrido	Dong, Z., Wu, L., Wang, L., Li, W., Wang, Z., & Liu, Z. (2022). Optimization of Fracturing Parameters with Machine-Learning and Evolutionary Algorithm Methods. Energies, 15(16), 6063. https://doi.org/10.3390/en15166063	Modelo que predice la producción y optimiza los parámetros de fractura.

The Local Outlier Factor – Santos, Brasil	Aranha, P.E., Policarpo, N.A. & Sampaio, M.A. Unsupervised machine learning model for predicting anomalies in subsurface safety valves and application in offshore wells during oil production. *J Petrol Explor Prod Technol* **14**, 567–581 (2024). https://doi.org/10.1007/s13202-023-01720-4	Modelo de machine learning para encontrar anomalías en los datos que envían los sensores del pozo
Sacha – T inferior	Altamirano-Cárdenas, A. I., & Lucero-Calvache, F. A. (2023). Predicción de Producción de Fluidos empleando Machine Learning en T Inferior del Campo Sacha. FIGEMPA (En LíNea), 16(2), 70-78. https://doi.org/10.29166/revfig.v16i2.4542	Modelo de predicción de producción aplicado en el campo Sacha en la arena T inferior.
Caso de studio - Weber Sandstone, Rangely Oil Field	Jahediesfanjani, H. (2017). Application of decline curve analysis to estimate recovery factors for carbon dioxide enhanced oil recovery. Scientific Investigations Report. https://doi.org/10.3133/sir20175062c	Uso de curvas de declinación para estimar el Factor de Recobro con CO2-EOR.

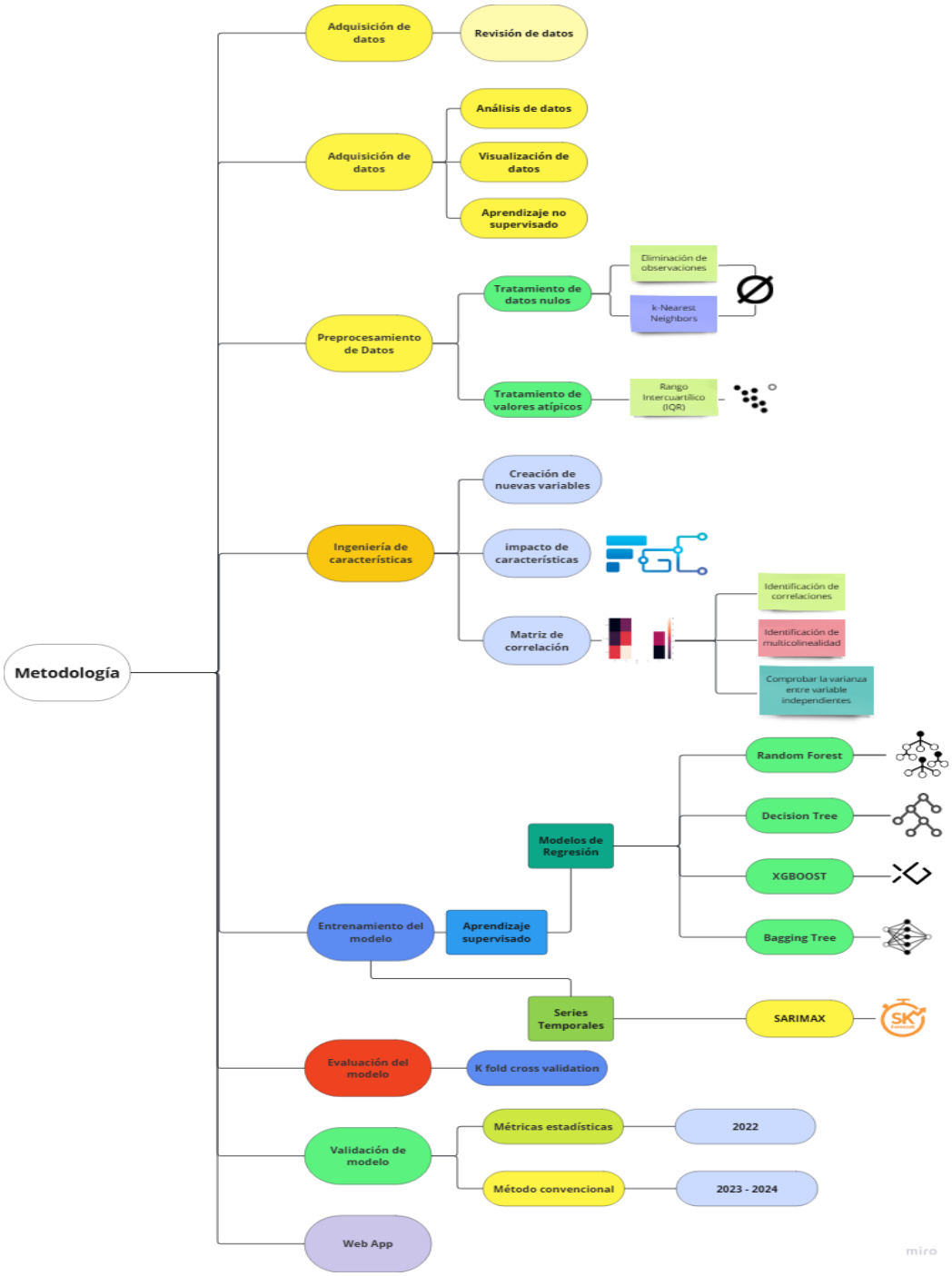
Capítulo 2

2. Metodología.

2.1. Esquema de trabajo

Figura 2.

Mapa conceptual del flujo de trabajo en el proyecto.



En la figura 2 se observa los procedimientos que fueron aplicados al conjunto de datos desde la adquisición de la información hasta la creación de la interfaz en formato Web App para los usuarios, con la finalidad de llevar un orden que permitiera optimizar las instrucciones del proyecto.

2.2. Integración de los datos.

Con la debida autorización de EP PETROECUADOR se obtuvieron los permisos para hacer uso de la información histórica de los pozos en producción de la arena U Inferior del campo Sacha (Bloque 60). En esta primera etapa solo se realizó una revisión y validación de los datos antes de utilizarla en el proyecto que se desarrolló. Los procesos principales que se aplicaron al conjunto de datos incluyeron una revisión de la información para asegurar que cumplan con ciertos criterios y estándares establecidos, verificando la consistencia, integridad y exactitud de los datos. (Batini & Scannapieca, 2006)

2.3. EDA (Exploratory Data Analysis)

El proceso de exploración que se le aplicó a los conjuntos de datos permite comprender la estructura, características y relaciones que pueden existir entre los datos. Además, se realizó una aplicación de técnicas de control de calidad como la identificación y corrección de errores, eliminación de duplicados, revisión de datos nulos y revisión de valores atípicos. (Batini & Scannapieca, 2006).

Algunas de las técnicas implementadas fueron:

- Análisis Exploratorio

Consistió en la descripción de los datos utilizando medidas básicas de dispersión como media, mediana, desviación estándar, mínimos y máximos para luego proceder en la búsqueda de valores faltantes y atípicos con la ayuda scripts que permiten la búsqueda de outliers. Por último, fue necesaria la exploración de la distribución de cada variable en el caso de tener presencia de asimetrías poco visibles. (Batini & Scannapieca, 2006)

- Visualización de los datos

  Graficar los datos de las variables numéricas permitió tener una mejor comprensión de la información con la que se trabajó, la relación de sus caracteristicas y la presencia de patrones ocultos, donde los gráficos de barra, dispersión, las matrices de correlación y mapas de calor dieron un mejor enfoque de las columnas que presentan más impacto para la predicción de la variable objetivo. (Batini & Scannapieca, 2006)

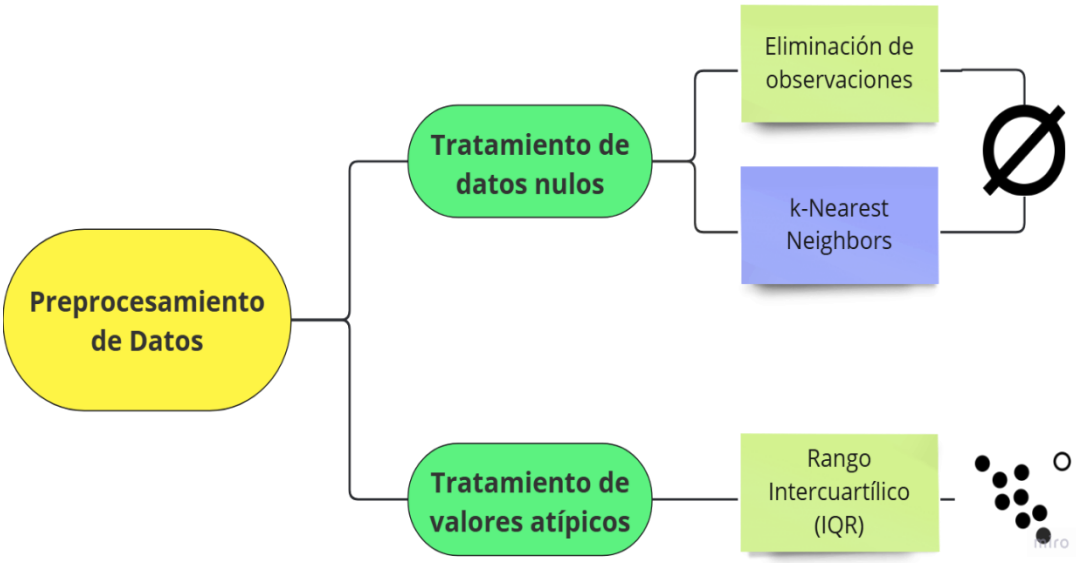
- Aprendizaje no supervisado

  Se utilizaron algoritmos de clustering, como DBSCAN y clustering jerárquico, para agrupar los datos en categorías naturales y descubrir subgrupos dentro del conjunto de datos. Al final de esta etapa, se evaluaron los resultados de los algoritmos de clustering para interpretar los grupos formados y su relevancia en el contexto del análisis. (Batini & Scannapieca, 2006)

2.4. Procesamiento de datos

Figura 3.

Esquema de trabajo para el procesamiento de datos.



