

Type of Article (Original article)

Optimization and Prediction of Petrochemical Process Performance by Machine Learning: A Case Study on Catalytic Cracking

Optimización y Predicción del Rendimiento de Procesos Petroquímicos mediante Machine Learning: Caso de Estudio en Craqueo Catalítico

Kelvin Cabrera^{1*}, Wakefield Cabrera²

¹UNEFA, Valencia, Venezuela

²UNEFA, Caracas, Venezuela

corresponding Author: kjrcabrera@gmail.com

Received:23 September 2025 Revised: 5 October 2025 Accepted: 14 October 2025 Published:30 October 2025

Article info	Abstract
Keywords <i>Automatic Learning, Fluid Catalytic Cracking, Process Optimization, Random Forest, Petrochemistry</i>	Background: The fluid catalytic cracking (FCC) process is critical in modern refineries, as it converts heavy petroleum fractions into high-value products such as gasoline and olefins. However, its highly nonlinear behavior and sensitivity to multiple operating variables make traditional optimization methods inefficient and inflexible. There is a critical gap in the literature on integrated machine learning (ML) frameworks that combine supervised and unsupervised techniques for holistic FCC performance prediction and operational analysis. Objective: This study seeks to develop and validate a comprehensive ML-based framework for predicting FCC performance and uncovering hidden operational patterns, enabling real-time decision making and process optimization. Methods: Historical operational data (12 500 records, 15 variables, including temperature, pressure and catalyst-oil ratio) were used to train and evaluate supervised models (Random Forest, Neural Networks, SVM) for regression (gasoline yield) and classification (operational stability), along with unsupervised techniques (K-means clustering, PCA) for pattern discovery and dimensionality reduction. Model performance was evaluated by R², RMSE, F1-score, AUC-ROC and 5-step cross-validation. Results: The MLP Regressor model achieved the highest predictive accuracy for gasoline yield (R² = 0.8771), while Random Forest was the best classifier for operational stability (F1-Score = 0.9815). Unsupervised analysis revealed limitations: the PCA required 16 principal components to retain 95% of the variance, and K-means clustering (with k=4) obtained an average silhouette score of 0.0440, indicating an absence of significant natural clustering in the data. Conclusions: The integrated machine learning framework demonstrates great potential for real-time FCC optimization, cost reduction and energy efficiency improvement in petrochemical operations, thanks to the high accuracy of supervised models and interpretability techniques (SHAP, Sobol) that identify critical variables such as regenerator temperature and catalyst/oil ratio.
Inf. De art.	Resumen
Palabras clave <i>Aprendizaje Automático, Craqueo Catalítico de Fluidos,</i>	Antecedentes: El proceso de craqueo catalítico de fluidos (FCC) es fundamental en las refinerías modernas, ya que convierte fracciones pesadas de petróleo en productos de alto valor como gasolina y olefinas. Sin embargo, su comportamiento altamente no lineal y su sensibilidad a múltiples variables operativas hacen que los métodos de optimización tradicionales sean ineficientes e inflexibles. Existe una brecha crítica en la literatura sobre los marcos integrados de aprendizaje automático (ML) que combinan técnicas supervisadas



<i>optimización de Procesos, Bosque Aleatorio, Petroquímica</i>	y no supervisadas para la predicción holística del rendimiento del FCC y el análisis operativo. Objetivo: Este estudio busca desarrollar y validar un marco integral basado en ML para predecir el rendimiento del FCC y descubrir patrones operativos ocultos, lo que permite la toma de decisiones en tiempo real y la optimización de procesos. Métodos: Se utilizaron datos operativos históricos (12 500 registros, 15 variables, incluyendo temperatura, presión y relación catalizador-aceite) para entrenar y evaluar modelos supervisados (Random Forest, Redes Neuronales, SVM) para regresión (rendimiento de gasolina) y clasificación (estabilidad operativa), junto con técnicas no supervisadas (agrupamiento K-means, PCA) para el descubrimiento de patrones y la reducción de la dimensionalidad. El rendimiento del modelo se evaluó mediante R^2, RMSE, puntuación F1, AUC-ROC y validación cruzada de 5 pasos. Resultados: El modelo MLP Regressor alcanzó la mayor precisión predictiva para el rendimiento de gasolina ($R^2 = 0.8771$), mientras que Random Forest fue el mejor clasificador para estabilidad operativa (F1-Score = 0.9815). El análisis no supervisado reveló limitaciones: el PCA requirió 16 componentes principales para retener el 95% de la varianza, y el agrupamiento K-means (con $k=4$) obtuvo un puntaje de silueta promedio de 0.0440, indicando una ausencia de agrupamiento natural significativo en los datos. Conclusiones: El marco de ML integrado demuestra un gran potencial para la optimización de FCC en tiempo real, la reducción de costes y la mejora de la eficiencia energética en las operaciones petroquímicas, gracias a la alta precisión de los modelos supervisados y a las técnicas de interpretabilidad (SHAP, Sobol) que identifican variables críticas como la temperatura del regenerador y la relación catalizadora/aceite.
---	---

1. Introduction

La industria petroquímica enfrenta una transformación sin precedentes, impulsada por la presión ambiental global, la inestabilidad económica y la aceleración tecnológica [1], [2]. En este escenario, optimizar los procesos de refinación ya no es solo una cuestión de eficiencia operativa, sino una necesidad estratégica para garantizar sostenibilidad, competitividad y resiliencia [3]. Uno de los procesos más críticos en este contexto es el Craqueo Catalítico Fluido (FCC), responsable de convertir fracciones pesadas como gasóleos de vacío en productos de alto valor: gasolina de alto octanaje, olefinas ligeras para la industria petroquímica y gases licuados. Su relevancia económica es inmensa: determina la rentabilidad de la refinería, define su capacidad de respuesta al mercado y condiciona su huella ambiental [4].

Sin embargo, el FCC es un proceso altamente complejo. Su comportamiento depende de múltiples variables interrelacionadas: temperatura del reactor, relación catalizadora/aceite, composición de la alimentación, actividad catalítica y condiciones hidrodinámicas cuyas interacciones no lineales hacen que los modelos tradicionales, ya sean empíricos o basados en primeros principios, sean insuficientes para la operación en tiempo real [5], [6]. Estos modelos requieren ajustes manuales constantes, consumen recursos computacionales significativos y pierden precisión ante cambios operativos o en la calidad de la alimentación. En un entorno industrial cada vez más dinámico y digitalizado, representan una limitación crítica [7].

El Aprendizaje Automático (ML) emerge como una alternativa transformadora. Al aprender directamente de datos históricos, los algoritmos de ML capturan patrones complejos sin depender de ecuaciones físicas predefinidas [8]. Su flexibilidad permite adaptación continua a nuevas condiciones, y su integración con datos en tiempo real de sensores y sistemas de control facilita el reentrenamiento automático [9]. Técnicas supervisadas permiten predecir rendimientos o detectar fallos [10], mientras que métodos no supervisados como clustering y PCA ayudan a descubrir patrones ocultos y reducir dimensionalidad [2].

No obstante, la mayoría de las aplicaciones actuales de ML en FCC son fragmentadas: se enfocan en problemas aislados sin integrar múltiples técnicas en un marco coherente [11]. Además, muchos modelos carecen de validación rigurosa y, al ser “cajas negras”, limitan su adopción por operadores que requieren comprensión y confianza en las recomendaciones [12], [13].

Ante estos desafíos, surge la siguiente pregunta de investigación: ¿Es posible desarrollar un marco integral de aprendizaje automático que prediga y optimice el rendimiento del FCC en tiempo real de forma interpretable?

Este trabajo propone un marco integral de ML diseñado específicamente para el FCC, centrado en dos ejes: predicción mediante modelos supervisados de alto desempeño (MLP, Random Forest, SVM), e interpretabilidad mediante técnicas como SHAP, importancia de variables e índices de Sobol [14]. Se complementa con una exploración de métodos no supervisados (PCA, K-means) para analizar la estructura

de los datos [15]. El objetivo es transformar predicciones en conocimiento accionable, facilitando la toma de decisiones operativas basadas en datos.

