PetroRenova
REVISTA CIENTÍFICA DE LA ENERGÍA
Vol. 1, núm. 1, abril-junio, 2025 INDEXED

Predicción de la concentración de hidrógeno en un reactor de polimerización basado en machine learning

Hydrogen concentration prediction in a polymerization reactor based on machine learning

DOI: https://doi.org/10.5281/zenodo.15643200

Recibido: 2025-01-23 **Aceptado**: 2025-03-08

Sabino Montero, Karla Valentina[1]
Correo: karlavsm6@gmail.com
Orcid: https://orcid.org/0009-0002-5853-4550

Noguera Hernández, José Ricardo[2]
Correo: josericardo95@hotmail.com
Orcid: https://orcid.org/0009-0008-1636-7823

Resumen

Se modeló la concentración de hidrógeno en un reactor de polimerización a través de Machine Learning en Python. Se emplearon métodos de preprocesamiento de datos (variables rezagadas, limpieza y detección de valores atípicos). Se aplicaron técnicas estadísticas para la visualización de correlación de variables a través de Heatmap. Se ajustaron los modelos Linear Regression, ARIMAX y GBR, obteniendo correlaciones de 0.7950, 0.6722 y 0.6395 respectivamente. Se seleccionó el modelo predictivo de Linear Regression por su mayor correlación, y se obtuvo una mejora del 14.96 % mediante agrupación de observaciones. Para el análisis de sensibilidad, se obtuvo un valor de predicción de la concentración de 0.65 y 0.44 para valores de 3 y 2 kg/h de flujo de hidrógeno crudo, respectivamente, con relación positiva en la variación del mismo. Los resultados confirman la efectividad del aprendizaje automático en el análisis predictivo de procesos industriales.

Palabras clave: python, aprendizaje automático, análisis predictivo, inteligencia artificial.

Abstract

The concentration of hydrogen in a polymerization reactor was modeled using Machine Learning in Python. Data preprocessing methods (lagged variables, cleaning, and outlier detection) were employed. Statistical techniques were applied for visualization of variable correlation using Heatmap. Linear Regression, ARIMAX and GBR models were adjusted, obtaining correlations of 0.7950, 0.6722 and 0.6395 respectively. Linear Regression predictive model was selected for its higher correlation, and a 14.96% improvement was obtained through observation grouping. For sensitivity analysis, concentration prediction was achieved with values of 0.65 and 0.44 for 3 and 2 kg/h raw hydrogen flow, respectively, showing a positive relationship with its variation. The

[1] Ingeniero Químico, Altamar Trading, C.A., Universidad Rafael Urdaneta. Maracaibo, Venezuela.
[2] Ingeniero Químico, Polipropileno de Venezuela, Propilven S.A., La Universidad del Zulia. Maracaibo, Venezuela.

Este artículo de acceso abierto está bajo **Licencia Creative Commons Attribution 4.0 Internacional**

results confirm the effectiveness of Machine Learning in the predictive analysis of industrial processes.

Keywords: python, machine learning, predictive analysis, artificial intelligence

Introducción

La Inteligencia Artificial (IA) es un campo de la informática que se ha convertido en un pilar fundamental para transformar grandes volúmenes de datos en información valiosa. Un subconjunto de la IA es el aprendizaje automático, que permite procesar grandes cantidades de datos de entrada para resolver problemas de modelado, lo que ofrece una visión de posibles futuros (Dubravova et al., 2024). A nivel mundial, el Machine Learning ha proporcionado diferentes técnicas o algoritmos para predecir situaciones de acuerdo con grandes cantidades de información que, a través de un buen procesamiento y filtrado de datos, pueden generar predicciones muy efectivas (Forero-Corba y Negre, 2024). Los modelos predictivos son herramientas estadísticas diseñadas para identificar patrones y relaciones, proporcionando la capacidad de anticipar comportamientos futuros en función del entrenamiento de datos históricos. Estas capacidades predictivas son fundamentales para la toma de decisiones en múltiples disciplinas.

En este contexto, la versatilidad y simplicidad del lenguaje de programación Python, son de utilidad para el procesamiento de datos y el ajuste de modelos predictivos complejos. Las extensas bibliotecas de Python, como NumPy, SciPy y pandas, proporcionan recursos poderosos para el análisis de datos, la visualización y el Machine Learning, por lo que se convierte en una herramienta invaluable para llevar a cabo los modelos predictivos (Kovac et al., 2024). Entre ellos se destacan el modelo de Regresión Lineal, esencial para examinar y modelar relaciones lineales entre variables; ARIMAX, especializado en el análisis de series temporales; por último, Gradient Boosting Regressor, un algoritmo de aprendizaje supervisado que construye árboles de decisión secuenciales.

A pesar de las dificultades técnicas y económicas para obtener suficientes datos experimentales (Ngige et al., 2022), la incorporación de tecnologías de inteligencia artificial, como el aprendizaje automático, podría ayudar a las empresas a no solo optimizar la eficiencia operativa y disminuir costos, sino también prever fallos, asegurar la seguridad en las plantas y mejorar la calidad de los productos. En base a esto, el presente estudio tuvo como objetivo el

modelado dinámico de la concentración de hidrógeno en un reactor de polimerización empleando técnicas de Machine Learning, mediante el procesamiento de datos históricos, y su posterior empleo para el entrenamiento de los modelos, generación de predicciones y evaluación de las métricas relevantes obtenidas con respecto a los valores reportados por la literatura.