El procesamiento de datos (DP) es necesario para preparar los datos antes de la aplicación de los modelos predictivos. Este proceso incluye varias etapas, tales como la limpieza de datos y tratamiento de valores atípicos

La aplicación de Rango Intercuartílico (IQR) fue calculado con los valores del cuartil 1 (Q1) y el cuartil 3 (Q3) con la finalidad de limitar el conjunto de datos entre ambos cuartiles y eliminar la presencia de las celdas que se encuentren fuera de rango. (Larson, 2006)

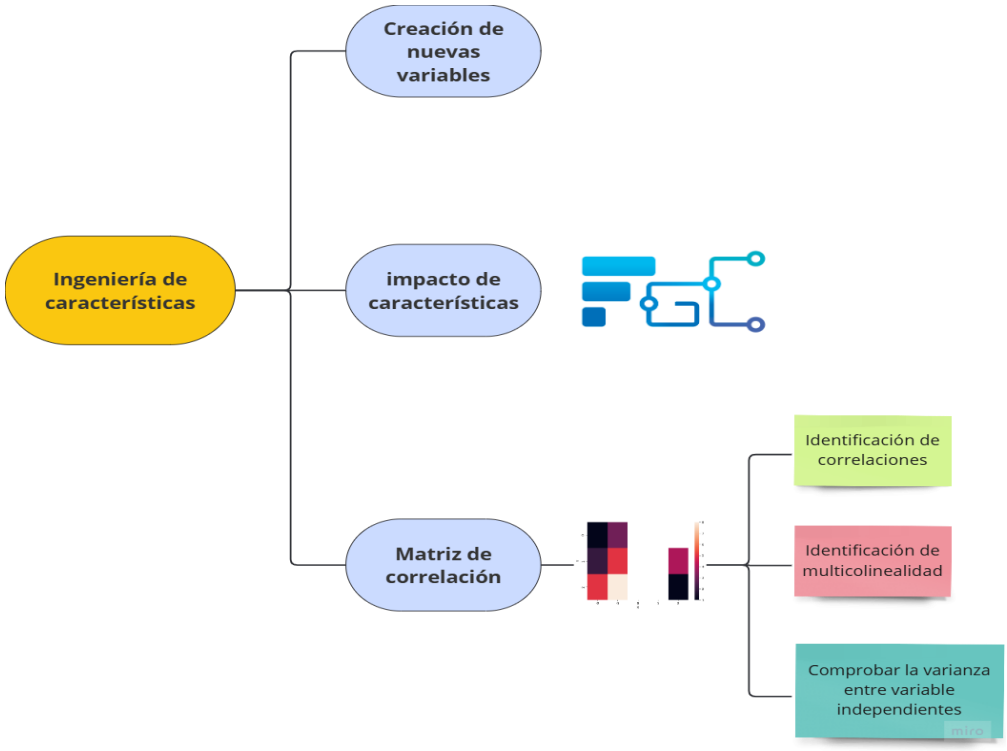
La limpieza de datos se llevó a cabo para asegurar que el conjunto de datos estuviera libre de errores, valores faltantes y anomalías. Los valores faltantes se imputaron utilizando métodos apropiados, como el k-Nearest Neighbors (k-NN). (Ilyas & Chu, 2019). En casos donde los valores faltantes sean numerosos y no puedan ser

imputados de manera confiable, se optó por aplicar la regla del 75%, si el número de valores en cada observación es igual o mayor a ese porcentaje se elimina la fila. Se realizó una revisión tanto manual como automática de los datos para identificar y corregir errores tipográficos e inconsistencias. (Walker, 2020)

2.5. Ingeniería de características

Figura 4.

Esquema de trabajo para la ingeniería de características.



A partir del conjunto de datos con el que se trabaja este proyecto, fue necesario elegir las variables que tienen mayor relevancia para el modelo. Se inició con la clasificación de las variables entre cuantitativas o cualitativas debido a que, dependiendo del tipo de parámetro, se aplicó el correspondiente arreglo o procesamiento. Luego, se realizó la creación de las posibles derivaciones, variables

creadas a partir de otras variables, que ayudaron al modelo a tener una mejor eficiencia y solides. (Bangert, 2021)

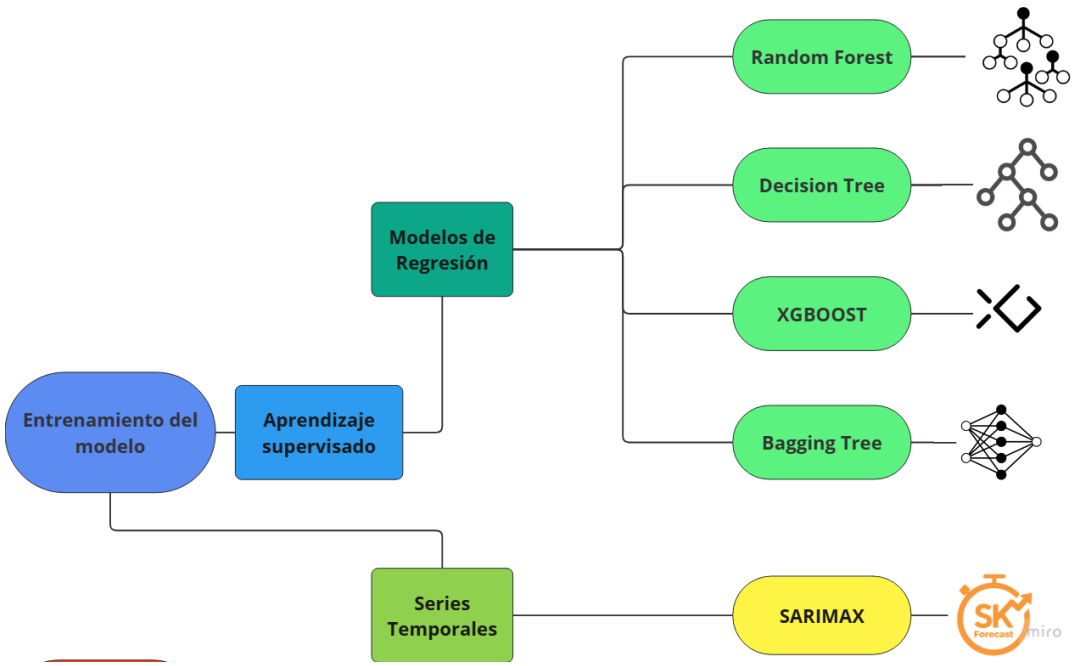
Otro proceso que se realizó fue la revisión de impacto de las características que sirvió para identificar y cuantificar la relevancia de cada variable, es decir, que características permiten tener un mejor rendimiento de los modelos predictivos. La librería Scikit-learn ofreció opciones para realizarlo de forma automática con funciones de los módulos RandomForest o DecisionTree. (Géron, 2019)

La matriz de correlación utilizada en esta sección fue el mapa de calor que se dividió en 3 subprocesos: identificación de correlaciones, multicolinealidad usando la correlación de Pearson, método que calcula el coeficiente de correlación de cada indicador e identificar las variables que tienen una fuerte correlación con los indicadores clave o variables independientes, abarcando tanto las correlaciones positivas como negativas, para completar el análisis de correlación, y evaluación de varianza entre variable independientes. (Bangert, 2021). En el modelo se aplicaron estos subprocesos para eliminar las variables que estén altamente correlacionadas entre sí para mejorar la interpretación y rendimiento del modelo. (Gerón, 2019)

2.6. Entrenamiento del modelo – Aprendizaje Supervisado con modelos de Regresión

Figura 5.

Esquema de trabajo del entrenamiento del modelo



En esta sección iniciaron los procesos de aprendizaje supervisado y se definió que la variable objetivo es cuantitativa, debido a esto se optó por modelos de regresión los cuales se utilizaron para predecir el caudal de petróleo de pozos en el campo SACHA haciendo uso de la analítica de datos. Por otro lado, se consideró la implementación de modelos de series temporales para poder predecir valores a futuros con una variable que dependa del tiempo y variables exógenas que ayuden al ajuste del algoritmo. El conjunto de datos para entrenamiento y evaluación consideró una partición porcentual 80-20 respectivamente en el periodo de tiempo 2014-2022. Posteriormente, se realizaron predicciones del periodo 2023 - 2024 usando el modelo de regresión óptimo y el modelo de series temporales para ser comparadas con los datos reales de producción y un método convencional - ecuaciones de declinación de Arps, serie de fórmulas empíricas utilizadas en la industria del petróleo y gas para

predecir la producción futura de pozos petroleros y de gas a partir de datos históricos de producción. (Arps, 1945).

Los modelos propuestos son:

- Random Forest (RF): es un conjunto de árboles de decisión entrenados de manera aleatoria y combinados. Cada árbol se crea a partir de una muestra aleatoria del porcentaje de datos específicos para entrenamiento y utiliza un subconjunto aleatorio de características en cada partición de nodo. (Géron, 2019)
- Decision Tree (DT): divide iterativamente el espacio de características en regiones más pequeñas, basándose en las características más importantes, para tomar decisiones o predecir valores. (Géron, 2019)
- XGBOOST (XGB): crea un conjunto de modelos en secuencia con el fin de que cada modelo intente corregir los errores de los modelos anteriores, optimizando una función de pérdida regularizada. (Géron, 2019)
- Bagging Tree(BT): consiste en capas de nodos (neuronas) conectados entre sí donde cada conexión tiene un peso ajustable que se optimiza en el entrenamiento para minimizar el error en las predicciones. (Géron, 2019)
- SARIMAX: sus siglas significan “Media móvil integrada autorregresiva estacional con factores exógenos” y es utilizado para la predicción de datos en factor de tiempo haciendo uso de la estacionalidad e implementación de variables exógenas. (Durbin & Koopman, 2012)

Se priorizó buscar un rendimiento óptimo en cada modelo para evitar problemas de subajuste o sobreajuste. (Bangert, 2021). Por esta razón se realizó la optimización de hiperparámetros. En la tabla 3 se especifican las modificaciones consideradas.

Tabla 3.

Hiperparámetros en modelos de machine learning.

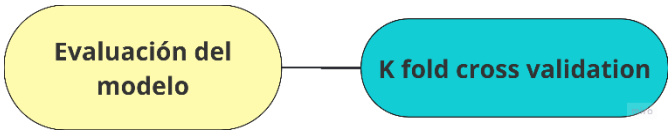
Modelo	**Hiperparámetros**
RF	• Número de árboles. • Profundidad máxima por árbol. • Número mínimos de muestras necesarias. • Número de características a considerar.
DT	• Profundidad máxima del árbol. • Número mínimos de muestras necesarias. • Número de características a considerar. • Calidad de la división.
XGB	• Número de árboles o estimaciones. • Profundidad máxima del árbol. • Tasa de aprendizaje. • Proporción de características para cada árbol.
BT	• Número de estimaciones. • Profundidad máxima del árbol. • Proporción de características para cada árbol.
SARIMAX	• Orden. • Estacionariedad. • Frecuencia de los datos. • Varibales exógenas.

Información tomada de Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Gerón, 2019.

2.7. Evaluación de modelos de regresión

Figura 6.

Evaluación de modelos de regresión.



Al finalizar el entrenamiento de cada modelo de regresión se utilizó la técnica K-fold cross-validation para valorar la exactitud y precisión. Se realizó este procedimiento en 24 particiones con el fin de tomar el resultado promedio de todas las ejecuciones. De forma porcentual la partición se fijó en 80% - 20% para entrenamiento y validación y se repitió una cantidad de veces configuradas en el algoritmo.

2.8. Validación del modelo

2.8.1. Métricas y desviación del rendimiento.

Las métricas que permiten valorar el rendimiento de cada propuesta son: Bagging Tree(BT) y coeficiente de determinación (R2). (Thabet et al., 2024).

$$RMSE = \sqrt{\frac{\sum(yi - \widehat{yi})^2}{N}} \qquad (2.1)$$

RMSE = raíz del error cuadrático medio.

yi = valores de la variable dependiente.

$\widehat{yi}$ = valores de predicción del modelo.

N = número de observaciones

Ecuación 1. Raíz del error cuadrático medio.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \widehat{yi})^2}{\sum_{i=1}^{n}(yi - \overline{yi})^2} \quad (2.2)$$

yi = valores de la variable dependiente.

$\widehat{yi}$ = valores de predicción del modelo.

$\overline{yi}$ = media de la variable dependiente.

N = número de observaciones.

Ecuación 2. Coeficiente de determinación

2.8.2. Comparación con método convencional

Por último, para validar el comportamiento de los modelos de datos, así como su precisión, se realizó una comparación con un método convencional que nos permite obtener predicciones similares a las de los algoritmos. Este método se conoce como análisis de curvas de declinación de producción previamente mencionado en la sección 2.6. En una gráfica se visualizan los valores reales de producción de los últimos 5 años como referencia para tener una mejor apreciación de la eficiencia en las predicciones del modelo y la técnica convencional. (Lin et al., 2017)

$$q(t) = q_i e^{D_i t} \quad (2.3)$$

q(t) = caudal en función del tiempo t.

qi = caudal inicial.

Di = tasa de declinación.

t = diferencia de tiempo inicial y tiempo final.

Ecuación 3. Declinación exponencial.

$$q(t) = \frac{q_i}{(1+bD_i)^{\frac{1}{b}}} \quad (2.4)$$

q(t) = caudal en función del tiempo t.

qi = caudal inicial.

Di = tasa de declinación.

t = diferencia de tiempo inicial y tiempo final.

b = exponente de declinación de Arps.

Ecuación 4. Declinación Hiperbólica.