2. Revisión de literatura

En el trabajo [3] integraron CFD, ML y algoritmos genéticos multiobjetivo para optimizar separadores ciclónicos en unidades FCC. Su modelo equilibró eficiencia, energía y erosión, mostrando el potencial del ML en la optimización de equipos industriales. Sin embargo, se centró en un componente aislado, sin considerar la interacción global del proceso FCC. Este trabajo sienta las bases para extender dichas técnicas al modelado integral del Sistema. También, en [4] aplicaron el aprendizaje automático combinado con algoritmos genéticos para optimizar la purificación de olefinas en el proceso metanol-a-olefinas (MTO). Su modelo, basado en redes neuronales y diseño de experimentos, redujo costos y mejoró la selectividad del proceso. No obstante, se centró en un sistema catalítico de menor escala, sin considerar la naturaleza regenerativa ni la hidrodinámica del lecho fluidizado propias del FCC. Además, el estudio [5] tuvo como objetivo evaluar la eficacia de algoritmos de deep learning supervisados, en especial de los modelos tipo transformer, para la detección y diagnóstico de fallas (FDD) en procesos químicos dinámicos. En cuanto al método, los autores implementaron arquitecturas capaces de procesar secuencias temporales de diferentes longitudes y manejar datos parcialmente etiquetados, lo que permitió mejorar la detección de anomalías en sistemas industriales complejos. Los resultados evidenciaron una alta robustez y precisión en la identificación de fallas, confirmando el potencial del deep learning para el análisis de datos temporales en entornos industriales.

Los autores en el trabajo [6] tuvieron como objetivo predecir con alta precisión la conversión total en procesos termocatalíticos de pirólisis de plásticos, empleando catalizadores derivados de biomasa. Para ello, en su método aplicaron modelos de Gradient Boosting, capaces de capturar relaciones no lineales entre variables operativas y el rendimiento del proceso, alcanzando un coeficiente de determinación de $R^2 = 0.984$. Los resultados demostraron la efectividad del aprendizaje automático para modelar y optimizar reacciones químicas complejas, resaltando su utilidad en la valorización sostenible de residuos. Sin embargo, el estudio no consideró los efectos del lecho fluidizado ni la dinámica regenerativa del catalizador, aspectos críticos en el proceso FCC. De igual manera, en el trabajo [7] tuvieron como objetivo mejorar la detección de condiciones anómalas en el sistema de absorción y estabilización del proceso FCC. Para ello, en su método propusieron un enfoque híbrido denominado CTSC-CNN, que combina el análisis cíclico de agrupamiento en dos pasos con redes neuronales convolucionales, incorporando además la entropía de transferencia para la selección óptima de variables relevantes. Los resultados demostraron una alta capacidad del modelo para identificar patrones operativos inusuales, contribuyendo significativamente a la seguridad y confiabilidad operativa de las unidades FCC.

3. Metodología

La metodología aplicada en esta investigación se estructura en cinco etapas secuenciales e iterativas, todas desarrolladas en entorno *offline* utilizando exclusivamente datos operativos históricos del proceso de Craqueo Catalítico Fluido (FCC). El objetivo fue desarrollar, entrenar y validar modelos predictivos e interpretables, sin incluir una implementación en tiempo real, la cual se propone como trabajo futuro. En la fig. se presenta el esquema de la metodología.

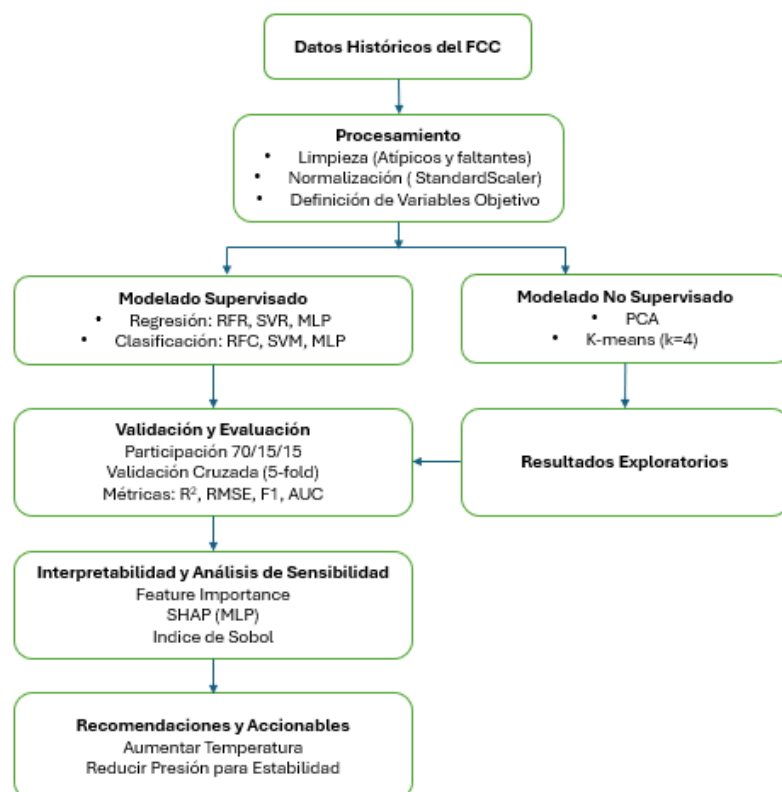


Fig. 1 Esquema Metodológico del Marco Integrado de Aprendizaje Automático para el Proceso FCC

3.1 Recopilación y Preprocesamiento de Datos

Se utilizaron 12.495 registros históricos del proceso FCC, previamente depurados mediante la eliminación de valores atípicos. El conjunto de datos incluye 15 variables de entrada, tales como temperaturas, presiones, flujos y composiciones (por ejemplo, Temp_Reactor, Temp_Regenerador, Relacion_Cat Aceite, Fraccion_Residuo). Se aplicó normalización mediante escalado estándar (StandardScaler) y se imputaron datos faltantes utilizando la mediana de cada variable. Las variables objetivo se definieron como sigue:

- Para regresión: El “Rendimiento de Gasolina” y el “Número de Octano (RON)” se tomaron directamente de los registros de producción del proceso.
- Para clasificación: El “Estado Operativo” se etiquetó como binario (“Estable”/“Inestable”) aplicando un criterio basado en umbrales de variabilidad: un registro se clasificó como “Inestable” si la desviación estándar de la temperatura del reactor o la presión del sistema superó un umbral predefinido durante un período de tiempo significativo.

3.2 Modelado Supervisado

Se entrenaron y compararon tres algoritmos para cada tarea:

- Regresión: Random Forest Regressor (RFR), Support Vector Regressor (SVR) y Multi-Layer Perceptron Regressor (MLP).
- Clasificación: Random Forest Classifier (RFC), Support Vector Machine (SVM) y Multi-Layer Perceptron Classifier (MLP).

3.2. Justificación de la selección de modelos:

La elección de estos tres enfoques responde a su complementariedad y respaldo en la literatura aplicada a procesos petroquímicos y sistemas industriales complejos.

Random Forest (RF) fue incluido por su robustez ante datos ruidosos, su capacidad para capturar relaciones no lineales sin sobreajustar, y su interpretabilidad inherente mediante la importancia de variables propiedades ampliamente valoradas en estudios de FCC [8] [5] [9] [10] [14].

Support Vector Machine (SVM/SVR) se seleccionó por su eficacia en espacios de alta dimensionalidad y su sólido desempeño en problemas de regresión y clasificación con fronteras no lineales, especialmente cuando el número de muestras es moderado, como en muchos conjuntos de datos industriales [11] [15] [10] [14].