1. Fundamentos teóricos

1.1. Bases teóricas

Machine Learning

El aprendizaje automático es una rama particular de la inteligencia artificial que enseña a una máquina cómo aprender, mientras que la Inteligencia Artificial (IA) es la ciencia general que busca emular las habilidades humanas. Un método de IA, llamado aprendizaje automático, enseña a las computadoras a aprender de sus experiencias pasadas. Los algoritmos de aprendizaje automático no dependen de una ecuación predeterminada como modelo, sino que "aprenden" información directamente de los datos utilizando técnicas computacionales. A medida que aumenta la cantidad de ejemplos de aprendizaje, los algoritmos mejoran adaptativamente en lo que hacen. Este documento proporciona una visión general del campo, así como una variedad de enfoques de ML, incluyendo el aprendizaje supervisado, no supervisado y por refuerzo, y varios lenguajes utilizados para el aprendizaje automático (Shaveta, 2023).

Modelo de regresión lineal (Linear regression)

De acuerdo a Qu (2024), la regresión lineal es un método estadístico utilizado para establecer una relación lineal entre una variable independiente X y una variable dependiente Y. El objetivo es encontrar una función lineal óptima, es decir, determinar un conjunto de coeficientes (pesos) de tal manera que la función pueda predecir el valor de la variable dependiente con la mayor precisión posible. El objetivo principal de los algoritmos de regresión lineal es encontrar la mejor estimación de parámetros, de manera que la diferencia entre el valor predicho por el modelo y los datos reales sea mínima. Cuando existen múltiples factores que afectan a la variable dependiente, se necesita un modelo de regresión lineal múltiple.

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

Modelo ARIMA con variables exógenas (ARIMAX)

Hyndman y Athanasopoulos (2021) definen los modelos ARIMA (AutoRegressive Integrated Moving Average) como una metodología estadística para pronosticar series temporales univariantes, combinando tres componentes esenciales: autorregresivo (AR), integrado (I) y media móvil (MA). De acuerdo a Alharbi y Csala (2022), el modelo ARIMAX es una evolución del ARIMA que emplea series temporales multivariadas para predecir la variable dependiente. A diferencia del ARIMA tradicional, incorpora múltiples series temporales como variables exógenas. Su diseño específico para series temporales distingue al ARIMAX de los modelos de aprendizaje supervisado, ya que considera la secuencia de las entradas como un factor crucial.

Modelo Gradient Boosting Regressor (GBR)

El Modelo Gradient Boosting Regressor emplea un enfoque de aprendizaje en conjunto, donde se construyen modelos de predicción robustos mediante la combinación de múltiples árboles de regresión individuales, conocidos como aprendices débiles. Este tipo de algoritmo disminuye la tasa de error de estos aprendices débiles (regresores o clasificadores). Los aprendices débiles son aquellos que presentan un alto sesgo hacia los datos de entrenamiento, con baja varianza y regularización, y cuyas predicciones solo muestran una ligera mejora en comparación con conjeturas aleatorias. En general, los algoritmos de impulso (boosting) constan de tres elementos clave: un modelo aditivo, aprendices débiles y una función de pérdida. El algoritmo es capaz de modelar relaciones no lineales (Singh et al., 2021).

Coeficiente de Determinación (R^2)

Según Chicco, Warrens y Jurman (2021), el coeficiente de determinación (R^2) se puede interpretar como la proporción de la varianza en la variable dependiente que es explicada o predecible por las variables independientes, es decir, indica qué porcentaje de la variación en la variable que se quiere predecir (la dependiente) puede explicarse por las variables que se utilizan para la predicción (las independientes). Es una métrica clave para evaluar la capacidad de un modelo de regresión para explicar la variabilidad de la variable objetivo. Un R^2 de 1 señala un ajuste perfecto, donde el modelo explica toda la varianza, mientras que un R^2 de 0 implica que el modelo no ofrece mejor predicción que la media de los datos.

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

1.2. Revisión de Antecedentes

Calofir et al. (2024) propuso una metodología innovadora para evaluar el daño sísmico en estructuras de marcos resistentes a momentos mediante el uso de algoritmos de aprendizaje automático que fueron entrenados y probados con un extenso conjunto de datos, generados a través de simulaciones numéricas, para replicar el índice de daño estructural de Park-Ang. Se ajustó el porcentaje de entrenamiento y prueba para optimizar la generalización, evitando tanto el subajuste como el sobreajuste, y utilizando la validación cruzada para asegurar la robustez del modelo. De esta investigación se verificó la metodología seguida, así como el porcentaje de la data empleada para el entrenamiento y las técnicas para evitar el sobreajuste y lograr buena generalización en la predicción.

