CNN PARA PREDICCION DE INDICE DE CETANO USANDO NIR

MODELO DE RED NEURONAL ARTIFICIAL CONVOLUCIONAL PARA LA PREDICCION DEL INDICE DE CETANO DE MUESTRAS DIÉSEL USANDO ESPECTROS NIR

Javier Danilo Aranda García

Trabajo de Grado para Optar el Título de Ingeniero Químico

Director

Helver Crispiniano Álvarez Castro

Doctor en Ingeniería Química

Codirector

Giovanni Morales Medina

Doctor en Ingeniería Química

Universidad Industrial de Santander

Facultad de Ingenierías Fisicoquímicas

Escuela de Ingeniería Química

Bucaramanga

2026

Dedicatoria

Este trabajo solo está dedicado a mis padres quienes me apoyaron y tuvieron paciencia en el proceso.

Agradecimientos

Agradezco a mis padres por apoyarme y darme todo lo que pudieron para que lograra este hito.

Agradezco a la universidad por permitirme definir el enfoque de mi carrera.

Agradezco a los profesores quienes me ayudaron a sintetizar este documento y tuvieron la suficiente paciencia.

Tabla de contenido

Tabla de tablas

Lista de figuras

Lista de apéndices

Resumen

Título: Modelo de red neuronal artificial convolucional para la predicción del índice de cetano de muestras diésel usando espectros NIR.

Autor: Javier Danilo Aranda Garcia

Palabras Clave: Índice de Cetano, Red Neuronal, Espectroscopia Infrarroja, PLS, Overfitting

Descripción:

Este trabajo buscó desarrollar un modelo predictivo para determinar el índice de cetano en diésel mediante espectroscopía de infrarrojo cercano (NIR) y aprendizaje profundo, con el objetivo de crear una alternativa rápida y económica a los métodos de laboratorio tradicionales.

La metodología se centró en un conjunto de 62 muestras de diésel. Los espectros NIR fueron corregidos con filtros Savitzky-Golay y la técnica Standard Normal Variate (SNV). Se evaluaron dos enfoques: Redes Neuronales Convolucionales (CNN) y una regresión de Mínimos Cuadrados Parciales (PLS). El rendimiento de los modelos fue medido con validación cruzada y un conjunto de prueba final para asegurar una evaluación imparcial.

Los resultados principales indicaron un fallo en el desarrollo del modelo. Las arquitecturas CNN sufrieron de un sobreajuste (overfitting) severo, incapaces de aprender debido al limitado número de datos. Por su parte, el modelo PLS, aunque mostró un rendimiento promisorio en la validación interna (R^2 de 0.51), demostró una incapacidad total de generalización en la prueba final, con un R^2 negativo de -0.1763.

La conclusión fundamental es que el objetivo de crear un modelo predictivo viable no se alcanzó. El factor crítico fue el insuficiente tamaño del conjunto de datos, que impidió construir un modelo robusto. Se demostró que la predicción de este parámetro es una tarea compleja que requiere una cantidad significativamente mayor de muestras para lograr resultados fiables y aplicables en la industria.

---

[*] Trabajo de Grado

[**] Facultad de Ingenierías Fisicoquímicas. Escuela de Ingeniería Química. Director: Helver Crispiniano Álvarez Castro. Doctor en Ingeniería Química. Codirector: Giovanni Morales Medina. Doctor en Ingeniería Química

Abstract

Title: Convolutional artificial neural network model for predicting the cetane index of diesel samples using NIR spectra.

Author: Javier Danilo Aranda Garcia

Key Words: Cetane Index, Neural Network, Spectroscopic Infrared, PLS, Overfitting

Description:

This work aimed to develop a predictive model to determine the cetane index in diesel using near-infrared (NIR) spectroscopy and deep learning, with the goal of creating a fast and cost-effective alternative to traditional laboratory methods. The methodology focused on a set of 62 diesel samples. The NIR spectra were corrected using Savitzky-Golay filters and the Standard Normal Variate (SNV) technique. Two approaches were evaluated: Convolutional Neural Networks (CNN) and Partial Least Squares (PLS) regression. Model performance was measured using cross-validation and a final test set to ensure an unbiased evaluation.