2.9. Aplicación Web

Una vez concluido el proyecto, se inició el diseño de una interfaz que permita al usuario interactuar con el modelo. La biblioteca de Streamlit en Python tiene la característica de que se relaciona bien con librerías de machine learning como scikit-learn. De forma más específica, en la configuración de la interfaz se agregó la opción de cargar base de datos y a partir de esta se creó un menú que diera acceso a la información por pozo, visualización de gráficas de producción y presentación de estadísticas. Se requirió importar los paquetes de Python para que el modelo trabajara dentro de la interfaz. Este tipo de presentación puede ser utilizada en cualquier tipo de sistema operativo. (Mensah, 2023)

Capítulo 3

3. Resultados y análisis

3.1. Integración de los datos

3.1.1. Carga de datos

La información adquirida por EP PETROECUADOR estaba separada en 3 conjuntos de datos: Pozos influenciados (inyección de agua), Pozos no influenciados y la Presión de reservorio de cada uno de los pozos. Cada archivo se subió a un notebook de JUPYTERLAB, utilizando codificación con Python.

La cantidad de pozos con los que se trabajó el modelo fue un total de 12 presentes en la arena "U inferior" del BLOQUE 60 – CAMPO SACHA con los siguientes nombres: SCHAL-442UI, SCHAL-443UI, SCHAL-444UI, SCHAL-445UI, SCHAL-579UI, SCH-118UI, SCHA-418UI, SCHAF-387S1U, SCHAF-499UI, SCHAF-519S1U, SCHAF-529HUI', SCHAF-539HS1.

3.1.2 Organización de datos

En los datos de pozos influenciados y no influenciados se encontraba el historial de producción de cada pozo en hojas diferentes por lo que se realizó un proceso de manipulación previo al preprocesamiento para concatenar toda la información en una sola hoja para cada archivo. Posteriormente, se concatenaron ambos conjuntos de datos con una verificación previa de que exista el mismo número de características o variables.

Por otra parte, fue necesario crear la variable categórica "Influencia" dentro del conjunto de datos concatenados para poder diferenciar los pozos que habían sido intervenidos en trabajos de reacondicionamiento e inyección de

agua. Si el valor de la característica es 1, significa que el pozo ha sido afectado, pero si es 0 se entiende que no se ha realizado algún proceso de intervención. El último proceso realizado en esta sección es con respecto al enlace que se creó entre la base de datos principal con las presiones de reservorio y presiones de punto de burbuja de cada pozo debido a la importancia de ambas características.

3.2. Análisis Exploratorio de Datos (EDA)

3.2.1. Exploración de datos

Teniendo en cuenta que la información adquirida se enlazó en una base de datos principal, se realizó la inspección de los tipos de datos de cada una de las variables con el fin de ir depurando aquellas características que no presenten datos numéricos o contengan información relevante.

Tabla 4.

Información de las variables presentes en la base de datos.

Columna	Datos no-nulos	Tipo de dato
Nombre de Pozo	2782	object
Fecha Efectiva	2782	datetime64[ns]
Fluido Total	2782	float64
Crudo	2782	float64
AGUA	2782	float64
Corte de Agua	2782	float64
Gas Total	2782	float64
GOR	2782	float64
API Crudo	2782	float64
Salinity	2782	float64
Pump Frequency	2782	float64

Presión de Succión	2782	float64
Presión de Descarga	2782	float64
Temperatura de Succión	2782	float64
Temperatura de Descarga	2782	float64
Presión de Tubing	2782	float64
Código SPT	2782	float64
Presión de Casing	2782	float64
Temp Tubing	2782	float64
Comentarios	2160	object
* Motor Temp	2782	float64
Tipo de Bomba	2758	object
Pump Model	2722	object
Pump Date	2755	object
Horas de Prueba	2782	float64
Influencia	2782	int64
Pr.	2782	int64
Pb	2782	int64

Nota. Información tomada por EMPRESA (2024).

En la Tabla 4 fue posible apreciar el tipo de datos que contenía cada una de las características que se encontraban en el conjunto de datos. Este fue el primer filtro que se aplicó ya que todas las variables que eran de tipo “object” se eliminaron además de aquellas de tipo “float64” o numéricas que no tenían mayor relación con el caudal de Petróleo, estas variables son: Codigo SPT, Pump date, Pump Model, Comentarios, Tipo de bomba, Horas de prueba y Motor Temp.

Previo a la **sección 3.2.2** se aplicó la función "describe( )" de la librería Pandas con el objetivo de conocer las propiedades estadísticas de cada característica. En la **tabla 5** se puede apreciar la información de 8 variables numéricas presentes en la base de datos.

Tabla 5

Propiedades estadísticas de la base de datos.

	Fluido Total (BFPD)	Crudo (BPPD)	AGUA (BAPD)	Corte Agua (%)	GOR (MMSCF /BPPD)	API	Pr. (PSI)	Pb (PSI)
Total	2782	2782	2782	2782	2782	2782	2782	2782
Prom	645.19	536.92	108.27	16.44	348.24	24.64	843.99	1106.35
std	412.82	373.88	165.31	20.74	390.39	4.27	55.96	105.42
min	32	0	1	0.39	0	0	700	930
25%	304.25	238	6	1.32	155	20.2	820	1126
50%	552	444	39	4.07	203.01	26.3	840	1126
75%	889	702.25	140	28.03	600	26.9	885	1185
max	1849	1710	1144	100	15527.59	33.2	1000	1226

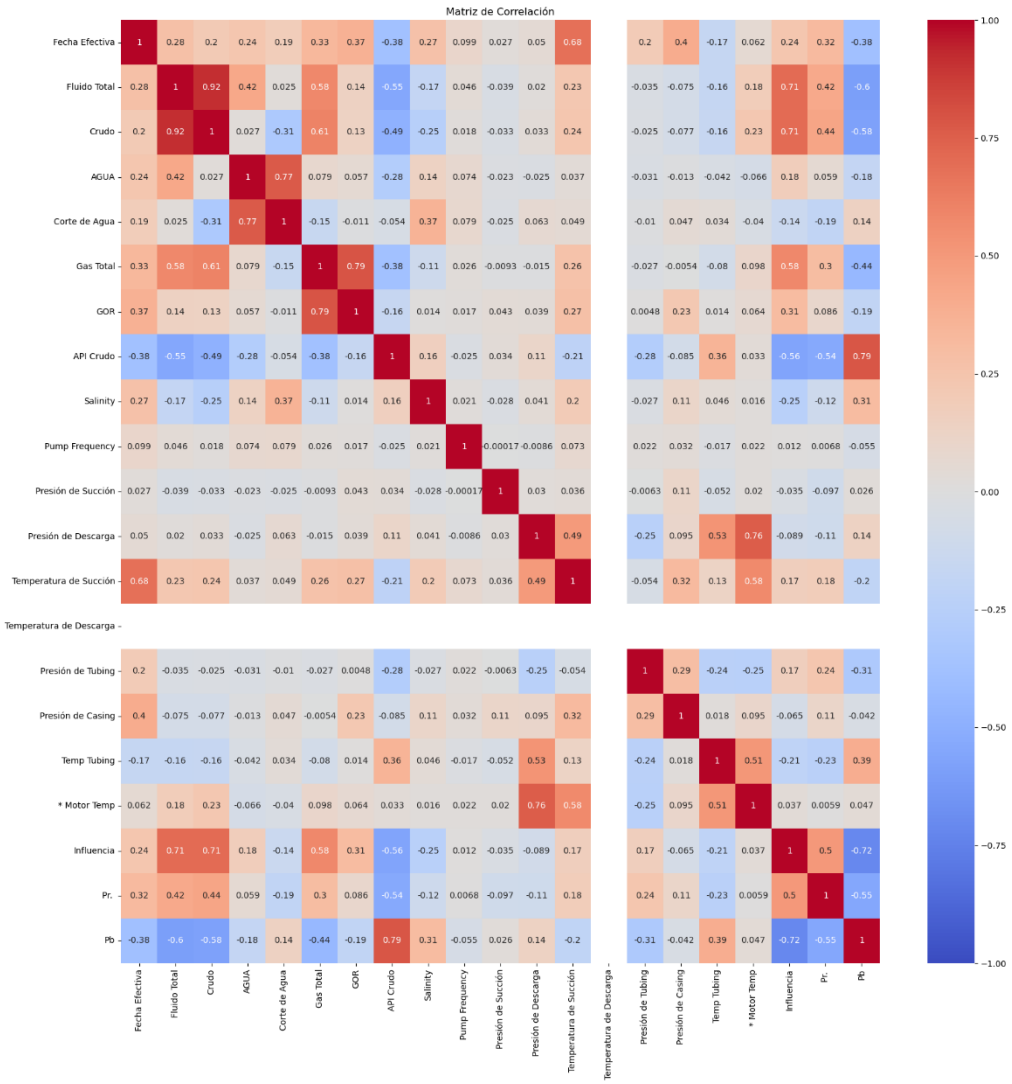
Nota. Información tomada de EP PETROECUADOR (2024).

A pesar de que la producción de los pozos se realiza en el mismo yacimiento "U Inferior", según la tabla 4, se puede apreciar una alta desviación estándar en las columnas de producción de fluidos. Por otro lado, la gravedad API del crudo en la zona tiene una desviación estándar de 4.27 pero su valor en el primer cuartil es 20.2 y el máximo 33.2 que puede ser un factor determinante para clasificar el petróleo en la escala API, es decir, si un pozo presenta valores superiores a 30 se lo puede considerar como petróleo liviano, pero si está por debajo de este límite y mayor a 20, se dice que es un petróleo mediano.

3.2.2. Visualización de los datos

Figura 7

Matriz de correlación de la base de datos.

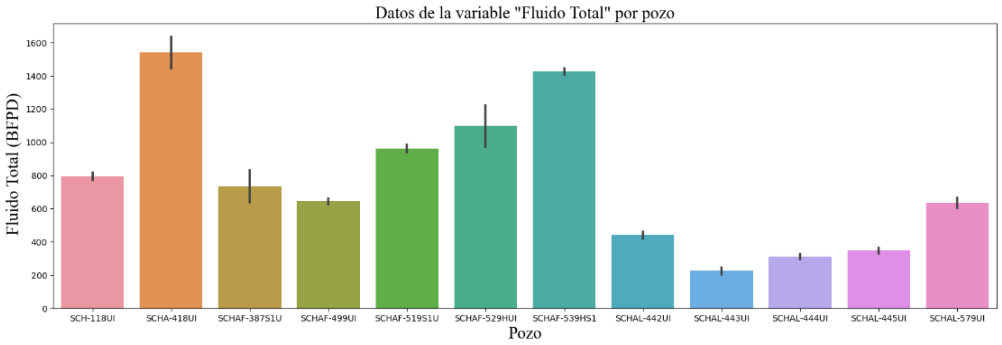


Con respecto a la figura 7, la matriz de correlación aplicada a las variables originales permitió conocer la correlación que existe entre sí mismas, aunque visualmente lo primero que se apreció fueron los datos nulos presentes en la característica “Temperatura de descarga”, por lo tanto, esta variable fue eliminada de forma manual.

Lo siguiente dentro del apartado **3.2.2.** es la visualización de los gráficos de barras de cada variable presente en la Tabla 4 con los respectivos pozos, donde fue posible apreciar la confiabilidad de los datos y su valor promedio con la finalidad de encontrar tendencias, patrones y otros comportamientos.

Figura 8

Gráfico de barras de la variable "Fluido Total".