Por otro lado, en el estudio realizado por Mansi et al. (2023) se propone un modelo de aprendizaje automático supervisado, basado en regresión lineal y redes neuronales artificiales (ANNs), para evaluar la eficiencia de la recuperación mejorada de gas (EGR) mediante inyección de CO2 en yacimientos de lutitas, un proceso complejo controlado por múltiples parámetros. Utilizando un amplio conjunto de datos de simulaciones y experimentos, el modelo buscó predecir el incremento de CH4 recuperado. De este trabajo, se logró verificar tanto la metodología como el empleo del coeficiente de correlación para la evaluación del desempeño de los modelos. Así mismo, Gou et al. (2024) entrenó cuatro modelos de aprendizaje automático (regresión lineal múltiple, árboles de decisión, regresores Adaboost y bagging) para predecir la producción y composición de biocarbón a partir de residuos orgánicos, superando las limitaciones de precisión y coste computacional de los modelos existentes. Entrenados con datos de pruebas de pirólisis, los modelos logran un R2 de hasta 0.96, demostrando una precisión predictiva significativamente superior. Destaca como aporte de investigación las técnicas de preprocesamiento de datos empleados, además de la metodología desarrollada y el empleo del coeficiente de determinación para la evaluación de los modelos.

2. Metodología

La presente investigación se caracteriza por ser de tipo correlacional, al determinar la relación o asociación entre las variables de proceso, y predictiva, por predecir valores futuros de la concentración de hidrógeno basándose en datos históricos. Su diseño es no experimental, cuantitativo y retrospectivo. La población estudiada es clasificada como accesible, referida a los

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

datos históricos de las variables de proceso seleccionadas. La muestra fue seleccionada en un rango de tiempo de 3 días, con intervalos de 5 minutos. Por otro lado, como técnica de recolección de datos se empleó la observación documental, y como instrumentos se utilizaron la Hoja de Excel para la recolección de datos registrados en el programa Uniformance Process Studio, y el lenguaje de programación Python, mediante el editor de código Visual Studio Code, para el procesamiento de los mismos. Esta metodología permitió verificar el rendimiento de los modelos predictivos, garantizando la precisión y relevancia de los resultados obtenidos.

3. Resultados

3.1. Preprocesamiento de datos

Se aplicaron las técnicas de creación de rezagos, limpieza de datos faltantes y detección de valores atípicos, obteniendo finalmente el marco de datos preprocesado con valores representativos del proceso. En la Tabla 1 se observan las variables seleccionadas.

Tabla 1. Variables seleccionadas para el marco de datos preprocesado

Variable	Descripción	Unidad
CH2	Concentración de hidrógeno en el reactor de polimerización	Fracción molar
FH2 CRUDO	Flujo de hidrógeno crudo hacia el reactor	kg/h
FCAT	Flujo de catalizador hacia el reactor	kg/h
FH2 REC	Flujo de hidrógeno recirculado hacia el reactor	kg/h
C PASTA	Concentración de la pasta en el reactor	Fracción molar
CH2 208	Concentración de hidrógeno en el tambor de reciclo	Fracción molar
C H2(i)	Rezago de la concentración de hidrógeno en el reactor para la posición i	Fracción molar

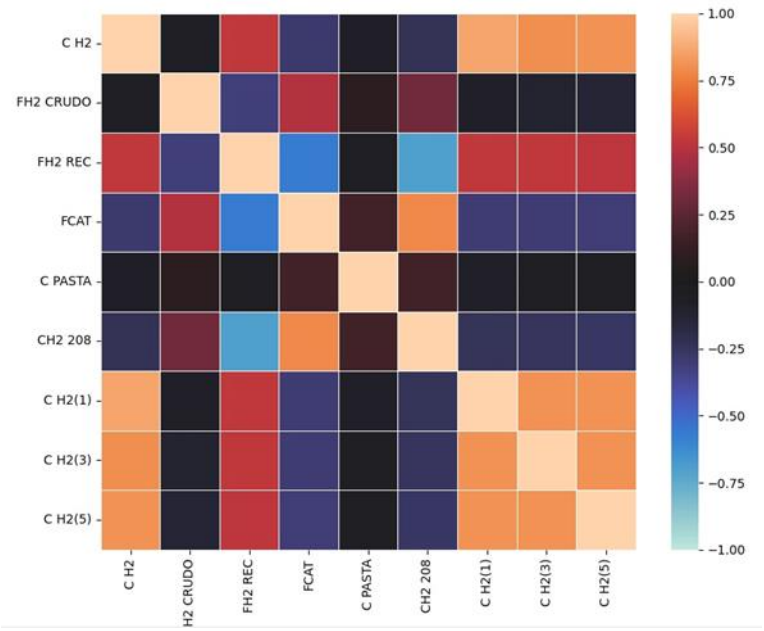
Fuente: elaborado por los autores, datos de la investigación

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

3.2. Análisis de correlación de variables

Se aplicaron técnicas para la verificación de la correlación entre las variables de proceso y los rezagos correspondientes a la concentración de hidrógeno. De esa manera, se generó la matriz de correlación visualizándose en forma de Heatmap. tal como se observa en la Figura 1, de esta forma es posible apreciar visualmente la correlación en una matriz de colores relacionados a escala, donde la correlación positiva perfecta fue representada por el valor 1, la correlación negativa por el -1, y la ausencia de correlación por el 0 (Khodabakhshi y Bijani, 2024).