The main results indicated that the CNN architectures suffered from severe overfitting, unable to learn due to the limited amount of data. The PLS model, on the other hand, although it showed promising performance in internal validation (R^2 of 0.51), demonstrated limited generalization ability in the final test, with a negative R^2 of -0.1763 attributed to overfitting.

The key conclusion is that the goal of creating a viable predictive model was not achieved. The critical factor was the insufficient dataset size, which prevented the construction of a robust model. It was shown that predicting this parameter is a complex task that requires a significantly larger number of samples to achieve reliable and industry-applicable results.

---

[*] Degree Work

[**] Facultad de Ingenierías Fisicoquímicas. Escuela de Ingeniería Química. Director: Helver Crispiniano Álvarez Castro. Doctor en Ingeniería Química. Codirector: Giovanni Morales Medina. Doctor en Ingeniería Química

Introducción

El creciente desafío de mejorar la eficiencia y reducir el impacto ambiental en el sector energético ha impulsado la búsqueda de métodos innovadores para el control de calidad de combustibles fósiles. En Colombia, el diésel sigue siendo un insumo esencial en diversas aplicaciones industriales y de transporte. Sin embargo, su uso plantea importantes retos, ya que la calidad de su combustión afecta directamente tanto al rendimiento del motor como a la emisión de contaminantes como óxidos de nitrógeno (NOx) y material (Herrera Susa et al., 2020; Ministerio de Ambiente y Desarrollo Sostenible, 2021).

Uno de los parámetros clave para evaluar esta calidad es el índice de cetano, que mide la capacidad del combustible para auto encenderse bajo compresión (ASTM International, 2021). Un índice de cetano óptimo asegura un retardo de ignición corto, lo que se traduce en un arranque más suave del motor, una combustión más completa, mayor eficiencia energética y una reducción significativa de emisiones nocivas (Herrera Susa et al., 2020).

Los métodos tradicionales para determinar el índice de cetano, como el estándar ASTM D613-23 (ASTM International, 2023) y el índice calculado ASTM D4737-21 (ASTM International, 2021), si bien son precisos, dependen de equipos de motor especializados, son destructivos y requieren procesos prolongados, incrementando considerablemente los costos y tiempos de análisis. Frente a estas desventajas, surge la necesidad de desarrollar alternativas analíticas rápidas, no destructivas y rentables.

En este contexto, el presente trabajo propone una alternativa que supera las limitaciones de los enfoques quimio métricos tradicionales. Dichos métodos, como la regresión PLS, a menudo requieren un preprocesamiento extenso y una ingeniería de características manual para aislar la

información relevante del ruido y las interferencias presentes en los espectros crudos. Este preprocesamiento, si bien es efectivo, introduce pasos adicionales y puede descartar información sutil pero valiosa.

Para abordar este desafío, se desarrolla un modelo predictivo basado en redes neuronales convolucionales (CNN) que analiza directamente los espectros de infrarrojo cercano (NIR) (Bishop, 1994; Ponce, 2010). Se opta por esta arquitectura precisamente porque está diseñada para trabajar con datos de alta dimensionalidad, como un espectro completo, aprendiendo a extraer características relevantes de forma automática. De esta manera, la CNN aprovecha la totalidad de la "huella espectral" del combustible, minimizando la necesidad de preprocesamientos complejos y eliminando la subjetividad asociada a la selección manual de variables.

Esta propuesta no solo busca ser una alternativa viable para reducir los costos y tiempos asociados a los métodos convencionales, sino que también aporta una herramienta validada que puede integrarse en procesos de control de calidad en tiempo real, promoviendo la transición de la industria hacia operaciones más eficientes y sostenibles (Domínguez et al., 2020).

1. Objetivos

1.1. General

Desarrollar un modelo predictivo basado en redes neuronales convolucionales para determinar el índice de cetano en muestras de diésel utilizando espectros de infrarrojo cercano (NIR), con el fin de optimizar tiempos y costos en comparación con métodos tradicionales.