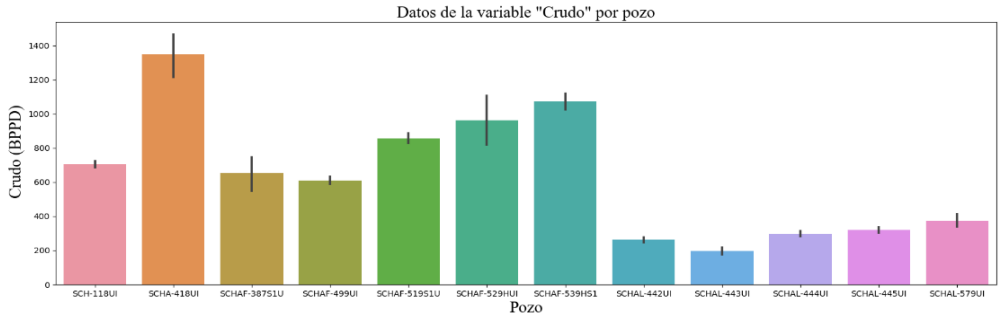


Nota. Información tomada de EP PETROECUADOR (2024).

La figura 8 mostró que el pozo SCHA-418UI tiene una mayor producción de fluidos totales (petróleo, gas y agua) además de una mayor confiabilidad en sus datos por la línea gris que se presenta en la parte superior de la columna, esto específica que el porcentaje de confianza de los datos es superior al 95%.

Figura 9

Gráfico de la variable "Crudo".

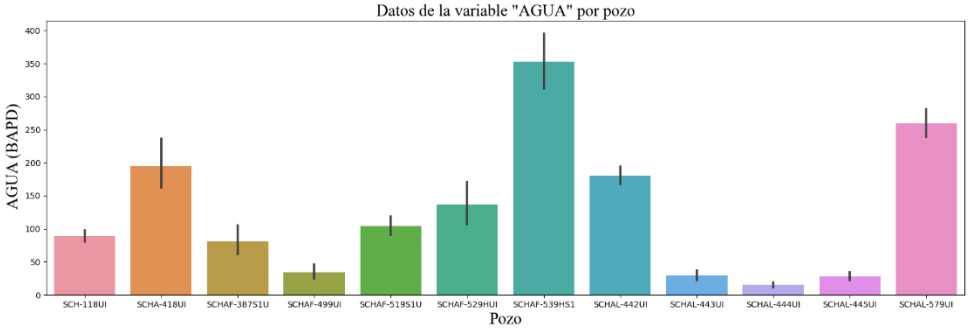


Nota. Información tomada de EP PETROECUADOR (2024).

La figura 9 expuso que el pozo SCHA-418UI cuenta con una mayor producción de petróleo y producción de fluidos totales. Este comportamiento permitió asumir que un mayor levantamiento de fluidos es mayor levantamiento de petróleo, además de que ambas variables se encuentren fuertemente correlacionadas según la Figura 7. Por otro lado, el SCHAF-539HS1 demuestra mayor confiabilidad en sus datos.

Figura 10

Gráfico de la variable "AGUA".

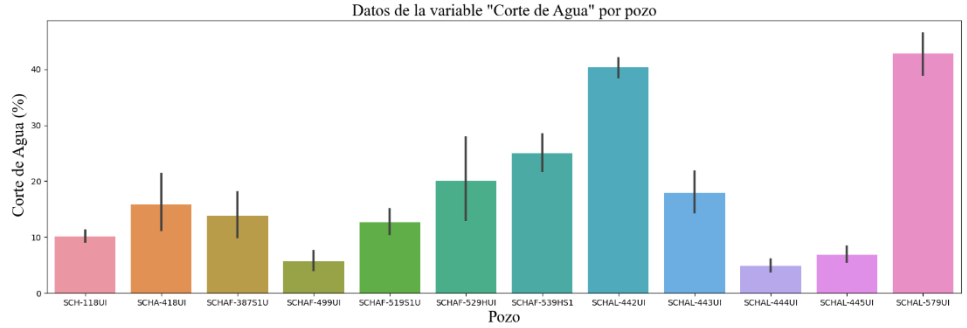


Nota. Información tomada de EP PETROECUADOR (2024).

En los datos de la Figura 10 se destacó un alto valor de producción de agua en el pozo SCHAF-539HS1.

Figura 11

Gráfico de la variable "Corte de Agua"

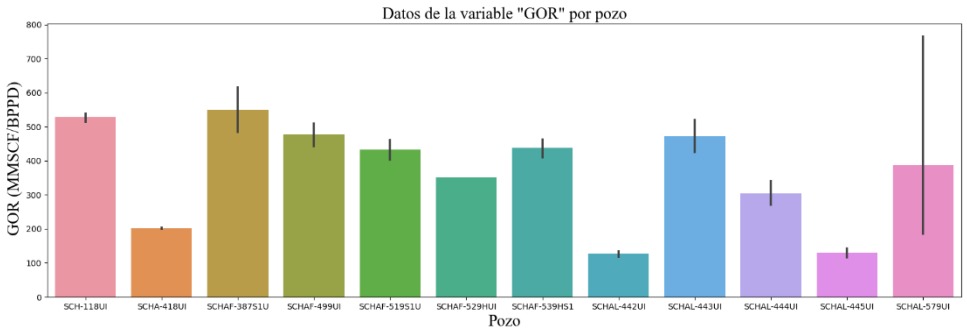


Nota. Información tomada de EP PETROECUADOR (2024).

Aunque en la figura 10 el pozo con mayor producción de agua es el SCHAF-539HS1 esto no significa un mayor corte de agua debido a que esta medida es proporcional a la producción total de fluidos en cada pozo. En la Figura 11 se apreció que el pozo con un mayor BSW es el SCHAL-579UI.

Figura 12

Gráfico de la variable "GOR"

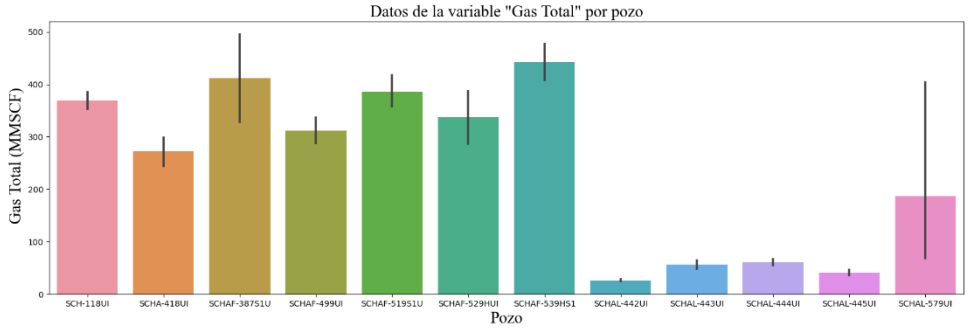


Nota. Información tomada de EP PETROECUADOR (2024).

La figura 12 expone que el pozo SCHAL-579UI tiene una mayor confiabilidad en los datos por tener un intervalo de confianza del 95% con mayor amplitud (línea negra vertical). Este pozo fue uno de los que mayor corte de agua y producción del mismo fluido demostraron, pero en la figura 12 se observa un menor valor de GOR en comparación al resto de pozo. Aunque estas variables tienen una mínima correlación entre sí según la figura 7, se puede apreciar que los pozos con un GOR más alto tienen un menor corte de agua.

Figura 13

Gráfico de la variable "Gas Total"

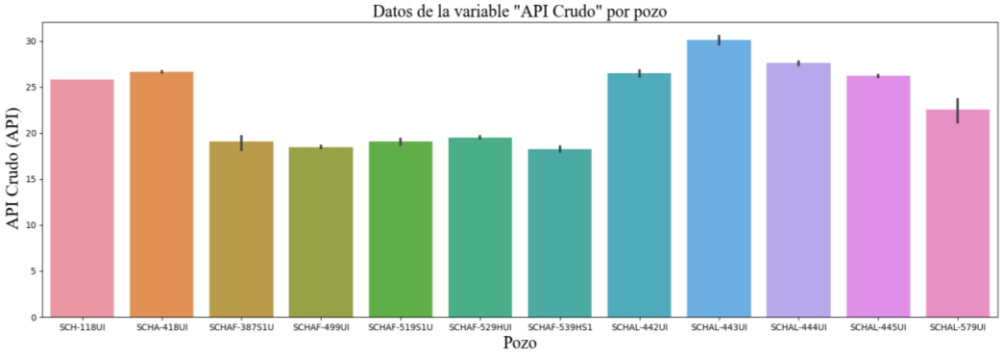


Nota. Información tomade de EP PETROECUADOR (2024).

La figura 13 describe el comportamiento de la variable “Gas Total” donde fue posible corroborar el comportamiento observado en la figura 12 que expuso la relación que existe entre la producción de agua y gas. Los primeros 6 pozos que se encuentran de izquierda a derecha son aquellos con los valores mayores de GOR y producción de gas, pero cuando se observaron las variables de corte de agua y producción en los mismos pozos sus valores eran menores en comparación al resto de pozos.

Figura 14

Gráfico de la variable “API”



Nota. Información tomade de EP PETROECUADOR (2024)

El crudo presente en la arena U Inferior en el campo Sacha tiene una gravedad API que según el pozo SCHAF-539HS1 tiene un valor mínimo menor a 20 API o un máximo mayor a 30 API. Por el comportamiento evaluado en cada

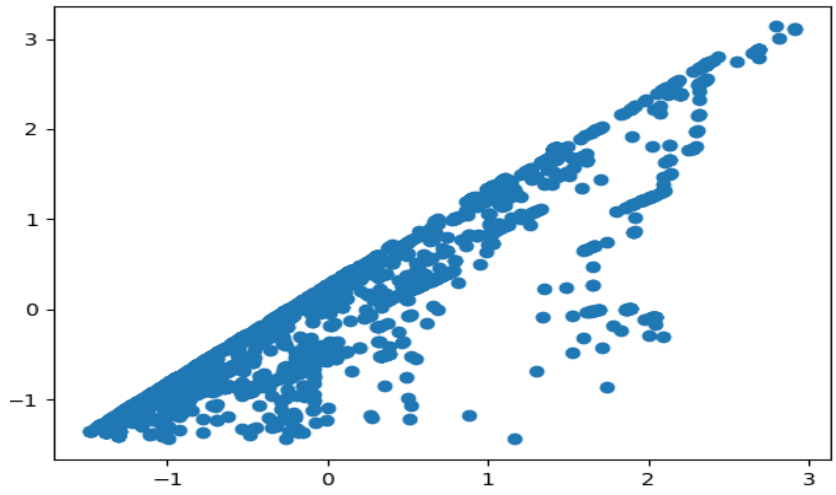
pozo, el crudo producido es de carácter mediano según la escala API para clasificar el petróleo.

3.2.3. Aprendizaje no Supervisado

En el conjunto de datos se inició con un proceso de escalamiento para mejorar la organización de la información bajo una escala previo a la aplicación de algoritmos de aprendizaje no supervisado: DBSCAN y Herarchical.

Figura 15

Escalamiento de datos

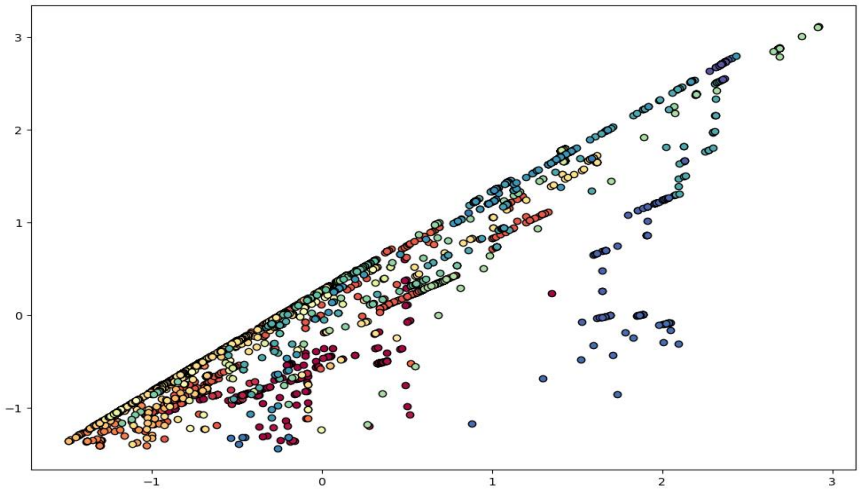


Nota. Información tomada por EP PETROECUADOR (2024)

La figura 15 permitió observar que existe una tendencia lineal entre la distancia y varianza que existe en los datos. El primer algoritmo de agrupamiento que se empleó fue DBSCAN exponiendo que en la información encontró 70 grupos diferentes por similitudes y 35 datos atípicos bajo un épsilon (distancia de un punto al centroide) de 1.70 como se puede apreciar en la figura 16.