Figura 1. Mapa de calor (Heatmat) de las variables de proceso



Fuente: elaborado por los autores, datos de la investigación

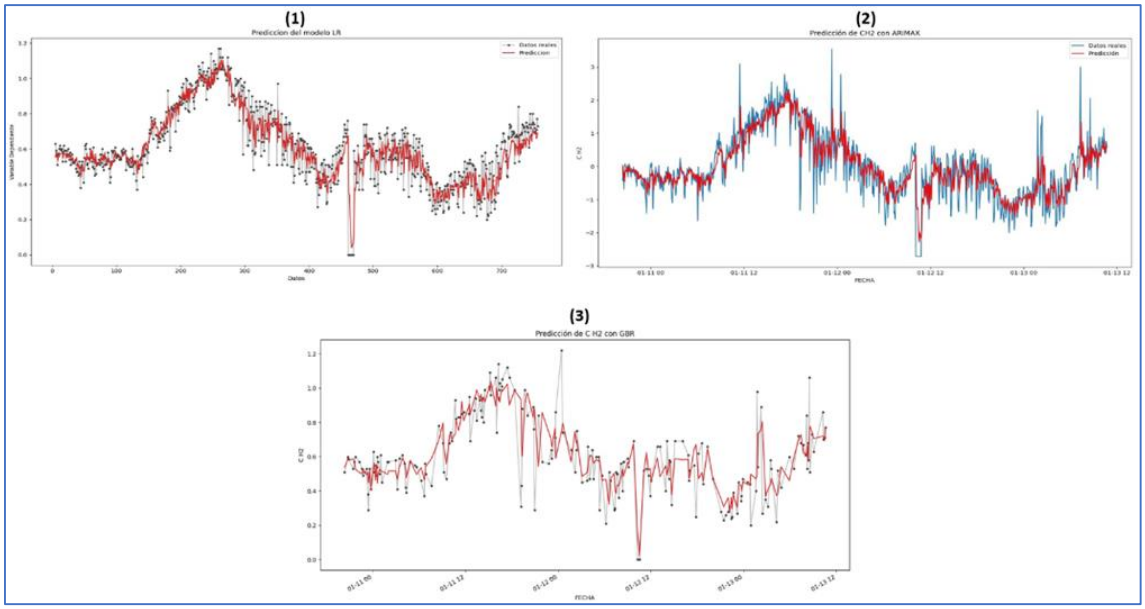
3.3. Selección del modelo de predicción

En la Figura 2, se puede apreciar la predicción de la concentración de hidrógeno comparada con los datos reales en cada uno de los modelos de predicción utilizados. Se destacan cuatro modelos en el análisis: Linear Regression, ARIMAX y Gradient Boosting Regressor (GBR). La visualización demuestra cómo cada modelo aborda la predicción de la concentración de hidrógeno, permitiendo identificar las diferencias y similitudes en sus aproximaciones. Es evidente

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

que el modelo de Linear Regression sigue más de cerca la tendencia de los datos reales, mostrando un ajuste más preciso y una menor desviación en comparación con los otros modelos.

Figura 2. Predicción de la concentración de hidrógeno en el reactor de polimerización basado en el modelo: (1) Linear Regression, (2) ARIMAX, (3) GBR



Fuente: elaborado por los autores, datos de la investigación

Se seleccionó el modelo predictivo de Linear Regression por su correlación de 0.7950, con respecto a las variables de proceso seleccionadas para el marco de datos preprocesado. Esto puede evidenciarse en la Tabla 2, donde se visualiza el coeficiente de determinación obtenido para cada modelo.

Tabla 2. Coeficiente de determinación obtenido para cada modelo predictivo

Modelo predictivo	Coeficiente de determinación
Linear Regression	0.7950
ARIMAX	0.6722
Gradient Boosting Regressor	0.6395

Fuente: elaborado por los autores, datos de la investigación

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

3.4. Ajuste del modelo

A partir del modelo seleccionado, se aplicaron nuevas técnicas con el fin de aumentar la precisión de la predicción. Se realizó entonces el agrupamiento de variables en intervalos de 2 y 3 mediciones temporales. Posterior a ello, se generaron las curvas de predicción y residuales; se visualizaron así mismo los coeficientes e intercepto de la ecuación de regresión. de manera que se lograra incrementar la correlación, verificar la magnitud de los residuales, y corroborar la relación entre variables. En cuanto al agrupamiento de observaciones, se determinó la correlación a diferentes porcentajes de datos seleccionados para el entrenamiento y, de acuerdo al agrupamiento entre cada dos y tres mediciones temporales, tal como se muestra en la Tabla 3. En base a estos resultados, se emplea un porcentaje de entrenamiento de 80% y un agrupamiento de dos mediciones para evitar la reducción excesiva en la cantidad de datos y capacidad de generalización del modelo.