1.2. Específicos

- Recopilar y preprocesar los espectros NIR de muestras de diésel, aplicando técnicas de normalización para garantizar la calidad y consistencia de los datos.
- Implementar y entrenar un modelo CNN en Python, utilizando técnicas de regularización para mejorar la capacidad predictiva del modelo en la estimación del índice de cetano (Ng, 2004).
- Evaluar el rendimiento del modelo mediante métricas estadísticas y validación cruzada (K-Fold), comparándolo con métodos tradicionales para determinar su viabilidad y aplicación práctica en el control de calidad de combustibles (Kohavi, 1995).

2. Estado del Arte

Hosseinpour et al., 2016. Desarrollo un modelo de predicción del índice de cetano, usando el contenido de esteres metílicos de ácidos grasos (FAME). Para ello usaron el enfoque aproximado de mínimos cuadrados parciales (PLS) adaptado a una red neuronal desarrollada en el software de cómputo matemático MATLAB, de esta manera se pudo determinar la relación entre el índice de cetano con el contenido de FAME de las muestras. Usando los criterios de coeficiente de determinación (R2), el error cuadrado medio (MSE) y el error porcentual (PE). Concluyendo

que el modelo planteado de PLS con ANN tiene una aproximación cercana del CN a comparación del modelo clásico de solo PLS, teniendo como resultados de MSE, PE, R2, de

0.7266,1.06% y 0.9934 respectivamente para todas las muestras analizadas, indicando que es una herramienta factible como alternativa al método tradicional de laboratorio para la determinación del CN (Hosseinpour et al., 2016).

Diaz, Julieth A., 2015. Analizo la predicción del índice de cetano del diésel usando los datos proporcionados por la refinería de Barrancabermeja. Mediante la aplicación de métodos de regresión lineal y no lineal (Redes Neuronales), en donde aplicaron la regresión lineal a datos macroscópicos (Densidad, Viscosidad, índice de refracción, etc.) usando el software R y la regresión no lineal a los datos proporcionados por espectroscopia infrarroja NIR usando el software MATLAB. Según los resultados, se pudo evidencian una aproximación del valor real reportado por las normas ASTM D4737. Concluyendo que esta es una forma alternativa a la norma ASTM D-976, ya que esta metodología muestra una mejor predicción y aproximación al valor real de la muestra (Díaz Serrano, 2015).

Sánchez et al., 2012. Desarrollaron un modelo analítico que permite estimar el número de cetano de Biodiésel y aceites vegetales de diferentes fuentes naturales, para ello usaron el software estadístico Statgraphics Centurion XV.I, en el cual se modelaron las ecuaciones de regresión lineal para la predicción de la variable objetivo en función del contenido de ácidos grasos presentes en la muestra. Para determinar el contenido de ácidos grasos se usó un Cromatógrafo de Gases Carlo Erba 8065 equipado con detector FID y columna capilar HT8, en el cual se analizaron las diferentes muestras de Biodiésel, ácidos grasos puros, y aceites vegetales. Concluyeron que el contenido de ácidos grasos en el Biodiésel influye en el índice de cetano, para lo cual el modelo propuesto indica

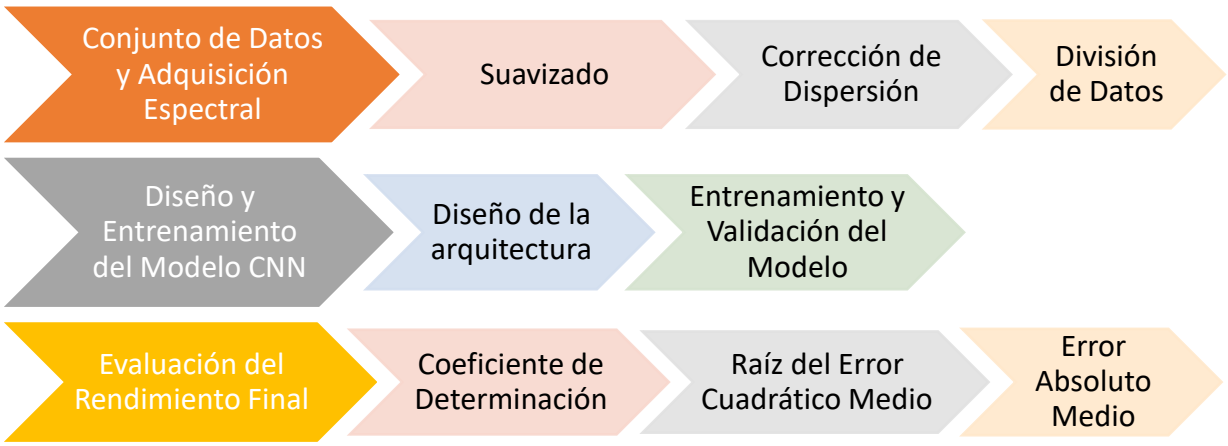
una precisión del 99.6% para Biodiésel y de 98.4 % para el caso de aceites vegetales (Sánchez et al., 2012).