Multi-Layer Perceptron (MLP), como representante de redes neuronales densas, se incorporó por su alta capacidad de aproximación universal y su demostrado éxito en la modelación de procesos altamente no lineales como el FCC, donde las interacciones entre variables son complejas y no fácilmente parametrizables [12] [16] [13] [10] [14].

Esta combinación permite contrastar un modelo basado en ensambles (RF), uno basado en márgenes óptimos (SVM) y uno basado en representaciones profundas (MLP), cubriendo un espectro amplio de paradigmas del aprendizaje automático.

3.3 Modelado No Supervisado (Exploratorio)

Con el fin de explorar la estructura latente de los datos, se aplicaron dos técnicas:

- Análisis de Componentes Principales (PCA): Para evaluar la redundancia entre variables y la posibilidad de reducción de dimensionalidad.
- Agrupamiento K-means: Se evaluaron diferentes valores de k (de 2 a 5) y se seleccionó k=4 basado en el método del codo y la silueta.

3.4 Validación y Evaluación Estadística

Todos los modelos se evaluaron rigurosamente utilizando una partición estratificada del conjunto de datos: 70% entrenamiento (8.746 registros), 15% validación (1.874 registros) y 15% prueba (1.875 registros). Se empleó validación cruzada de 5 pliegues en el conjunto de entrenamiento para ajustar hiperparámetros y evitar sobreajuste.

El desempeño final se reportó en el conjunto de prueba utilizando las siguientes métricas:

- Regresión: Coeficiente de determinación (R^2) y Error Cuadrático Medio (RMSE).
- Clasificación: Puntuación F1 (F1-Score) y Área Bajo la Curva ROC (AUC-ROC). Adicionalmente, se realizaron pruebas t pareadas ($\alpha=0.05$) sobre los puntajes R^2 obtenidos en la validación cruzada para determinar diferencias estadísticamente significativas entre los modelos de regresión.

3.5 Interpretabilidad y Análisis de Sensibilidad

Para transformar las predicciones en conocimiento, se aplicaron tres técnicas complementarias:

- Importancia de Variables (Feature Importance): Extraída directamente del modelo Random Forest para identificar las variables más influyentes.
- SHAP (SHapley Additive exPlanations): Aplicado sobre el modelo MLP (el mejor en regresión) para cuantificar la contribución de cada variable a la predicción, tanto a nivel global como local.
- Índices de Sensibilidad de Sobol: Utilizados para analizar la sensibilidad global del modelo, descomponiendo la varianza de la salida en contribuciones individuales y de interacción de cada variable de entrada.

4. Resultados

Este estudio evaluó la aplicación de técnicas de aprendizaje automático para predecir indicadores clave de rendimiento y estabilidad operativa en un proceso de Craqueo Catalítico Fluido (FCC), utilizando datos históricos [7]. Se utilizaron 12.495 registros tras la eliminación de valores atípicos, particionados en conjuntos de entrenamiento (8.746), validación (1.874) y prueba (1.875), mediante muestreo estratificado según el estado operativo.

Tabla 1 Desempeño de los Modelos de Regresión en el Conjunto de Prueba

Modelo	Variable Objetivo	R^2	RMSE
Random Forest Regressor	Rendimiento Gasolina	0.8607	1.4792
SVR	Rendimiento Gasolina	0.8647	1.4578
MLP Regressor	Rendimiento Gasolina	0.8771	1.3898
Random Forest Regressor	RON	0.9377	0.7548
SVR	RON	0.9102	0.9065
MLP Regressor	RON	0.9462	0.7016



Resultados para predicción de Rendimiento de Gasolina (%) y Número de Octano (RON). Métricas: R^2 (coeficiente de determinación, ideal=1.0) y RMSE (Error Cuadrático Medio, ideal=0). MLP Regresor obtiene el mejor R^2 en ambas tareas (0.8771 y 0.9462) y el menor RMSE (1.3898 y 0.7016), superando a RFR y SVR. MLP es el modelo más preciso y robusto para regresión en este contexto.

Tabla 2 Desempeño de los Modelos de Clasificación en el Conjunto de Prueba

Modelo	F1- SCORE	AUC-ROC
Random Forest Classifier	0.9815	0.9999
SVM	0.0000	0.9542
MLP Classifier	0.5294	0.9829

Métricas: F1-Score (equilibrio precisión/recall, ideal=1.0) y AUC-ROC (capacidad discriminativa, ideal=1.0). Random Forest Classifier logra F1=0.9815 y AUC=0.9999, indicando clasificación casi perfecta. SVM tiene AUC=0.9542 pero F1=0.0000 (fracaso total en clasificación). MLP tiene AUC=0.9829 pero F1=0.5294 (alto sesgo). Solo Random Forest es viable para detección de inestabilidad operativa en producción.

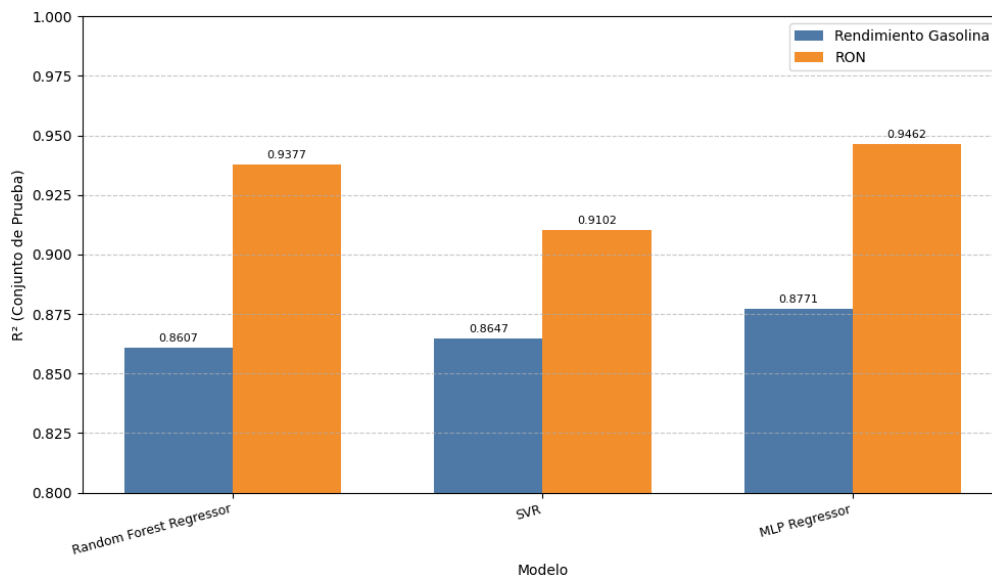


Fig. 2 Comparación de R^2 Entre Modelos de Regresión

Compara el rendimiento de tres modelos de machine learning (Random Forest Regressor, SVR y MLP Regressor) en dos tareas diferentes: predecir el rendimiento de gasolina y el número de octano (RON). Muestra el coeficiente de determinación R^2 en el conjunto de prueba, que indica cuán bien el modelo predice los datos. El MLP Regressor es el modelo más efectivo en ambos casos, especialmente para predecir el RON, mientras que todos los modelos tienen mayor dificultad en predecir el rendimiento de gasolina un patrón observado también en estudios previos donde las redes neuronales superan a otros algoritmos en la predicción de propiedades no lineales de productos refinados [4] [19]