Figura 16.

DBSCAN aplicado al conjunto de datos

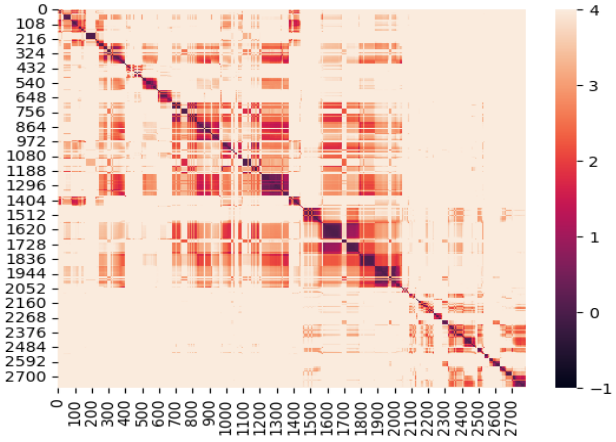


Nota. Información tomada por EP PETROECUADOR (2024)

En la figura 16 se observó la cantidad de agrupamientos encontrados por lo que se infirió la existencia de subgrupos dentro del conjunto de datos, esto se debe a la unión del historial de producción de los 12 pozos en un solo conjunto de datos.

Figura 17

Matriz de distancia aplicando Herarchical Clustering



Nota. Información tomada por EP PETROECUADOR (2024)

En la figura 17 se realizó la matriz usando la distancia euclidiana que demostró de forma gráfica los agrupamientos presentes en el conjunto de datos y por otra parte también la presencia de datos atípicos en aquellos que tienen distancias superiores mayor o igual a 4 según la escala de la matriz.

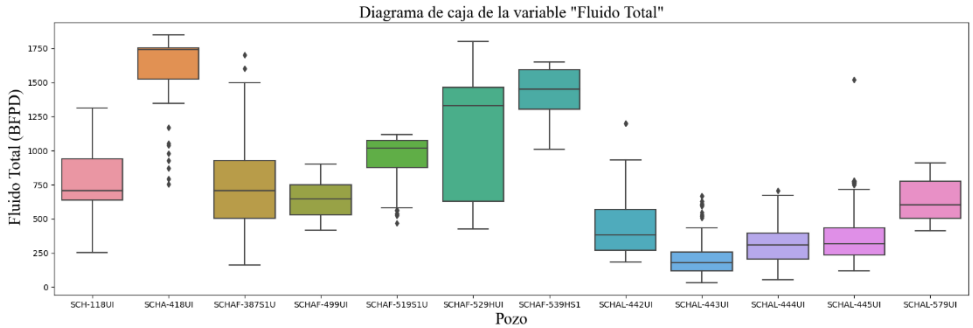
3.3. Procesamiento de datos

3.3.1. Tratamiento de valores atípicos

El conjunto de datos no contaba con valores nulos hasta este apartado por lo que se continuó con el tratamiento de valores atípicos usando el método del rango intercuartílico. Para observar la presencia de las celdas que se encontraban fueran de rango luego del procedimiento mencionado anteriormente, se utilizaron diagramas de caja.

Figura 18

Diagrama de caja de la variable "Fluido Total" posterior a tratamiento de valores atípicos.



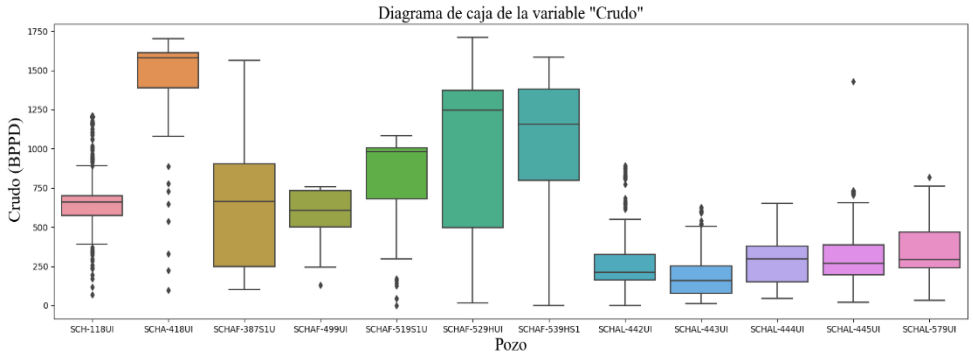
Nota. Información tomada por EP PETROECUADOR (2024)

La figura 18 mostró que la cantidad de valores atípicos en la variable son pocos y están presentes solo en ciertos pozos, para observar este fenómeno se debe tomar en cuenta los puntos que están fuera del diagrama de caja. Este gráfico permitió ver el rango de producción en que se maneja cada pozo

demostrando que el posible con mayor intervalo de levantamiento es el SCHAF – 529HUI.

Figura 19

Diagrama de caja de la variable "Crudo" posterior a tratamiento de valores atípicos.

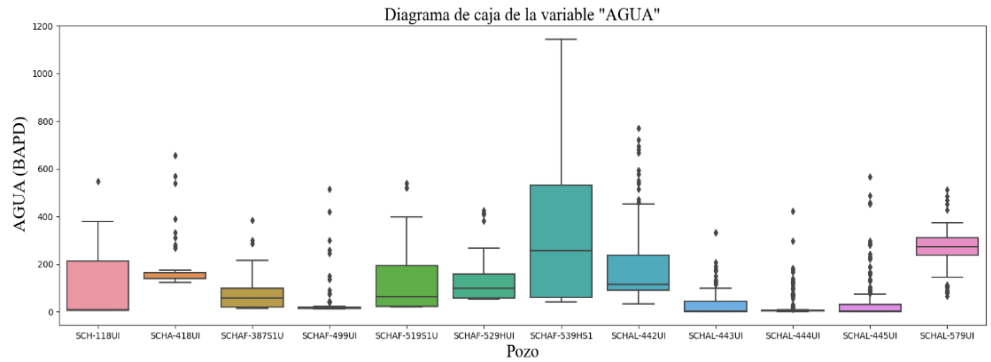


Nota. Información tomada por EP PETROECUADOR (2024)

En la figura 19 se pudo observar que el pozo SCHAF – 529HUI también tiene un mayor intervalo de levantamiento de crudo además de fluidos totales y no cuenta con presencia de valores atípicos.

Figura 20

Diagrama de caja de la variable "AGUA" posterior a tratamiento de valores atípicos.

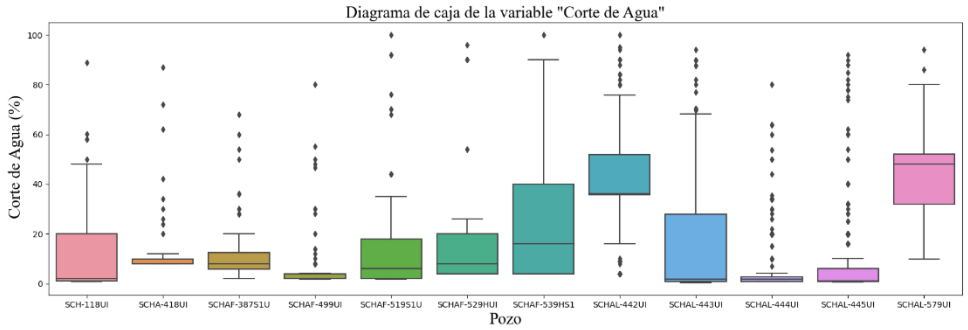


Nota. Información tomada por EP PETROECUADOR (2024)

El diagrama de caja de cada pozo presente en la figura 20 permitió analizar la presencia de valores atípicos en la mayoría de los pozos. Este dato no significa que sean datos erróneos, al contrario, pueden ser situaciones en que la producción de agua esté fuera del intervalo común de cada uno.

Figura 21

Diagrama de caja de la variable "Corte de Agua" posterior a tratamiento de valores atípicos.

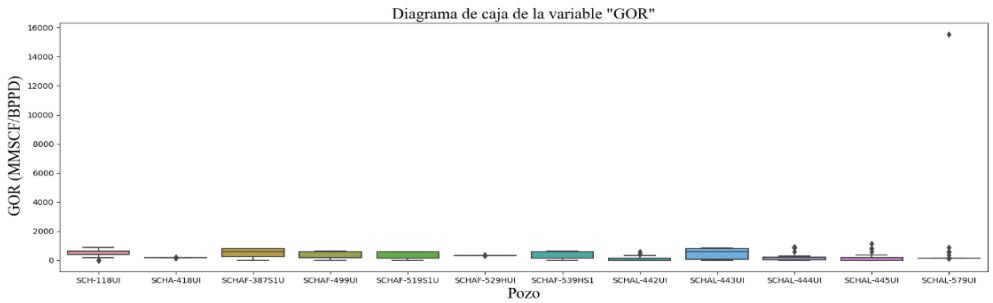


Nota. Información tomada por EP PETROECUADOR (2024)

En la figura 21 se pudo apreciar a simple vista la presencia de valores atípicos en todos los pozos debido a que el corte de Agua o BSW es un valor que no siempre se mantiene constante ni en un mismo rango, esta variable aumenta con el pasar del tiempo.

Figura 22

Diagrama de caja de la variable "GOR" posterior a tratamiento de valores atípicos.

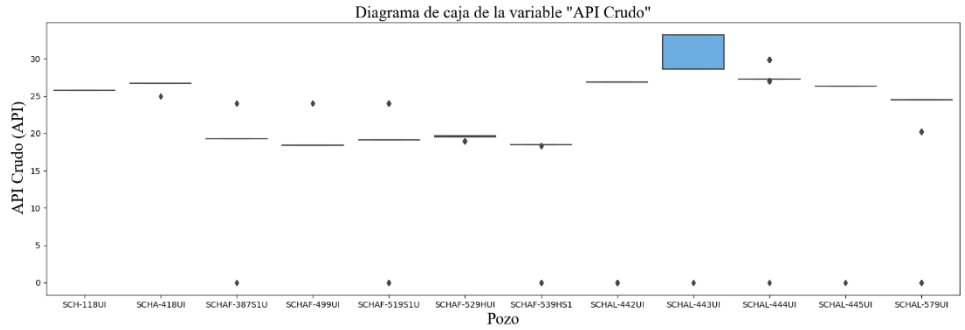


Nota. Información tomada por EP PETROECUADOR (2024).

Según la figura 22, el diagrama de cajas de la variable “GOR” en cada pozo permite inferir que el GOR de los pozos estudiados se encuentra entre valores mayores a 200 y valores menores a 2000 lo que podría describir la clasificación del crudo en esta característica.

Figura 23

Diagrama de caja de la variable “API” posterior a tratamiento de valores atípicos.



Nota. Información tomada por EP PETROECUADOR (2024)

La figura 23 demostró el comportamiento de la Gravedad API, existen pocos valores atípicos que se los puede omitir tomando en cuenta que los intervalos en cada uno de los pozos son cortos, es decir que el valor tiene pequeñas variaciones casi nulas.

Los valores atípicos presentes en cada uno de los diagramas de caja que eran iguales a cero son por la aplicación del método del rango intercuartil que al encontrar un valor fuera de rango lo elimina y la celda queda con valor nulo.

3.3.2. Tratamiento de valores nulos

Al principio de la sección 3.3. se mencionó que no existía la presencia de datos nulos, pero al finalizar la sección 3.3.1. se identificó celdas vacías por el tratamiento aplicado, debido a esto se realizó una imputación de datos

utilizando el algoritmo de K Nearest Neighbors (KNN) que reemplaza los datos nulos según el comportamiento de los valores vecinos o cercanos.