Tabla 3. Coeficientes de determinación del modelo de Linear Regression en agrupamientos entre cada dos y tres mediciones temporales

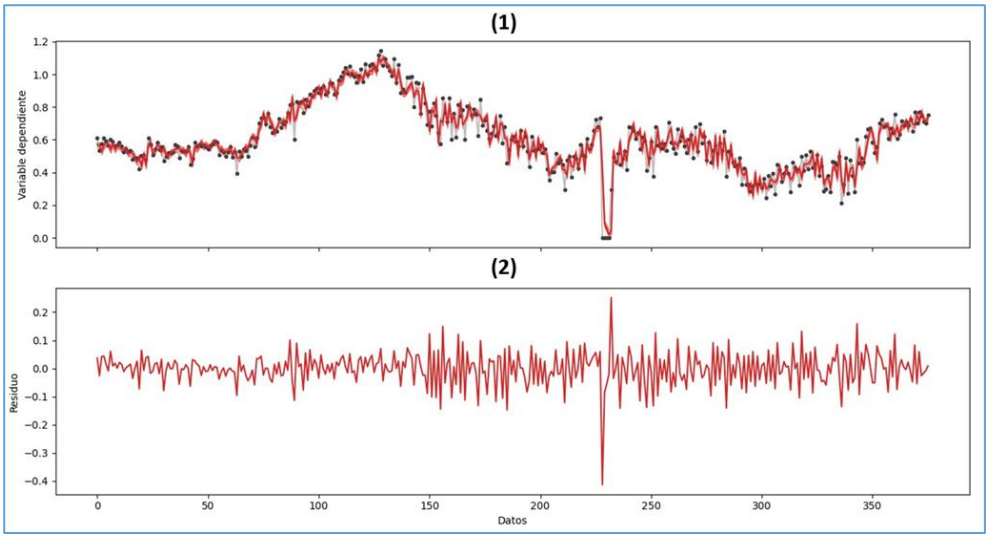
Entrenamiento (%)	r2 (Original)	r2 (2 Mediciones)	r2 (3 Mediciones)
10	0.7591	0.8940	0.9204
20	0.7739	0.8793	0.9237
30	0.7828	0.9077	0.9458
40	0.7824	0.9077	0.9535
50	0.7908	0.9091	0.9536
60	0.7925	0.9078	0.9536
70	0.7871	0.9084	0.9538
80	0.7928	0.9102	0.9542
90	0.7933	0.9106	0.9544

Fuente: elaborado por los autores, datos de la investigación

En la Figura 3 se observa el gráfico de la predicción de los datos reales por el modelo de Linear Regression, junto al respectivo grafico de residuales en la parte inferior, a partir del cual se obtuvo una correlación en el rango de 0.88 – 0.92 para los distintos porcentajes de entrenamiento en conjunto con los cambios realizados.

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

Figura 3. Predicción de la concentración de hidrógeno en el reactor de polimerización basado en el modelo Linear Regression (1) junto al grafico de residuales (2)



Fuente: elaborado por los autores, datos de la investigación

Por otro lado, se determinaron los coeficientes del modelo de regresión lineal multivariable ajustado en la Ec. 1, junto al término de intersección referido a la ordenada en el origen.

$$Y = 0.0616 + 1.0319 \cdot 10{-}2X1 + 1.3326 \cdot 10{-}4X2 - 1.7982 \cdot 10{-}3X3 - 3.9423 \cdot 10{-}5X4 + 3.1326 \cdot 10{-}2X5 + 8.0524 \cdot 10{-}1X6 + 6.4393 \cdot 10{-}2X7 + 8.1450 \cdot 10{-}2X8 \quad \boldsymbol{Ec.\ 1}$$

3.5. Análisis de sensibilidad

En la Tabla 4, se observa el valor inicial y final, además del valor de alcanzado antes y después del cambio en el flujo de hidrógeno crudo con respecto a la predicción de la concentración de hidrógeno, lo cual se registró a partir de los gráficos de la Figura 5. Por otro lado, los valores de referencia basados en la data histórica extraída pueden visualizarse en la Figura 6.

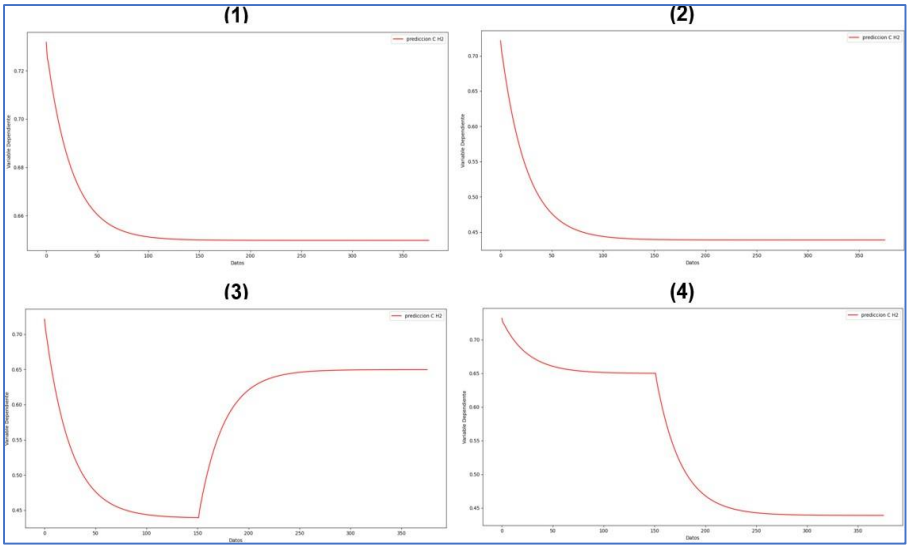
PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