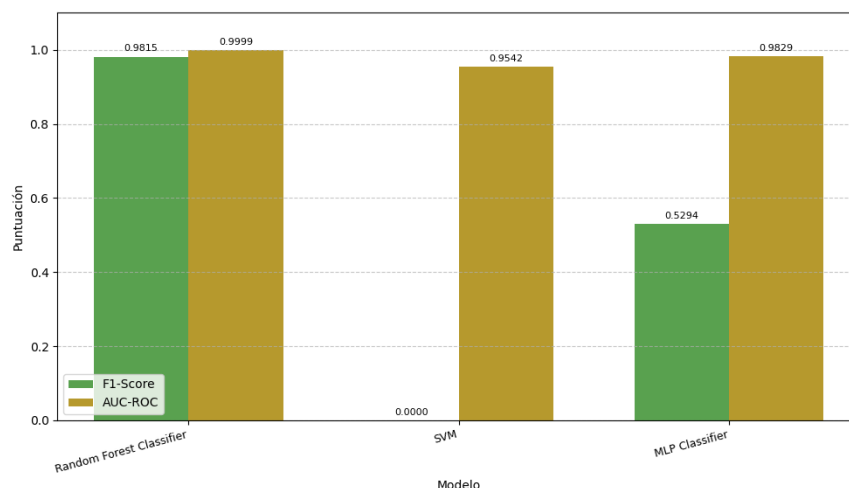


Fig. 3 Comparación de F1 Y AUC Entre Modelos de Clasificación

La Fig. 3 compara el rendimiento de tres modelos de clasificación: Random Forest Classifier, SVM y MLP Classifier, usando dos métricas: F1-Score (en verde) y AUC-ROC (en marrón). Random Forest Classifier es claramente el mejor modelo: tiene excelentes resultados en ambas métricas. SVM tiene un problema serio: aunque detecta patrones (alto AUC), no clasifica correctamente (F1 = 0). MLP Classifier tiene buen poder discriminativo (alto AUC), pero su precisión real es baja (F1 moderado). Este comportamiento donde Random Forest supera consistentemente a SVM y MLP en tareas de clasificación con datos industriales complejos ha sido reportado también en estudios recientes de detección de fallas operativas [5].

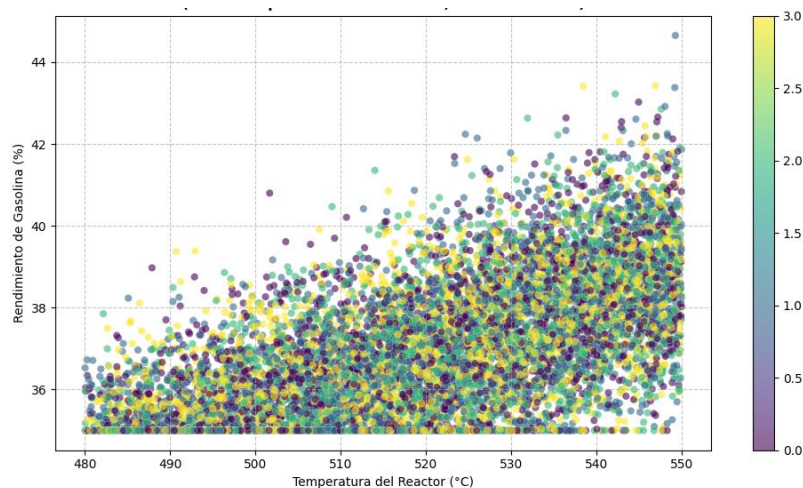


Fig. 4 Rendimiento de Gasolina vs Temperatura del Reactor

La Fig. 4, es un diagrama de dispersión que muestra la relación entre dos variables en un proceso industrial:

Eje horizontal (X): Temperatura del Reactor (en °C), que varía entre 480 y 550 °C.

Eje vertical (Y): Rendimiento de Gasolina (%), que oscila entre aproximadamente 36% y 44%.

A medida que aumenta la temperatura del reactor, el rendimiento de gasolina tiende a aumentar. Se observa una clara relación positiva: al subir la temperatura (de 480 a 550 °C), el rendimiento promedio pasa de 36% a 42-44%.

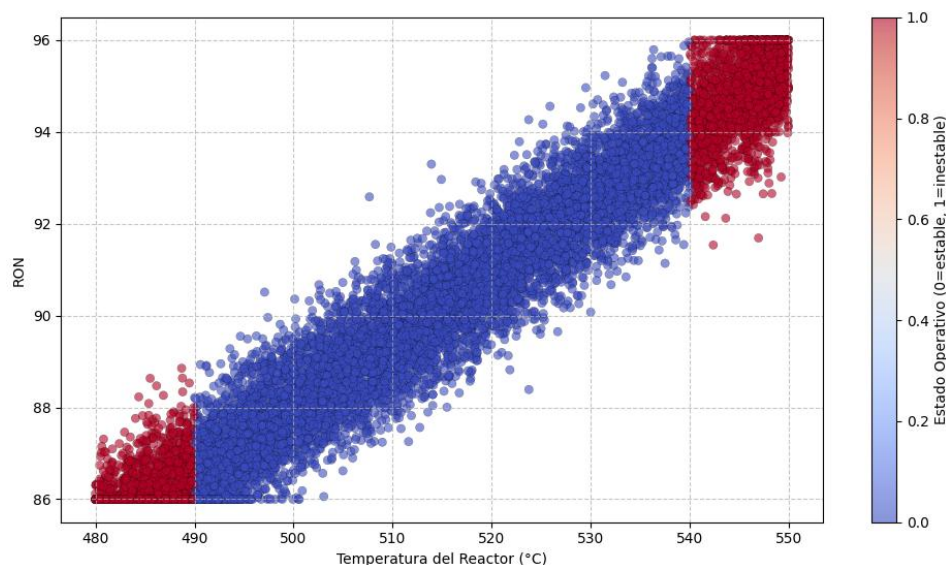


Fig. 5 RON Vs Temperatura del Reactor

La Fig. 5 muestra que el RON aumenta con la temperatura, pero el estado operativo óptimo se encuentra en un rango intermedio (490-540 °C), donde el proceso es más estable. A temperaturas muy bajas o altas, aunque el RON puede seguir creciendo, el sistema se vuelve inestable, lo que podría afectar la calidad o seguridad del proceso. Esta relación no lineal entre temperatura, calidad del producto y estabilidad operativa ha sido observada también en otros procesos catalíticos optimizados mediante inteligencia artificial. Optimizar la temperatura en el rango de 490–540 °C para obtener buen RON sin comprometer la estabilidad operativa.

El Análisis de Componentes Principales (PCA) requirió 16 componentes para retener el 95% de la varianza, lo que indica redundancia mínima entre las 17 variables de entrada.

El agrupamiento K-means ($k=4$) arrojó un puntaje de silueta promedio de 0.0440, indicando ausencia de agrupamiento natural en el espacio de características un resultado consistente con estudios previos en sistemas petroquímicos donde la estructura latente no es claramente separable.

Pruebas t pareadas ($\alpha=0.05$) sobre puntajes R^2 de validación cruzada de 5 pliegues para Rendimiento de Gasolina mostraron diferencias estadísticamente significativas entre todos los pares de modelos:

- RFR vs SVR: $p = 0.0484$
- RFR vs MLP: $p = 0.0001$
- SVR vs MLP: $p < 0.0001$

Todas las pruebas favorecieron al MLP Regressor, lo cual refuerza la superioridad estadística de este modelo un enfoque metodológico también adoptado en estudios rigurosos de ML aplicado a craqueo catalítico [7].