Figura 24

Visualización del conjunto de datos previo al proceso de imputación.

.

	Fluido Total	Crudo	AGUA	Corte de Agua	Gas Total	GOR	API Crudo	Salinity
0	782.0	344.0	NaN	56.01	53.32	155.00	26.9	NaN
1	782.0	344.0	NaN	56.01	53.32	155.00	26.9	NaN
2	784.0	345.0	NaN	55.99	53.48	155.01	26.9	NaN
3	784.0	345.0	NaN	55.99	53.48	155.01	26.9	NaN
4	784.0	345.0	NaN	55.99	53.48	155.01	26.9	NaN

Nota. Información tomada de EP PETROECUADOR (2024)

En la figura 24 se pudo apreciar que en la columna de las variables “AGUA” y “Salinity” tienen el valor “NaN” que hace referencia a que esa celda está vacía, es decir, valor nulo.

Figura 25

Visualización del conjunto de datos posterior al proceso de imputación.

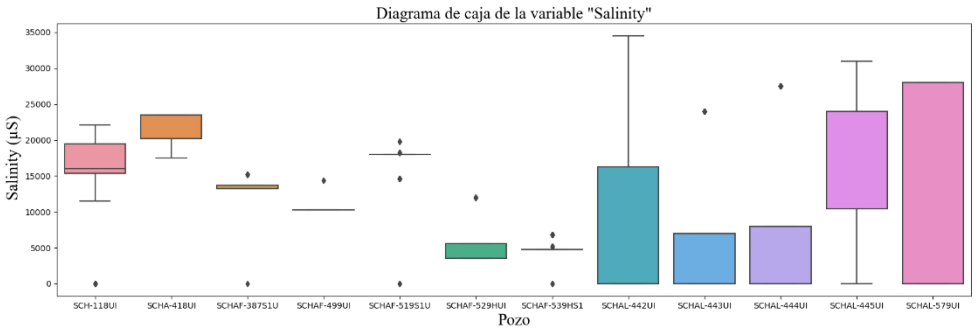
	Fluido Total	Crudo	AGUA	Corte de Agua	Gas Total	GOR	API Crudo	Salinity
0	782.0	344.0	293.5	56.01	53.32	155.00	26.9	22400.0
1	782.0	344.0	293.5	56.01	53.32	155.00	26.9	22400.0
2	784.0	345.0	293.5	55.99	53.48	155.01	26.9	22400.0
3	784.0	345.0	293.5	55.99	53.48	155.01	26.9	22400.0
4	784.0	345.0	293.5	55.99	53.48	155.01	26.9	22400.0

Nota. Información tomada de EP PETROECUADOR (2024)

La figura 25 permitió a simple vista ver que las columnas de “AGUA” y “Salinity” ya no tienen celdas nulas, al contrario, tienen valores específicos que fueron otorgados por el algoritmo KNN con una configuración que solo tomó en cuenta un número de vecinos igual a 2 por cada valor nulo.

Figura 26

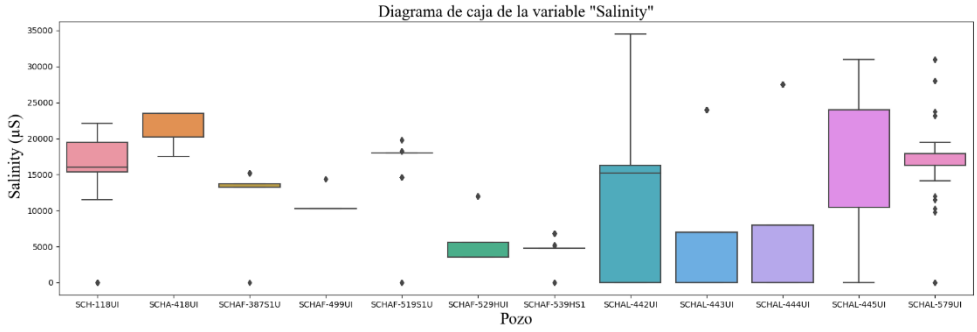
Diagrama de cajas de la variable “Salinity” previo al proceso de imputación.



Nota. Información tomada por EP PETROECUADOR (2024)

Figura 27

Diagrama de cajas de la variable “Salinity” posterior al proceso de imputación.



Nota. Información tomada por EP PETROECUADOR (2024)

Observando la figura 26 y la figura 27 fue posible observar como este proceso de imputación afectó al conjunto de datos en la variable “Salinity” para

reconocer los posibles cambios. De forma específica, se identificó que con el algoritmo aplicado cambió la mediana del pozo SCHAL – 442UI y el rango de valores que tenía el pozo SCHAL-579UI, es decir que este procedimiento ayudará a tener valores más relacionados al conjunto de datos.

3.4. Ingeniería de Características

3.4.1 Creación de nuevas variables

En la sección 3.1.2 se mencionó la creación de una nueva variable con el nombre “Influencia” para especificar si un pozo estaba o no con proceso de inyección de agua. Previo a la sección 3.5 fue necesario añadir otra variable categórica “pozo_categ” que permitió etiquetar cada pozo con un número único, con el fin de que los algoritmos puedan identificar que dentro del conjunto de datos existían pozos diferentes. No fue posible utilizar la columna con el nombre de los pozos debido a que es una característica de tipo objeto o texto y era necesario que fuera de tipo numérico ya sea entero o decimal.

Tabla 6

Etiqueta numérica de los pozos.

Pozo	**Etiqueta**
SCHAL-442UI	0
SCHAL-443UI	1
SCHAL-444UI	2
SCHAL-445UI	3
SCHAL-579UI	4
SCH-118UI	5
SCHA-418UI	6
SCHAF-387S1U	7

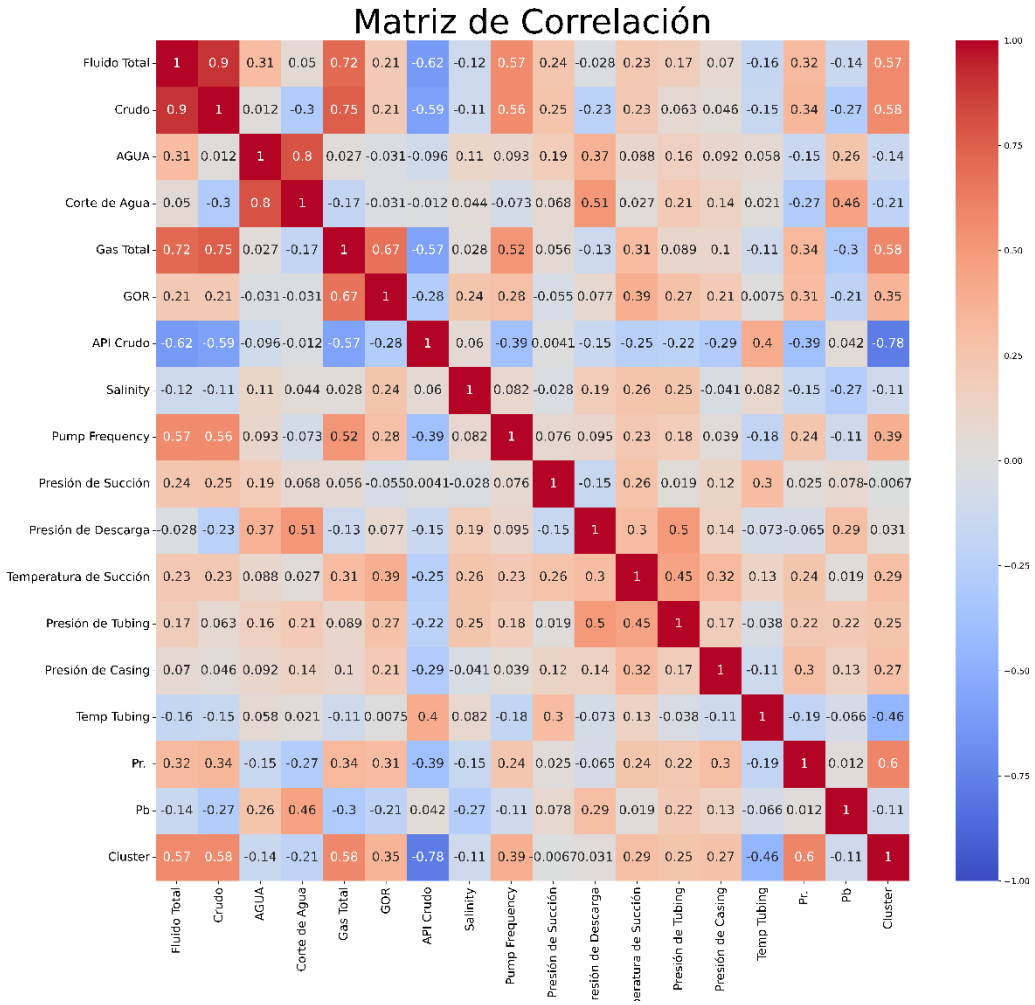
SCHAF-499UI	8
SCHAF-519S1U	9
SCHAF-529HUI	10
SCHAF-539HS1	11

Nota. Información tomada por EP PETROECUADOR (2024)

3.4.2 Impacto de características y Matriz de correlación

Figura 28

Matriz de correlación con datos procesados



Nota. Información tomada por EP PETROECUADOR (2024)

En la figura 28 la matriz de correlación permitió realizar una selección manual de las variables para entrenar el modelo de predicción, lo principal fue revisar que las variables seleccionadas no estuvieran fuertemente correlacionadas con la variable objetivo, es decir, la variable “Crudo”. Las primeras variables descartadas fueron las que tenían una correlación mayor a 0.7 entre sí como los fluidos totales y cantidad de gas producido. Luego, se realizó una verificación de multicolinealidad entre variables independientes para descartar aquellas que tuvieran un valor de relación mayor a 0.8; bajo este criterio no se podía seleccionar la variable “AGUA” y “Corte de Agua” por lo que solo se seleccionó la segunda.

Al finalizar el impacto de características del conjunto de datos, las variables independientes seleccionadas fueron: “Corte de Agua”, "GOR","API Crudo", "Salinity", "Pb", "Pump Frequency. Sin embargo, se tenían tres variables categóricas: “Nombre de pozo”, “pozo_categ” e "Influencia", y una variable de tiempo “Fecha” que ayudaron al modelo a tener un mejor contexto de los datos.

3.5. Entrenamiento de los modelos

3.5.1 Partición del conjunto de datos.

El conjunto de datos se separó en dos intervalos de tiempo, para el entrenamiento y evaluación se utilizaron los datos en un intervalo de tiempo desde el 2014 hasta el 2022 en una relación 80-20 donde el 80% de los datos se ubicaban entre 2014 y 2021 y el 20% en el año 2022. La evaluación con métricas estadísticas se realizó con los datos del 2022, a diferencia de la

comparativa con métodos convencionales donde se utilizaron los datos disponibles desde el 2023 hasta el 2024.

3.5.2 Evaluación de modelos de regresión por K fold cross validation

Tabla 7

Validación de modelos de regresión.

Modelo	RMSE	Desv std
RF	36.31	11.76
DT	43.80	20.04
XGBOOST	30.4	11.04
Bagging Tree	34.52	19.31

Nota. Resultados por modelos de creación propia.

En la tabla 6 se puede apreciar que el modelo con mejor resultado es XGBOOST que fue configurado con un valor de estimaciones igual a 1000, es decir, el número de árboles de decisión a construidos con una tasa de aprendizaje de 0.01 y un máximo valor de 9 niveles en cada árbol de decisión. Este modelo de regresión tuvo un RMSE bajo considerando que el rango de valores a predecir se encuentra en una escala 0-1000, por lo tanto, no es significativo en comparación a los valores reales de predicción, además la mínima desviación estándar de 11.06 demostró que los resultados del modelo no varían mucho. Ambas métricas indicaron que existe una tendencia sistemática con un sesgo menor en las predicciones de XGBOOST.