Tabla 4. Flujo de hidrógeno crudo inicial y final ajustado, junto a la predicción de la concentración de hidrógeno antes y después del punto de cambio

Prueba	FH2 Crudo Inicial (kg/h)	FH2 Crudo Final (kg/h)	CH2 Inicial (fracción molar)	CH2 Final (fracción molar)
1	3	3	0.65	0.65
2	2	2	0.44	0.44
3	2	3	0.44	0.65
4	3	2	0.65	0.44

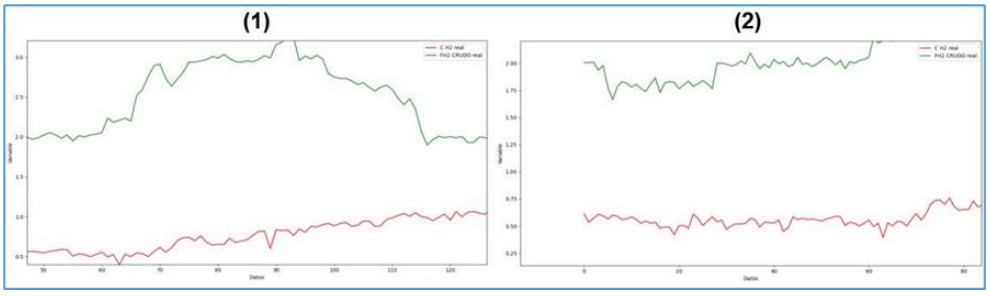
Fuente: elaborado por los autores, datos de la investigación

Figura 5. Análisis de sensibilidad para la prueba: (1) 3 kg/h, (2) 2 kg/h, (3) Incremento, (4) Disminución



Fuente: elaborado por los autores, datos de la investigación

Figura 6. Concentración de hidrógeno en el reactor de polimerización frente al flujo de hidrógeno crudo, datos históricos: (1) 3 kg/h, (2) 2 kg/h



Fuente: elaborado por los autores, datos de la investigación

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

4. Análisis y discusión de los resultados

Se llevaron a cabo técnicas de preprocesamiento en los datos industriales extraídos, fundamentales debido a la naturaleza incompleta, inconsistente o inesperada de estos registros, como evidenció el estudio de Gou et al. (2024), que reportó numerosos valores faltantes en los datos recuperados, resaltando la importancia crítica de este paso para garantizar la limpieza, estructuración adecuada y optimización del rendimiento de modelos de aprendizaje automático. Inicialmente, se abordó la limpieza de datos faltantes y la detección de valores atípicos mediante el método estadístico de puntuación Z, sustituyendo dichos valores por la media calculada en una ventana móvil de referencia para preservar la integridad temporal de los datos. Posteriormente, se generaron variables rezagadas, considerando que los cambios en una variable de perturbación pueden influir en la concentración de hidrógeno del reactor de polimerización con cierto retraso, inherente a la dinámica del proceso.

Por otro lado, el análisis de correlación es uno de los enfoques más utilizados para indicar la relación entre dos o más variables cuantitativas en el modelado predictivo, siendo esenciales para comprender la dependencia entre las variables de entrada/salida, para la selección de los predictores, y para evitar el sobreajuste del modelo (Mansi et al., 2023). De esa manera, se observa en el Heatmap realizado (Figura 1) que existe una fuerte relación positiva de la variable de entrada en relación a su variable de rezago de un paso temporal, seguida del resto de retrasos. Puede destacarse una fuerte relación negativa de la concentración de hidrógeno con respecto al flujo de catalizador hacia el reactor de polimerización, siendo coherente con la relación teórica.

Para la obtención de los tres modelos de predicción para su evaluación, se decidió trabajar con un porcentaje de entrenamiento del 80 % de la data suministrada, a modo de seleccionar el modelo con el mejor comportamiento para el análisis predictivo, tal como fue ejecutado en el estudio de Calofir (2024). Los resultados obtenidos pueden compararse con el estudio realizado por Mansi et al. (2023) sobre la predicción de la recuperación mejorada de CH4 por inyección de CO2, donde se registraron correlaciones de 0.68 para el modelo de regresión lineal, y de 0.778 para el segundo modelo evaluado, referido al de redes neuronales artificiales, mientras que en la presente investigación se logró un mayor coeficiente de determinación para el modelo Linear Regression, (0.7950) siendo seleccionado como el más adecuado para la predicción de la concentración de hidrógeno (Figura 2) en comparación con los modelos de ARIMAX (0.6722) y

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

Gradient Boosting Regressor (0.6395). Esta precisión se evidencia en la menor desviación de las predicciones respecto a los valores reales.