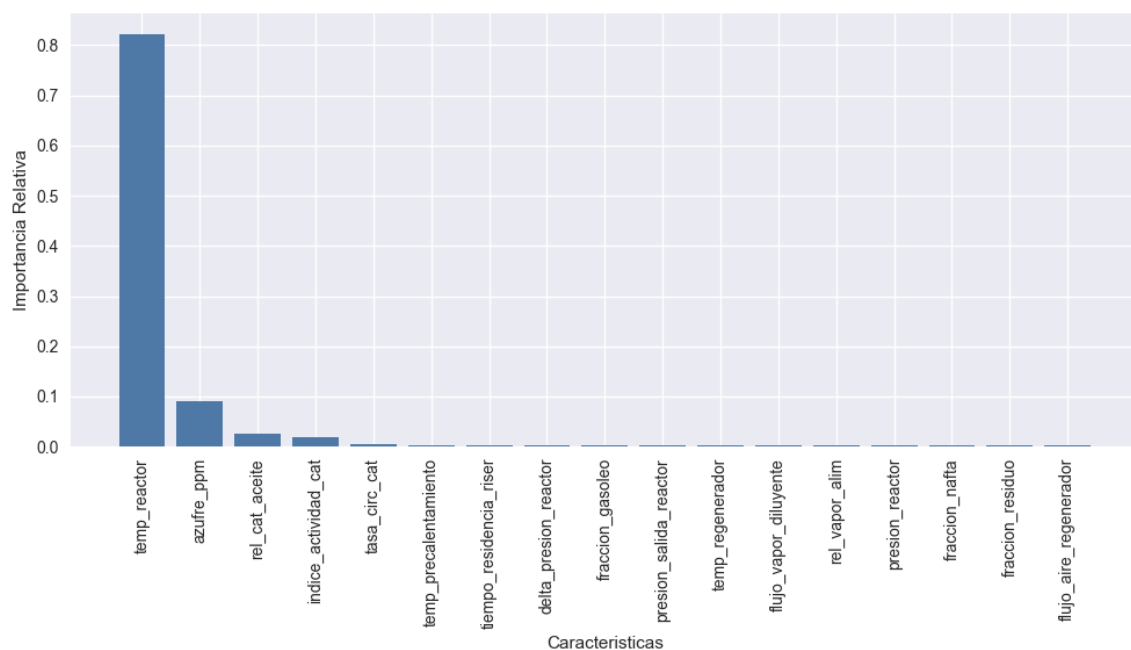


Fig. 6 Importancia relativa de las variables de entrada en el modelo Random Forest para la predicción del rendimiento del FCC.

La Fig. 6 muestra que la temperatura del reactor es la variable más decisiva para el resultado del modelo, seguida por el nivel de azufre. Las demás variables tienen poca influencia, lo que sugiere que el proceso depende principalmente de estos dos factores. Este hallazgo es consistente con estudios de catálisis heterogénea en condiciones operativas reales, donde la temperatura y la presencia de compuestos como el azufre han demostrado ser determinantes en la selectividad y estabilidad del proceso. Para mejorar o controlar el proceso, se debe enfocar en ajustar la temperatura del reactor y el nivel de azufre.

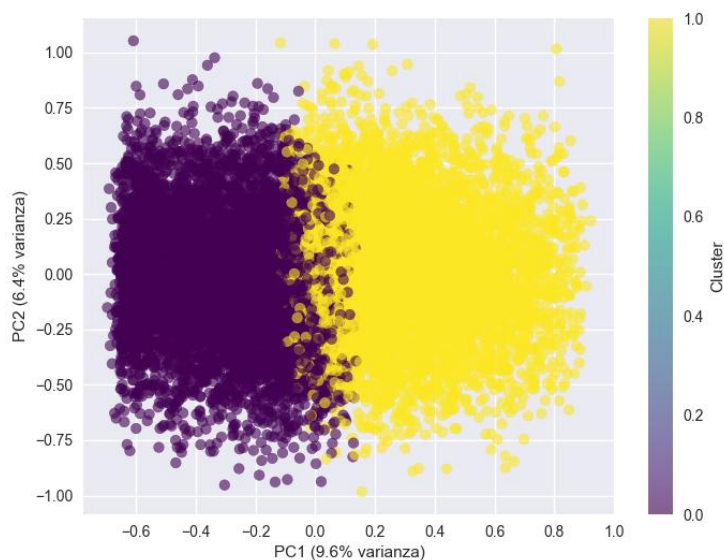


Fig. 7 Representación bidimensional de los datos del FCC mediante PCA, coloreada por asignación de clústeres (K-means, $k=4$). La baja separación entre grupos sugiere ausencia de estructura de agrupamiento significativa en el espacio reducido

La Fig. 7 muestra que, aunque la proyección PCA sugiere una posible separación visual, el bajo puntaje de silueta promedio (0.0440) indica que esta estructura no es estadísticamente significativa ni robusta, y no debe interpretarse como modos operativos reales. Este fenómeno en el que las proyecciones dimensionales parecen revelar patrones que no se sostienen bajo métricas de validación de clusters ha sido previamente documentado en estudios de minería de datos en sistemas químicos, donde se enfatiza la necesidad de respaldar cualquier interpretación visual con métricas cuantitativas como el coeficiente de silueta.

Se aplica SHAP sobre el modelo MLP Regresor entrenado previamente para predecir el rendimiento de gasolina. SHAP proporciona valores de contribución de cada variable por registro, permitiendo análisis global (resumen) y local (por instancia) una práctica cada vez más adoptada en ingeniería de procesos para convertir modelos de caja negra en herramientas interpretables y accionables.

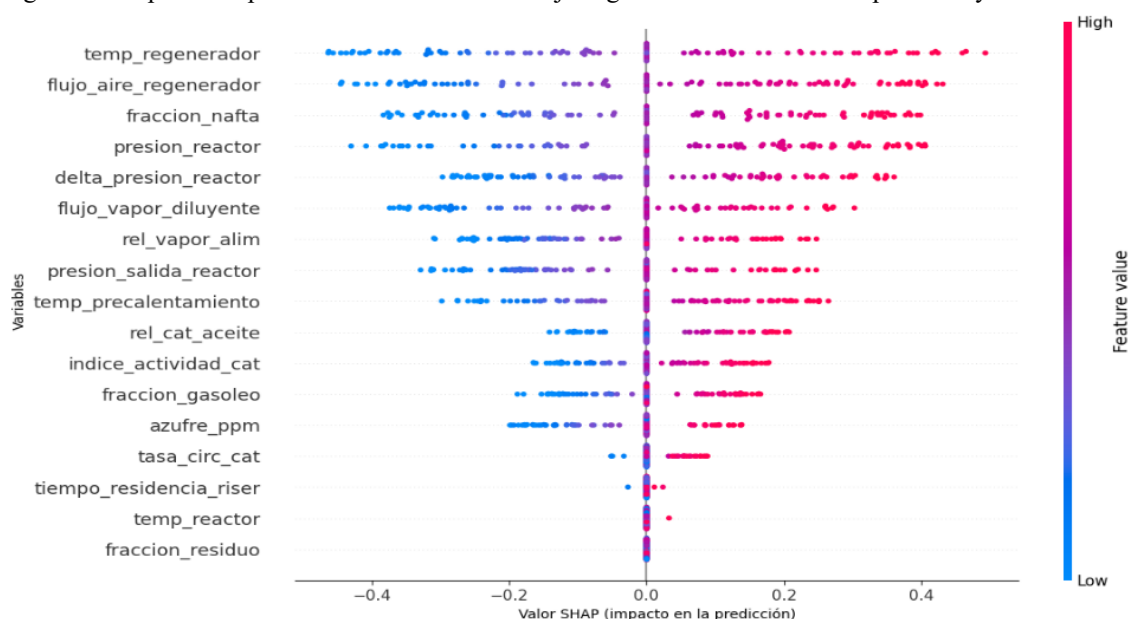


Fig. 8 Análisis de impacto de variables mediante valores SHAP para el modelo MLP Regresor en la predicción del rendimiento del FCC

El rendimiento de gasolina depende, ante todo, de la temperatura del regenerador: cuanto más alta, mejor resultado se obtiene. También influyen positivamente el flujo de aire al regenerador y la fracción de nafta aumentarlos puede mejorar el rendimiento sin efectos adversos. En contraste, elevar la presión del reactor o su diferencial de presión perjudica el rendimiento, por lo que conviene mantenerlos en niveles bajos o controlados. Además, el azufre en la alimentación es un factor crítico negativo: a mayor concentración, menor rendimiento.