3.6. Validación del modelo

3.6.1 Validación con métricas estadísticas

La evaluación con métricas estadísticas se realizó con el fin de comparar la eficiencia del modelo de regresión con mejores resultados en la validación y

un modelo de series temporales (SARIMAX) el cual fue configurado con la variable “Fecha” como índice del conjunto de datos para que el algoritmo pueda detectar el tiempo de cada celda de la variable objetivo y una modificación de hiperparámetros donde se detectó que los valores tenían un mejor comportamiento en las series de primer orden y no contaban con una frecuencia o estacionalidad definida. Así mismo, se utilizaron las características “pozo_categ”, “Influencia” y “Pump Frequency” como variables exógenas y optimizar el ajuste del modelo.

Tabla 8

Validación del modelo seleccionado con métricas estadísticas.

Modelo	**RMSE**	**R2**
RF	26.32	0.98
SARIMAX	------	0.72

Nota. Resultados por modelos de creación propia.

En la tabla 7 fue posible identificar que el RMSE obtenido por el método k fold cross validation es diferente que, aplicando las métricas estadísticas con predicciones de la base de datos de evaluación del modelo, pero no están tan alejados. Por otro lado, el R2 tuvo un valor de 0.98 aproximadamente demostrando la fiabilidad del modelo elegido, ya que los modelos con resultados cercanos a 1 en este parámetro tienen un ajuste correcto y mejor precisión en sus predicciones.

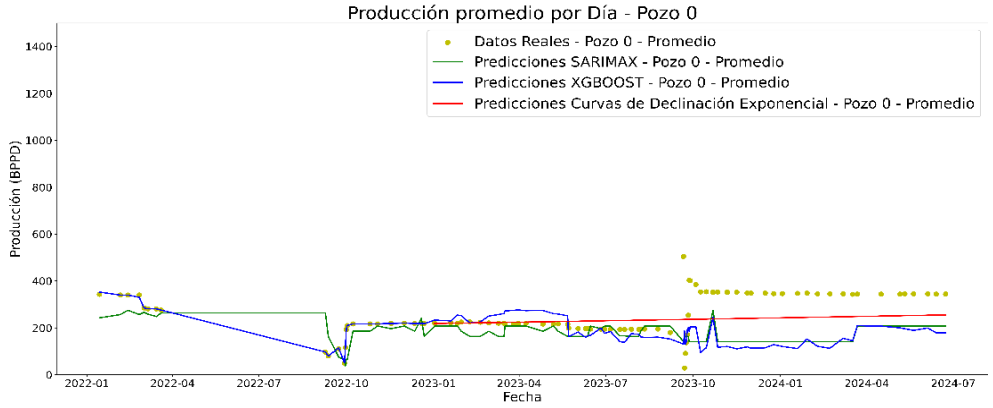
3.6.2. Validación con método convencional

Con la finalidad de conocer el comportamiento de los modelos XGBOOST y SARIMAX en períodos de tiempo más largos y comprobar su eficiencia con

respecto a los métodos convencionales y datos reales de la producción diaria de los pozos del BLOQUE 60 – Campo Sacha se realizaron predicciones hasta el año 2024 como se lo mencionó en la **sección 3.5.1;** A continuación, se pueden apreciar los gráficos de dispersión y línea que permitieron el análisis y aplicación de cada uno de los algoritmos.

Figura 29

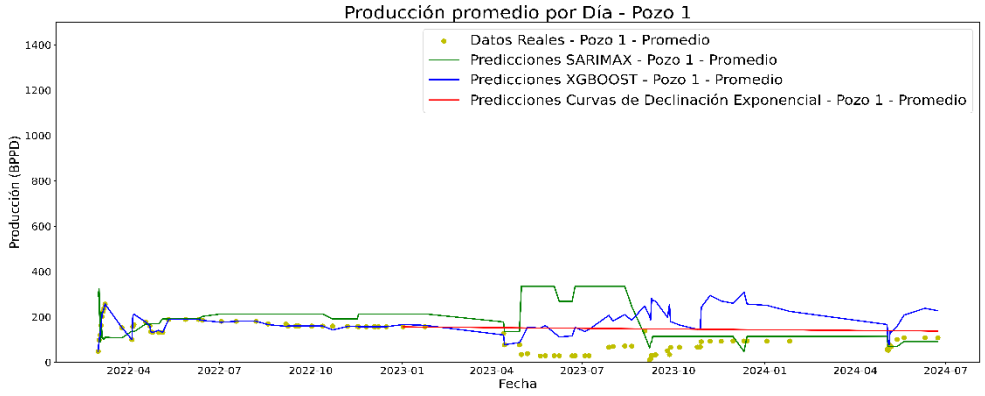
Comparación de predicciones entre método convencional, modelos predictivos y datos reales en el Pozo 0 - SCHAL-423UI.



Nota. Datos reales: gráfico de dispersión, datos predichos: gráfico de línea.

Figura 30

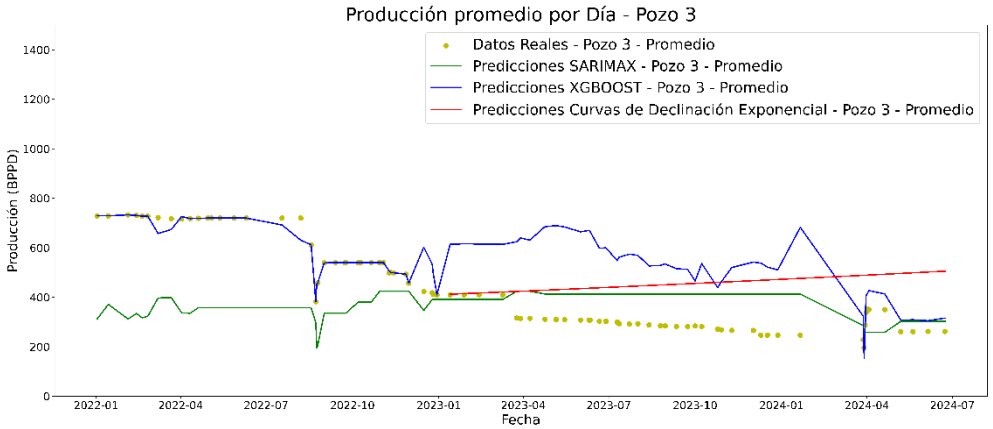
Comparación de predicciones entre método convencional, modelos predictivos y datos reales en el Pozo 1 - SCHAL-443UI .



Nota. Datos reales: gráfico de dispersión, datos predichos: gráfico de línea.

Figura 31

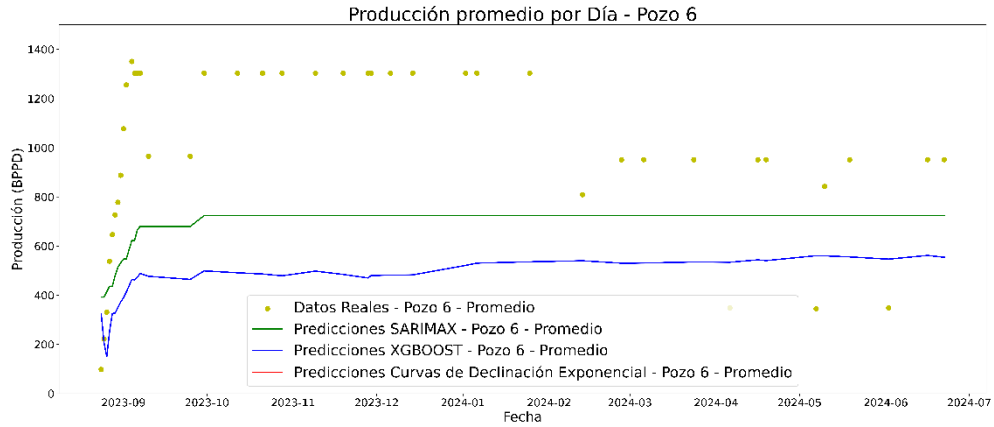
Comparación de predicciones entre método convencional, modelos predictivos y datos reales en el Pozo 3 - SCHAL-445UI .



Nota. Datos reales: gráfico de dispersión, datos predichos: gráfico de línea.

Figura 32

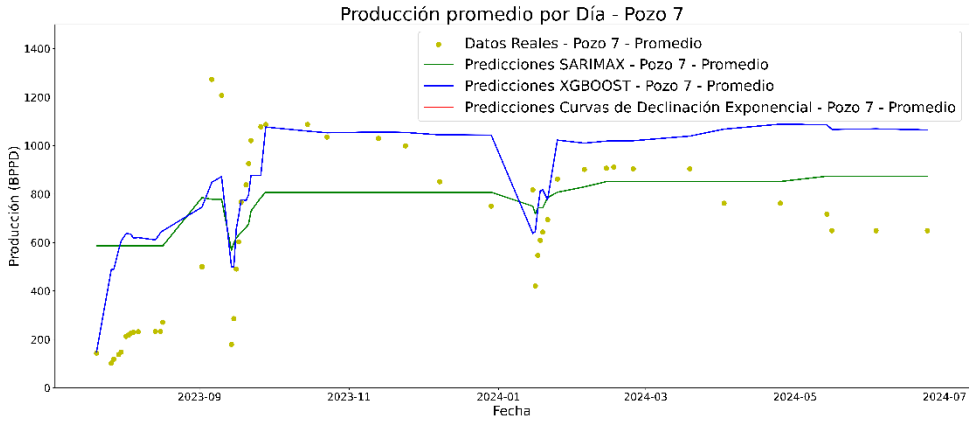
Comparación de predicciones entre método convencional, modelos predictivos y datos reales en el Pozo 6 - SCHA-418UI.



Nota. Datos reales: gráfico de dispersión, datos predichos: gráfico de línea.

Figura 33

Comparación de predicciones entre método convencional, modelos predictivos y datos reales en el Pozo 7 - SCHAF-387S1U .



Nota. Datos reales: gráfico de dispersión, datos predichos: gráfico de línea.

A partir de la figura 30, se pudo analizar el comportamiento de la declinación exponencial y los modelos predictivos con respecto a los datos reales en el periodo comprendido desde el 2022 al 2024. En la comparativa de ambos modelos con la declinación exponencial fue posible observar que la mayor diferencia se encontraba en el año 2023 donde los modelos predictivos tuvieron un mayor acercamiento con sus estimaciones que el método convencional. En 2024 se observó que ambos métodos tanto algoritmos de predicción como método convencional tuvieron predicciones casi similares pero el comportamiento y tendencia de los datos a través del tiempo fue diferente.

En la revisión individual de cada punto predicho en el año 2022 se pudo confirmar la eficiencia del 98% del modelo de regresión XGBOOS y 72% de SARIMAX mientras que en el periodo 2023-2024, el modelo de series temporales obtuvo un 53% de eficiencia mayor a XGBOOST siendo ambas opciones recomendables para estudiar el comportamiento de la producción de petróleo en el campo Sacha.

El modelo SARIMAX es la opción recomendable en predicciones para periodos de tiempo más largos en el BLOQUE 60 debido a la eficiencia conseguida con un menor número de variables. Sin embargo, dado los resultados de XGBOOST, su utilidad se vió enfocada en análisis de sensibilidad para la producción de crudo actual.

3.7. Aplicativo Web

La implementación de Streamlit permitó crear una interfaz optimizada para la interacción amigable con los usuarios. En el aplicativo web se desarrollaron opciones para cargar diferentes conjuntos de datos con cierta especificación; luego se añadieron las opciones de análisis de sensibilidad donde la interfaz trabaja con el modelo XGBOOST y predicción a largo plazo en la cual se utiliza el modelo SARIMAX; para cada una de las acciones configuradas el aplicativo se programó para que le permita al usuario visualizar los resultados con gráficos de dispersión y una tabla de datos para conocer los valores de forma puntual.