Posteriormente se logró verificar que el agrupamiento de los datos entre cada dos y tres mediciones temporales mejoró la correlación en un 14.55 % y 20.33 % respectivamente en relación al marco de datos original, tomando como referencia el porcentaje de entrenamiento del 80 %. Sin embargo, se decidió trabajar con un agrupamiento de dos mediciones, ya que no es conveniente la supresión masiva de datos para ajustar y entrenar el modelo predictivo, tomando en consideración que la mejora con respecto al agrupamiento entre tres mediciones no es significativa.

En el gráfico de residuales (Figura 3) se visualizó una variación entre la predicción y el comportamiento real en un rango aproximado de -0.2 y 0.2, con un pico máximo en -0.4, punto en el cual coincide con un cambio abrupto en el comportamiento de la data real, producto de un corte repentino del flujo de hidrógeno alimentado. Sin embargo, se aprecia de igual forma que el modelo predictivo reproduce con efectividad el comportamiento esperado ante la perturbación mencionada. También se observa su dificultad de predicción sobre la disminución de la concentración de hidrógeno, ya que en el rango de observaciones del marco de datos agrupado entre 140 y 210, donde se reduce la concentración tanto en la data real como en la línea de predicción, se visualizan mayores valores de error, al igual que en el rango de 330 a 340.

Por otro lado, en la ecuación del modelo de regresión lineal multivariable (Ec. 1), se aprecia una relación positiva de la concentración de hidrógeno en el reactor (Y) con respecto a las variables independientes de flujo de hidrógeno crudo (X1), flujo de hidrógeno recirculado (X2), y la concentración de hidrógeno de reciclo (X5), lo que implica que, al aumentar estas variables de perturbación, debería incrementarse la variable dependiente. De igual forma, el flujo de catalizador (X3) y la concentración de pasta en el reactor (X4), registraron un coeficiente negativo, indicando una relación inversa con respecto a la variable dependiente estudiada. En cuanto a las variables de rezago de la concentración de hidrógeno, se obtuvo que en el retraso número 1, 3 y 5, la relación fue positiva, correspondiente a las variables X6, X7 y X8 respectivamente. Estos resultados fueron los esperados según la relación teórica, confirmando la validez del modelo propuesto y su capacidad para capturar el comportamiento de la data histórica.

En cuanto al análisis de sensibilidad efectuado, se aprecia que para un valor constante de flujo de hidrógeno crudo hacia el reactor de 3 kg/h, se obtuvo una predicción de la concentración

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

de hidrógeno final de 0.65, tal como se observa en la sección (1) de la Figura 5. Por otro lado, para un valor constante de flujo de hidrógeno crudo de 2 kg/h, se obtuvo una predicción de la concentración de hidrógeno final de 0.44, tal como se observa en la sección (2) de la Figura 5. Esto concuerda con los valores registrados de la data histórica, ya que para un valor de flujo de hidrógeno de 3 kg/h, se visualizó en la sección (1) de la Figura 6 una concentración en el rango de 0.6 a 0.8, mientras que para un valor de 2 kg/h, se obtuvo una concentración en el rango de 0.4 a 0.6, lo cual logra apreciarse en la sección (2) de la Figura 6.

Para la tercera prueba realizada, donde se ejecutó un cambio del flujo de hidrógeno fresco desde 2 hasta 3 kg/h, se observó en la sección (3) de la Figura 5 una estabilización de la concentración de hidrógeno en un valor de 0.44, y luego de pasar el punto de cambio (observado en el eje de las abscisas), se incrementó hasta estabilizarse en el valor de 0.65, lo cual concuerda con los valores reportados anteriormente, apreciando una relación positiva entre ambas variables. Para la última prueba, donde se ejecutó una reducción del flujo de hidrógeno fresco en el mismo punto desde 3 hasta 2 kg/h, se observó en la sección (4) de la Figura 5 una estabilización de la concentración de hidrógeno a un valor de 0.65, y luego se redujo hasta estabilizarse a un valor de 0.44, lo cual concuerda con los valores reportados anteriormente, por lo que se aprecia de igual manera una relación positiva entre ambas variables.

Conclusiones

Se obtuvo una estructura de datos con valores representativos del proceso, por medio de la aplicación de métodos para el preprocesamiento de datos de las variables, incluyendo la creación de rezagos, la limpieza de datos faltantes y la detección de valores atípicos.

Se identificaron las variables con mayor sensibilidad y aporte de información al modelo, a través de la implementación de técnicas estadísticas de correlación entre variables y visualización de la matriz de correlación en forma de Heatmap.

Se seleccionó el modelo de Linear Regression para el análisis predictivo dado su mayor coeficiente de determinación luego del entrenamiento (0.7950), frente a lo reportado para ARIMAX (0.6722) y GBR (0.6395).

Se obtuvo una mejora del 14.55 % para la correlación del modelo predictivo seleccionado, mediante la agrupación de observaciones entre cada dos mediciones temporales, con

PetroRenova Indexed. Revista Científica de la Energía
Vol. 1, núm. 1, 53-68, abril-junio, 2025.
ISSN-e: 3080-6666/Maracaibo (4005), Zulia - Venezuela

coeficientes de determinación en el rango de 0.88 – 0.92 al variar el porcentaje de datos empleado para el entrenamiento.