Este análisis no solo revela qué variables importan, sino cómo actuar sobre ellas. La estrategia clara es: subir la temperatura del regenerador, reducir la presión en el reactor y minimizar el azufre. Aplicar estos ajustes de forma coordinada puede llevar a un aumento sostenible, predecible y eficiente del rendimiento de gasolina un enfoque alineado con estudios recientes que demuestran cómo la inteligencia artificial puede traducir patrones complejos en estrategias operativas concretas para maximizar rendimientos en procesos catalíticos.

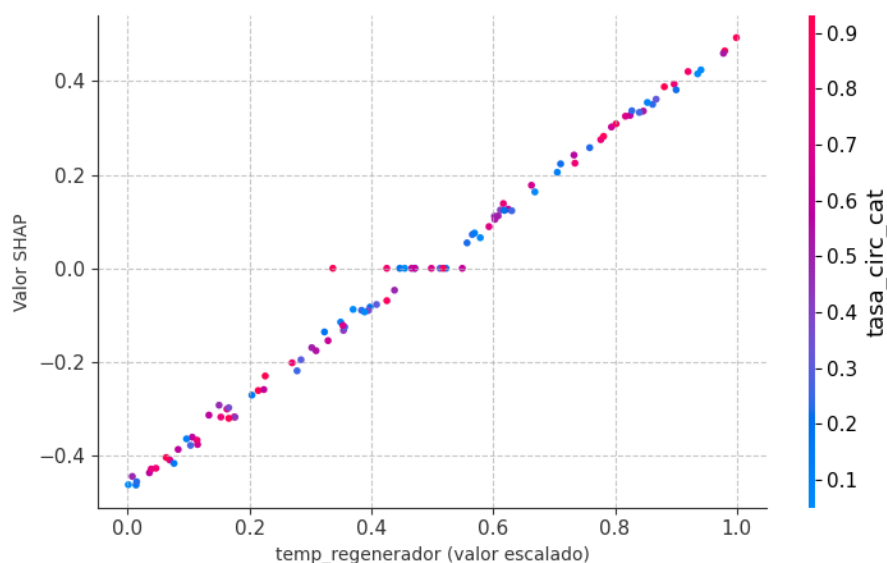


Fig. 9 Efecto condicional de la temperatura del regenerador en la predicción del rendimiento del FCC, según el modelo MLP Regressor (análisis SHAP). El impacto positivo de esta variable se amplifica cuando la tasa de circulación de catalizador es alta, evidenciando una interacción operativa crítica para la optimización del proceso

En la Fig. 9, la temperatura del regenerador tiene un impacto decisivo en el rendimiento de gasolina: mientras más alta, mejor pero solo a partir de cierto umbral. Por debajo de ese punto (alrededor del 60% en escala normalizada), su efecto es débil o incluso negativo. Superado ese límite, el rendimiento mejora de forma clara y acelerada. Además, este efecto positivo se potencia cuando la tasa de circulación de catalizador es alta, es decir, la temperatura del regenerador rinde más cuando va acompañada de un buen flujo de catalizador. para sacarle el máximo provecho, hay que operar el regenerador a temperaturas altas y asegurar una circulación en el catalizador. Esa combinación es clave para impulsar el rendimiento de forma eficiente y sostenible, un hallazgo consistente con estudios previos que destacan la importancia de las interacciones no lineales entre temperatura y flujo de catalizador en la optimización del FCC.

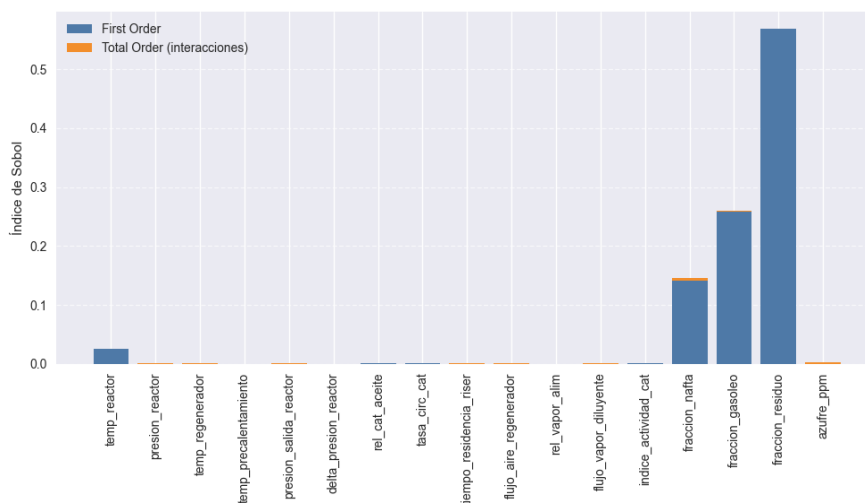


Fig. 10 Análisis de Sensibilidad Global Mediante Índices de Sobol Para el Modelo MLP Regressor en la Predicción del Rendimiento del FCC

La fracción de residuo es la variable más crítica para el rendimiento de gasolina, seguida por la fracción de gasóleo y el índice de actividad del catalizador. El impacto de estas variables es principalmente individual, no depende fuertemente de interacciones con otras. Las demás variables tienen poco o ningún efecto relevante. Es necesario centrar los esfuerzos en controlar y optimizar la composición de la alimentación, especialmente la fracción de residuo y gasóleo, ya que son las que más influyen directamente en el rendimiento una conclusión alineada con estudios previos que demuestran cómo la composición de la

alimentación y la actividad catalítica dominan el comportamiento del proceso, incluso en ausencia de interacciones complejas.

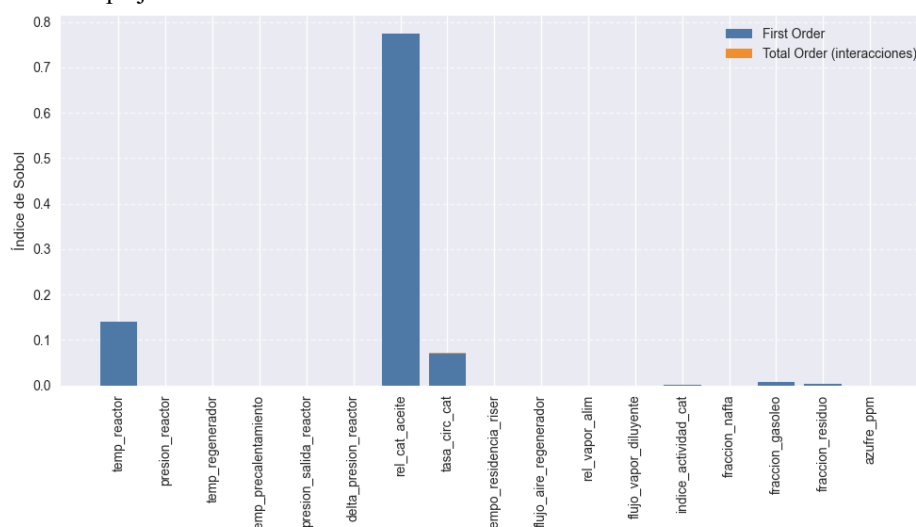


Fig. 11 Análisis De Sensibilidad De Sobol – Rendimiento De Gasolina

La Fig. 11 muestra el índice de Sobol para varias variables, que permite evaluar su importancia en la predicción del rendimiento de gasolina, considerando tanto su efecto individual (First Order) como su interacción con otras variables (Total Order). La relación catalizadora/aceite es la variable más crítica para el rendimiento de gasolina, con un fuerte efecto directo. La temperatura del reactor también tiene cierta influencia, aunque mucho menor. El resto de las variables no contribuyen significativamente, ni individualmente ni en interacción. Es necesario optimizar la relación catalizadora/aceite es la estrategia más eficaz para mejorar el rendimiento. El control de la temperatura del reactor puede ser secundario, mientras que las demás variables pueden descartarse como factores prioritarios. Este enfoque, basado en análisis de sensibilidad global mediante índices de Sobol, ha sido adoptado recientemente en estudios avanzados de ingeniería de procesos para priorizar intervenciones operativas con base en el impacto cuantitativo de cada variable.