Capítulo 4

4.1 Conclusiones y recomendaciones

4.1.1 Conclusiones

- El preprocesamiento aplicado al conjunto de datos permitió que la calidad de la información adquirida aumente a través de procesos de aprendizaje no supervisado, tratamiento de datos fuera de rango e imputación de datos faltantes. Disponer de un mayor rango de confianza optimizó la eficiencia y tendencia de los algoritmos en comparación a los datos reales.
- Los modelos de regresión obtuvieron mejores resultados que el modelo SARIMAX con un porcentaje mayor al 20% en la evaluación con métricas estadísticas tomando como referencia R^2 en el año 2022. Este resultado demuestra que los modelos de series temporales se ajustan mejor a intervalos de tiempo más largos mientras que los modelos de regresión ayudan en los análisis de sensibilidad para conocer como la variación de ciertos factores afecta a la producción actual de petróleo.
- SARIMAX obtuvo una mejor predicción con respecto a las pruebas de predicción en el período 2023-2024. Su eficacia mayor al 50% en comparación con XGBOOST y el método convencional demostró que es una opción viable para la estimación del comportamiento de los pozos en el campo Sacha, además de utilizar menos variables que los modelos de regresión.
- La creación de un aplicativo usando Streamlit permitió que la aplicación de modelos de predicción pueda ser utilizada por diferentes usuarios debido a su interfaz amigable e instrucciones específicas.

4.1.2 Recomendaciones

Al finalizar el proyecto, se hacen las siguientes recomendaciones para la optimización y replicabilidad:

- El conjunto de datos debería tener una frecuencia establecida para poder determinar la estacionalidad de la información. Ambos parámetros pueden mejorar el rendimiento en los modelos de series temporales.
- En las etapas de preprocesamiento de los datos se podría aplicar otros métodos para el tratamiento de valores fuera de rango o nulos con el fin de mejorar la calidad del conjunto de datos utilizado para entrenar los modelos predictivos.
- Se podría trabajar con un modelo de redes neuronales aplicando los mismos procesos de limpieza y tratamiento previo al entrenamiento y validación del mismo para comparar el rendimiento con los modelos propuestos en este proyecto.

Bibliografía

- Ayauca, A., Goyburo, C., Medina, A., Mendez, J. I., Guerrero, A., Saltos, R., Priscila, V. A., & Gutierrez, L. (2023, August 11). *ADVANCED TREATMENT OF FORMATION WATERS FROM THE OIL INDUSTRY TO MITIGATE CORROSION*. https://laccei.org/LACCEI2023-BuenosAires/meta/FP1479.html
- Bangert, P. (2021). Data Science, Statistics, and Time-Series. En Elsevier eBooks (pp. 13-36). https://doi.org/10.1016/b978-0-12-820714-7.00002-9
- Department of Chemical and Petroleum Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada Energies 2023, 16(3), 1392; https://doi.org/10.3390/en16031392
- European Knowledge Center for Information Technology. (2023, 28 noviembre). Analítica de datos (data analytics): ¿Cómo tener una visión global de los datos y las tendencias? TIC Portal. Recuperado en 4 de junio de 2024 de https://www.ticportal.es/glosario-tic/analitica-datos
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. «O'Reilly Media, Inc.»
- Ilyas, I. F., & Chu, X. (2019). Data cleaning. Morgan & Claypool.
- International Petroleum Industry Environmental Conservation Association. (2021). Water Management in the Oil and Gas Industry. https://www.ipieca.org/work/nature/water-management
- Larson, M. G. (2006). Descriptive Statistics and Graphical Displays. Circulation, 114(1), 76-81.

- Lin, Y., Guthrie, G., Coblentz, D., Wang, S., & Thiagarajan, J. (2017). Towards real-time geologic feature detection from seismic measurements using a randomized machine-learning algorithm. OnePetro. https://onepetro.org/SEGAM/proceedings-abstract/SEG17/All-SEG17/SEG-2017-17775517/102664
- Mensah, A. O. (2023). Real-Time Lithology Prediction While Drilling Using Machine Learning Algorithms: A Web Application Based Solution. ATCE. https://doi.org/10.2118/217479-stu
- Mensah, A. O. (2023). Real-Time Lithology Prediction While Drilling Using Machine Learning Algorithms: A Web Application Based Solution. ATCE. https://doi.org/10.2118/217479-stu
- Sayani, J.K.S., Lal, B. (2023). Machine Learning in Oil and Gas Industry. In: Lal, B., Bavoh, C.B., Sahith Sayani, J.K. (eds) Machine Learning and Flow Assurance in Oil and Gas Production. Springer, Cham. https://doi.org/10.1007/978-3-031-24231-1_2
- Singh, V., & Jain, A. (2018). Machine learning for reservoir characterization: A review. Journal of Petroleum Exploration and Production, 6(2), 279-292.
- Thabet, S., Zidan, H. M., Elhadidy, A., Helmy, A., Yehia, T., Elnaggar, H., & Elshiekh, M. (2024). Prediction of Total Skin Factor in Perforated Wells Using Models Powered by Deep Learning and Machine Learning. GOTECH. https://doi.org/10.2118/219187-ms
- U.S. Environmental Protection Agency. (2021). Groundwater and Drinking Water. https://cfpub.epa.gov/ncea/hfstudy/recordisplay.cfm?deid=332990
- VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.".

- Walker, M. (2020). Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights. Packt Publishing Ltd.

Anexos

https://github.com/ESPOL-FICT-PETROLEOS/OilProductionML.git

Este link permite visualizar en la plataforma GitHub el código realizado con el lenguaje de programación Python.

Escuela Superior Politécnica del Litoral
OBJETIVOS DE DESARROLLO SOSTENIBLE
La ESPOL promueve los Objetivos de Desarrollo Sostenible

Roberto Antonio Saltos Ortega
robsorte@espol.edu.ec

PREDICCIÓN DE LA PRODUCCIÓN DE POZOS DE PETRÓLEO DEL CAMPO SACHA UTILIZANDO ANALÍTICA DE DATOS.

PROBLEMA

Se abordan dos problemas principales de la industria petrolera: la excesiva cantidad de agua producida por los pozos petrolíferos y lo difícil que resulta predecir su caudal a lo largo del tiempo con los métodos convencionales además sesgo que existe en estos métodos. Por otro lado, la dificultad para estimar o predecir el potencial activo de un yacimiento.

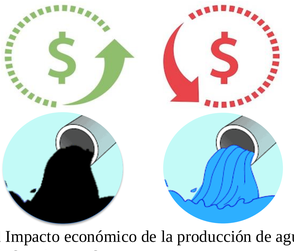


Fig. 1 Impacto económico de la producción de agua en la industria petrolera.

OBJETIVO GENERAL

Estimar la cantidad de petróleo que producirá un pozo en la arena U inferior antes de los disparos usando la información histórica de producción para el análisis de los futuros caudales de crudo y riesgo de inversión.

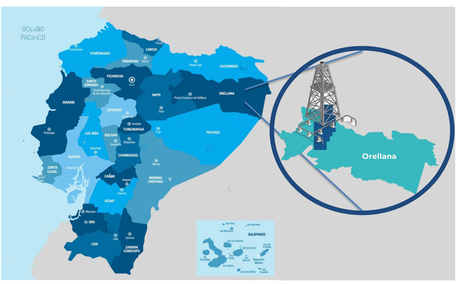


Fig. 2 Área de estudio, BLOQUE 60 - Campo de Sacha.

PROPUESTA

Encontrar la relación que existe entre las diversas variables que se estudian diariamente en campo aplicando machine learning que comparado con estudios como curvas de declinación tienen un menor uso de recursos siempre y cuando el porcentaje de eficiencia sea lo más cercano al 100%, con el fin de proporcionar una herramienta que pueda predecir el comportamiento productivo de un yacimiento sabiendo que las formaciones no tienen una tendencia lineal y su naturaleza no es fácil de interpretar.

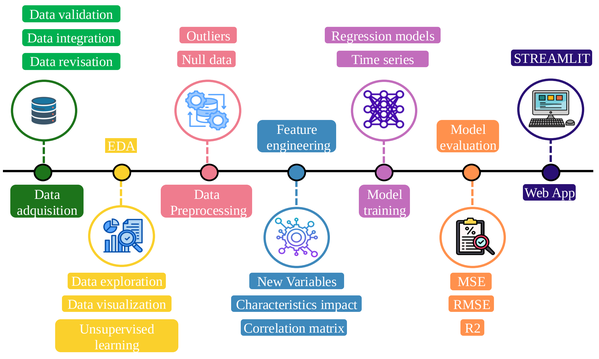


Fig. 3 Esquema de trabajo.

RESULTADOS

TRATAMIENTO DE DATOS Y SELECCIÓN DE VARIABLES

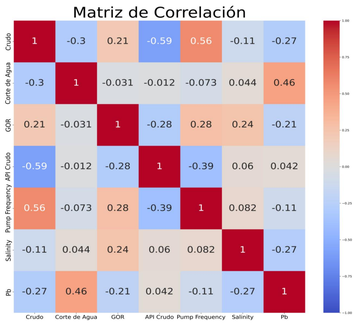


Fig. 4 Matriz de correlación entre variables independientes y variable objetivo.

EVALUACIÓN DE MODELOS

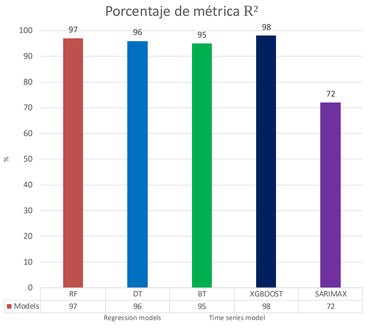


Fig. 5 Evaluación de modelos con R².

VALIDACIÓN DE MODELOS

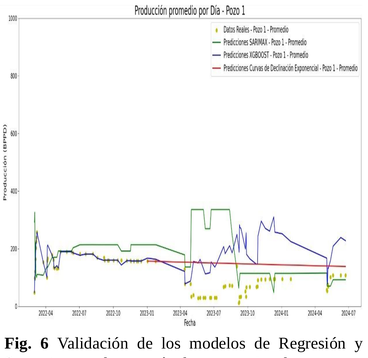


Fig. 6 Validación de los modelos de Regresión y Series temporales vs método convencional.

Posterior al **análisis y pre-procesamiento** se hizo una reducción a 6 variables independientes **cuantitativas**, 3 variables independientes **cualitativas**, 1 variable independiente de tiempo y la variable objetivo o **producción de crudo.**

Los **modelos de regresión** y de **series temporales** se entrenaron y validaron con una proporción **80-20** respectivamente en el periodo comprendido entre 2014 y 2022.
Un porcentaje del **98%** según la métrica R² demuestra la eficacia de **XGBOOST** con un ajuste moderado.

Las predicciones de **XGBOOST** y **SARIMAX** se compararon con los resultados de las **ecuaciones ARPS** - modelo exponencial y las pruebas de producción del campo Sacha.
Las series temporales obtuvieron un rendimiento superior al 50% en comparación con los demás modelos.

CONCLUSIONES

- El preprocesamiento aplicado al conjunto de datos permitió disponer de un mayor **rango de confianza** para optimizar los modelos predictivos.
- Los modelos de regresión obtuvieron mejores resultados en las secciones de validación que el modelo SARIMAX. Este resultado demuestra que **los modelos de series temporales se ajustan mejor a intervalos de tiempo más largos.**
- **SARIMAX** obtuvo una mejor predicción con respecto a las pruebas de predicción acumulativa. Su eficacia mayor al 50% en comparación con **XGBOOST** y el método convencional demostró ser una opción viable, además de utilizar menos variables que los modelos de regresión.

INGE-2439
Código Proyecto

Facultad de Ingeniería en Ciencias de la Tierra

12 PRODUCCIÓN Y CONSUMO RESPONSABLES