Durante el análisis de sensibilidad, se observó que la concentración de hidrógeno en el reactor presentó valores de 0.65 y 0.44 al variar el flujo de hidrógeno crudo a 3 kg/h y 2 kg/h, respectivamente. Estos resultados evidenciaron una relación directamente proporcional entre el flujo de hidrógeno y su concentración en el reactor, tanto al incrementar como al reducir el flujo.

Referencias

Alharbi, F. y Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. *Inventions*, 7(4), 94. https://doi.org/10.3390/inventions7040094

Calofir, V., Munteanu, R., Simoiu, M. y Lemnaru, K. (2024). Innovative approach to estimate structural damage using linear regression and K-nearest neighbors machine learning algorithms. *Results in Engineering*, 22. https://doi.org/10.1016/j.rineng.2024.102250

Chicco, D., Warrens, M. y Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, 623. https://doi.org/10.7717/peerj-cs.623

Dubravova, H., Cap, J., Holubova, K. y Hribnak, L. (2024). Artificial Intelligence as an Innovative Element of Support in Policing. *Procedia Computer Science*, 237, 237-244. https://doi.org/10.1016/j.procs.2024.05.101

Forero-Corba, W. y Negre, F. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *Revista Iberoamericana de Educación a Distancia*, 27(1), 209-253. https://doi.org/10.5944/ried.27.1.37491

Gou, J., Sajid, G., Sabri, M., El-Meligy, M., El Hindi, K. y Othman, N. (2024). Optimizing biochar yield and composition prediction with ensemble machine learning models for sustainable production. *Ain Shams Engineering Journal*, 16. https://doi.org/10.1016/j.asej.2024.103209

Hyndman, R. y Athanasopoulos, G. (2021). *Forecasting: principles and practice*. (3.ª ed.). OTexts. https://otexts.com/fpp3/

Khodabakhshi, M. y Bijani, M. (2024). Predicting scale deposition in oil reservoirs using machine learning optimization algorithms. *Results in Engineering*, 22. https://doi.org/10.1016/j.rineng.2024.102263

Kovac, N., Ratkovic, K., Farahani, H. y Watson P. (2024). A practical applications guide to machine learning regression models in psychology with Python. *Methods in Psychology*, 11. https://doi.org/10.1016/j.metip.2024.100156

Mansi, M., Almobarak, M., Ekundayo, J., Lagat C. y Xie, Q. (2023). Application of supervised machine learning to predict the enhanced gas recovery by CO2 injection in shale gas reservoirs. *Petroleum*, 10, 124-134. https://doi.org/10.1016/j.petlm.2023.02.003

Ngige, G., Ovuoraye, P., Igwegbec, C., Fetahi, E., Okekec, J., Yakubud, A. y Onyechi, P. (2022). RSM optimization and yield prediction for biodiesel produced from alkali-catalytic transesterification of pawpaw seed extract: Thermodynamics, kinetics, and Multiple Linear Regression analysis. *Digital Chemical Engineering*, 6. https://doi.org/10.1016/j.dche.2022.100066

Qu, K. (2024). Research on linear regression algorithm. *MATEC Web of Conferences*, 395. https://doi.org/10.1051/matecconf/202439501046

Shaveta, N. (2023). A review on machine learning. *International Journal of Science and Research* Archive, 9(1), 281–285. https://doi.org/10.30574/ijsra.2023.9.1.0410

Singh, U., Rizwan, M., Alaraj, M. y Alsaidan, I. (2021). A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. *Energies*, 14(16), 5196. https://doi.org/10.3390/en14165196

Declaración de conflicto de interés y originalidad

Conforme a lo estipulado en el C*ódigo de ética y buenas prácticas* publicado en **PetroRenova Indexed***, **Revista Científica de la Energía***, los autores ***Sabino Montero, Karla Valentina y Noguera Hernández, José Ricard****o, declaran al Comité Editorial que n*o tienen situaciones que representen conflicto de interés real, potencial o evidente, de carácter académico, financiero, intelectual o con derechos de propiedad intelectual relacionados con el contenido del artículo: ***Predicción de la Concentración de Hidrógeno en un Reactor de Polimerización basado en Machine Learning***, en relación con su publicación. De igual manera, declaran que el trabajo es original, no ha sido publicado parcial ni totalmente en otro medio de difusión, no se utilizaron ideas, formulaciones, citas o ilustraciones diversas, extraídas de distintas fuentes, sin mencionar de forma clara y estricta su origen y sin ser referenciadas debidamente en la bibliografía correspondiente. Consienten que el Comité Editorial aplique cualquier sistema de detección de plagio para verificar su originalidad.

Para citar este artículo:

Sabino, K. y Noguera, J. (2025). Predicción de la Concentración de Hidrógeno en un Reactor de Polimerización basado en Machine Learning. *PetroRenova, Revista Científica de la Energía*. Vol. 1, núm. 1, abril-junio. https://doi.org/10.5281/zenodo.15643200