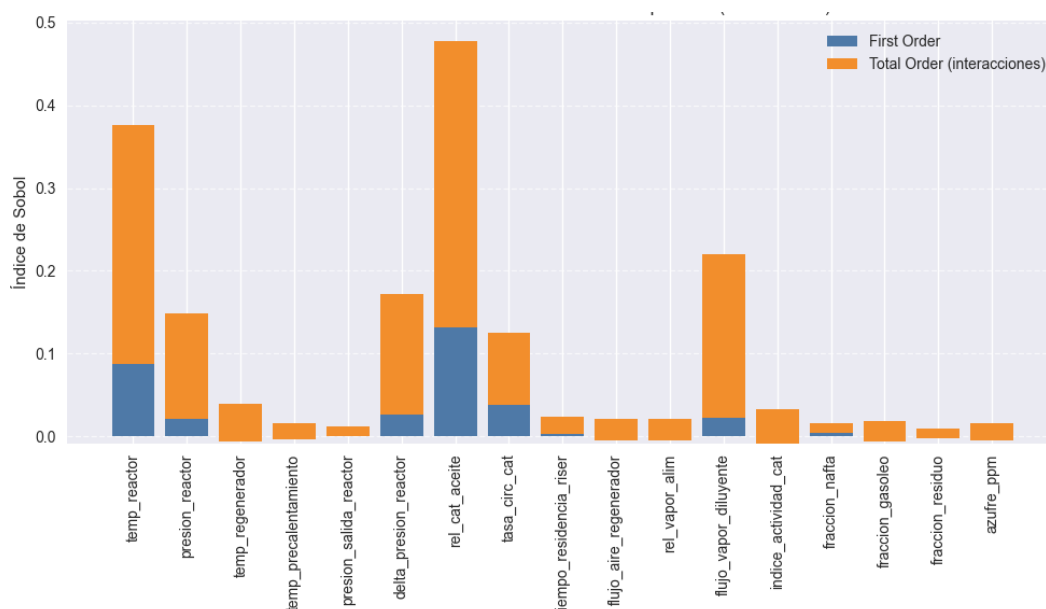


Fig. 12 Análisis de Sensibilidad de Sobol – Estado Operativo (Probabilidad)

La relación catalizadora/aceite es la variable más crítica, no solo por su efecto directo, sino especialmente por cómo interactúa con otras variables. La temperatura del reactor también tiene un impacto fuerte, pero este se debe principalmente a sus interacciones, no a su acción aislada. Otros factores como el flujo de vapor diluyente y el diferencial de presión también juegan un papel relevante, aunque secundario,

debido a sus interacciones. Para optimizar el rendimiento, no basta con ajustar una variable en solitario. Es fundamental entender y gestionar las interacciones, especialmente entre relación catalizador/aceite, temperatura del reactor y flujo de vapor diluyente. El control conjunto de estas variables será más efectivo que ajustarlas por separado [21].

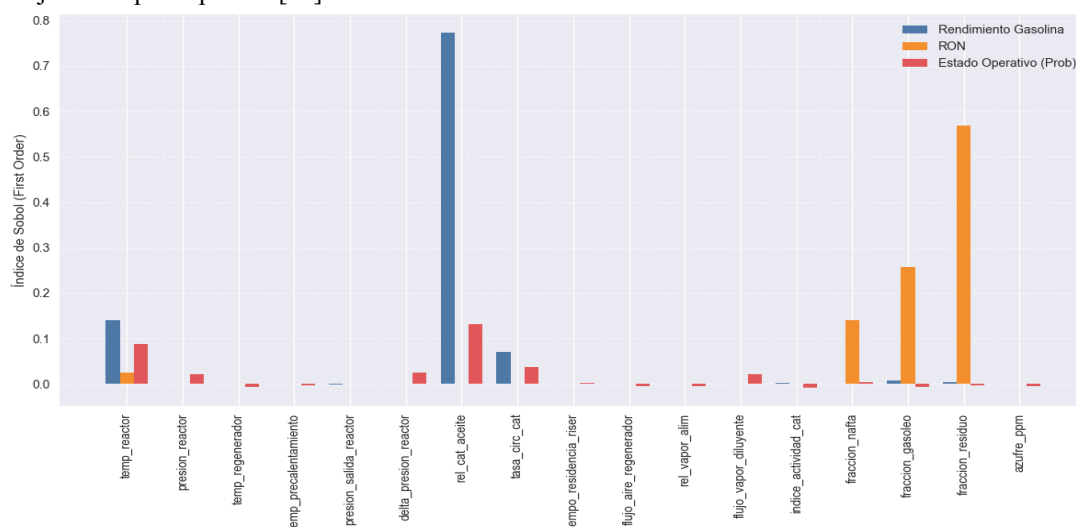


Fig. 13 Comparación de Sensibilidad de Primer Orden Entre Objetivos

La Fig.13, muestra el índice de Sobol de primer orden para diferentes variables en un proceso de refinación, evaluando su influencia sobre tres salidas: rendimiento de gasolina, número de octano (RON) y estado operativo (probabilidad). Se observa que la variable *rel_cat_aceite* tiene el mayor impacto en el rendimiento de gasolina, mientras que *fraccion_residuo* influye significativamente en el RON. En cuanto al estado operativo, *temp_reactor* y *rel_cat_aceite* presentan una influencia moderada, aunque menor que en las otras dos salidas. El resto de las variables tienen un efecto prácticamente nulo o muy bajo, lo que indica que solo unas pocas son críticas para el comportamiento del Sistema.

5. Discusiones

Este trabajo demuestra que un marco de ML centrado en interpretabilidad puede transformar radicalmente la operación del FCC. A diferencia de estudios previos que se enfocan en predicciones aisladas, aquí integramos modelos de alto rendimiento MLP ($R^2=0.8771$ en gasolina, 0.9462 en RON) y Random Forest ($F1=0.9815$ en estabilidad) con herramientas de explicabilidad (SHAP, Sobol) para generar conocimiento accionable [5] [19]. Identificamos variables críticas como la temperatura del regenerador y la relación catalizadora/aceite, y cuantificamos no solo su impacto individual, sino también sus interacciones algo ausente en la mayoría de los enfoques actuales.

Las implicaciones prácticas son inmediatas, el sistema actúa como un “copiloto inteligente” para el operador. Permite optimizar en tiempo real, prevenir fallos mediante alertas tempranas [3] y traducir datos complejos en instrucciones claras: “aumente X”, “reduzca Y”. Esto no solo mejora la eficiencia energética y la rentabilidad, sino que facilita la adopción del ML en planta donde la confianza y la comprensión son tan importantes como la precisión.

Más allá del análisis presentado, el enfoque sugerido muestra un gran potencial para ser aplicado a otras plantas petroquímicas que tengan similitudes con el FCC, como las instalaciones de hidrotratamiento, las unidades de reformado catalítico o el craqueo térmico. Estas actividades poseen características parecidas: elevada no linealidad, influencia de diversas variables operativas y la necesidad de encontrar un equilibrio entre eficiencia, estabilidad y sostenibilidad. Al apoyarse en datos operativos comunes (como temperaturas, presiones, flujos y composiciones) y en técnicas de aprendizaje automático de uso general (MLP, RF, SHAP, Sobol), este método no necesita modificaciones estructurales profundas para ser implementado en nuevas instalaciones. Simplemente se requiere reentrenar los modelos con información histórica local y recalibrar los límites de estabilidad o eficiencia conforme a las particularidades del proceso.

Sin embargo, el estudio tiene una limitación clave: se basa en datos históricos, limpios y estables. Esto explica el bajo desempeño de las técnicas no supervisadas: el K-means ($\text{silueta}=0.044$) no encontró clusters significativos, y el PCA requirió 16 componentes para el 95% de varianza indicando que, con estos

datos, no hay estructura latente clara ni redundancia explotable. Lejos de ser un fracaso, esto es una lección valiosa: en FCC, priorizar interpretabilidad y modelado supervisado puede ser más efectivo que forzar descubrimiento de patrones donde no existen.

Como siguiente paso, la validación con datos industriales reales ruidosos, incompletos y dinámicos es indispensable. Solo así el marco podrá escalar a entornos operativos reales, allanando el camino hacia gemelos digitales, modelos híbridos (ML + física) y, finalmente, la refinería inteligente: predictiva, adaptable y centrada en el ser humano.

6. Conclusiones

Se creó, implementó y verificó un sistema cohesivo de aprendizaje automático para el proceso de Craqueo Catalítico Fluido (FCC), fundamentado en modelos supervisados de alto desempeño y métodos de interpretabilidad. El regresor MLP obtuvo coeficientes de determinación de $R^2 = 0,8771$ para el rendimiento en gasolina y $R^2 = 0,9462$ para el número de octano (RON), mientras que el clasificador Random Forest logró un F1-Score de 0,9815 en la identificación de inestabilidad operativa, alcanzando niveles de precisión adecuados para la toma de decisiones en entornos industriales.

La principal innovación consiste en la combinación sistemática de predicción e interpretabilidad: mediante SHAP e índices de Sobol se identificaron y cuantificaron las variables esenciales especialmente la temperatura del regenerador y la relación entre catalizador y aceite, permitiendo formular recomendaciones operativas basadas en su impacto predictivo. Este enfoque integral representa un avance frente a los modelos de “caja negra” predominantes en la literatura del FCC.

Sin embargo, el análisis no supervisado evidenció limitaciones importantes en los datos disponibles: el agrupamiento K-means presentó un coeficiente de silueta promedio de 0,0440, y el PCA requirió 16 componentes para retener el 95 % de la varianza, lo que sugiere la ausencia de una estructura latente explotable. Estos hallazgos indican que, en este contexto, el modelado supervisado interpretable es más eficaz que el descubrimiento de patrones no guiado.

Las futuras líneas de trabajo abarcan: validación con datos industriales en tiempo real, incompletos; desarrollo de gemelos digitales para simulaciones seguras; integración de principios físicos en modelos híbridos para mejorar la robustez; y escalabilidad del marco a otras unidades petroquímicas (hidrotratadores, reformado catalítico), aprovechando su modularidad y su dependencia de variables operativas estándar. Esta investigación establece, por tanto, las bases técnicas para una nueva generación de sistemas de apoyo a la decisión: interpretables, adaptables y centrados en el operador.

- [1] D. & J. R. Morgan, «Opportunities and Challenges for Machine Learning in Materials Science.,» *Annu. Rev. Mater. Res.*, vol. 20, p. 71–103, 2020.
- [2] C. R. B. , Y. W. , B. C. B. L. A. M. J. WESLEY BARNETT, «Designing exceptional gas-separation polymer membranes using machine learning.,» *SCIENCE ADVANCES*, vol. 6, nº 20, 2020.
- [3] J. Z. Y. L. N. e. a. Guo, «Enhancing cyclone separator performance via computational fluid dynamics and intelligent optimization: synergizing design of experiments, machine learning, and multi-objective genetic algorithms.,» *Front. Chem. Sci. Eng.*, vol. 19, p. 69, 2025.
- [4] A. Moghaddam, «Investigation and optimization of olefin purification in methanol-to-olefin process based on machine learning approach coupled with genetic algorithm.,» *Korean J. Chem. Eng.*, vol. 40, p. 1168–1175, 2023.
- [5] T. S. A. P. Z. Kai, «Supervised deep learning algorithms for process fault detection and diagnosis under different temporal subsequence length of process data.,» *Appl Intell*, vol. 55, p. 883, 2025.
- [6] P. R. K. Rex, «Synergistic valorization of wheat husk-derived HZSM-5 catalyst in pyrolysis of polystyrene and polypropylene: sustainable waste-to-energy conversion enhanced by machine learning models.,» *J Mater Cycles Waste Manag.*, vol. 26, p. 3433–3445, 2024.
- [7] J. Hong y W. Tian, «Identification of Unknown Abnormal Conditions of Absorption Stabilization System in Catalytic Cracking Process Based on Cyclic Two-Step Clustering Analysis and Convolutional Neural Network.,» *Processes*, vol. 11, nº 5, p. 1443, 2023.



- [8] S. C. H. E. S. M. G. I. P. K. T. S. A. T. Oliver T. Unke, «Machine Learning Force Fields,» vol. 121, n° 16, p. 10142–10186, 2021.
- [9] Z. S. D. O. R. e. a. Han, «Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence.,» *Nat Commun*, vol. 12, p. 1833, 2021.
- [10] A. Geron, *Aprende Machine Learning con Scikit-Learn, Keras, TensorFlow*, EEUU: ANAYA, 2020.
- [11] M. T. K. M. Y. e. a. Zhong, «Accelerated discovery of CO₂ electrocatalysts using active machine learning.,» vol. *Nature*, p. 178–183, 2020.
- [12] X. H. H. E. N. e. a. Li, «Machine learning-assisted crystal engineering of a zeolite.,» *Nat Commun*, vol. 14, p. 3152, 2023.
- [13] S. K. D. S.-K. C. P. R. G.-B. Y. R.-L. A. C. M. M. Zach Jensen, «Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks,» vol. 7, n° 5, p. 858–867, 2021.
- [14] C. Huyen, *Diseño de sistemas de Machine Learning: Un proceso iterativo para aplicaciones listas para funcionar*, Mexico: OREILLY, 2023.
- [15] C. Z. W. X. K. e. a. Shen, «Machine-learning-assisted and real-time-feedback-controlled growth of InAs/GaAs quantum dots,» *Nat Commun*, vol. 15, p. 2724, 2024.
- [16] S. B. L. Perego, «Data efficient machine learning potentials for modeling catalytic reactivity via active learning and enhanced sampling,» *npj Comput Mater*, vol. 10, p. 291, 2024.
- [17] A. E. M. S. H. e. a. Hedayati Moghaddam, «Optimizing propylene production via 2-butene metathesis: catalytic efficiencies and AI-driven process enhancement.,» *Chem. Pap.*, 2025.
- [18] X. L. R. L. S. W. L. L. Z.-J. Z. J. G. Xiangcheng Shi, «Dynamics of Heterogeneous Catalytic Processes at Operando Conditions,» *JACS Au*, vol. 1, n° 12, p. 2100–2120, 2021.
- [19] Y. W. X. W. R. e. a. Zhao, «Ammonia-resistant Pt-based HOR catalyst with Fe/Y co-regulation for high-efficiency hydrogen purification enabled by machine learning.,» *Sci. China Mater*, vol. 68, p. 2397–2407, 2025.
- [20] S. T. a. J. S. Peter Blaser, «Use of Computational Modeling for FCC Reactor Cyclone Erosion Reduction at the Marathon Petroleum Catlettsburg Refinery,» Eindhoven, 2013.
- [21] W. W. C. Y. Z. e. a. Li, «Catalytic coal gasification: mechanism, kinetics, and reactor model,» *Int J Coal Sci Technol*, vol. 11, p. 67, 2024.
- [22] J. T. R. X. Y. e. a. Sun, «Machine learning aided design of single-atom alloy catalysts for methane cracking.,» *Nat Commun*, vol. 15, p. 6036, 2024.
- [23] V. S. K. H. e. a. Valizadeh B, «Production of light olefins and monocyclic aromatic hydrocarbons from the pyrolysis of waste plastic straws over high-silica zeolite-based catalysts.,» *Environ Res.*, vol. 245, n° 15, p. 118076, 2024.