3. Metodología

El desarrollo del modelo predictivo para el índice de cetano se llevó a cabo siguiendo un flujo de trabajo sistemático, diseñado para garantizar la robustez y replicabilidad del estudio. Las etapas comprendieron la preparación de los datos, el diseño y entrenamiento del modelo CNN, y la evaluación rigurosa de su rendimiento.

Figura 1

Descripción secuencial de la metodología.



3.1. Conjunto de Datos y Adquisición Espectral

Se utilizo un conjunto de 62 muestras de diesel comercial, obtenidas de una fuente confidencial dentro de la industria. Adicionalmente, se contó con el valor de referencia del índice de cetano para cada muestra determinado por medio de laboratorio por la fuente. Para

cada muestra, se registró la intensidad del espectro en un rango de [100.8 a 11998.764 nm] con un rango máximo de absorbancia de 30.

Antes de la construcción del modelo, los espectros crudos fueron sometidos a una serie de etapas de preprocesamiento con el fin de corregir interferencias físicas y químicas. El proceso, implementado en Python con las librerías Scikit-learn y SciPy, incluyó:(Chollet, 2015; Géron, 2019)

3.1.1. Suavizado (Smoothing)

Se aplicó un filtro de **Savitzky-Golay** (ventana de 20 puntos, polinomio de grado 1, primera derivada) para reducir el ruido instrumental.

3.1.2. Corrección de Dispersión (Scatter Correction)

Se utilizó la técnica Standard Normal Variate (SNV) para normalizar cada espectro, minimizando los efectos de la dispersión de la luz y las variaciones en la longitud del camino óptico.

3.1.3. División de Datos

El conjunto de datos completo (espectros y valores de cetano) se dividió aleatoriamente en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%). El conjunto de prueba se mantuvo separado y no se utilizó en ninguna fase del entrenamiento o validación para asegurar una evaluación final imparcial del modelo.

3.2. Diseño y Entrenamiento del Modelo

Se diseño un modelo que usa PLS que ofrece mejor rendimiento en conjuntos de datos con pocos registros.

3.2.1. Diseño de la arquitectura

Se definieron los parámetros de entrada al modelo, así como la estructura de entrenamiento usando la validación cruzada (K-Fold) (Kohavi, 1995). Así como las métricas que se evaluaran para cada Fold.

3.2.2. Entrenamiento y Validación del Modelo

El modelo se compiló utilizando el optimizador Adam con una tasa de aprendizaje de 0.0005 y la función de pérdida de Error Cuadrático Medio (MSE).

3.3. Evaluación del Rendimiento Final

La capacidad predictiva del modelo CNN final fue evaluada utilizando el conjunto de prueba (20%), que el modelo no había visto previamente. Se calcularon las siguientes métricas de rendimiento para cuantificar la precisión y el ajuste del modelo.

3.3.1. Coeficiente de Determinación (R2)

Para medir la proporción de la varianza en el índice de cetano que es predecible a partir de los espectros.

3.3.2. Raíz del Error Cuadrático Medio (RMSE)

Para evaluar la desviación promedio entre los valores predichos y los valores reales, en las mismas unidades que el índice de cetano.

4. Análisis de resultados

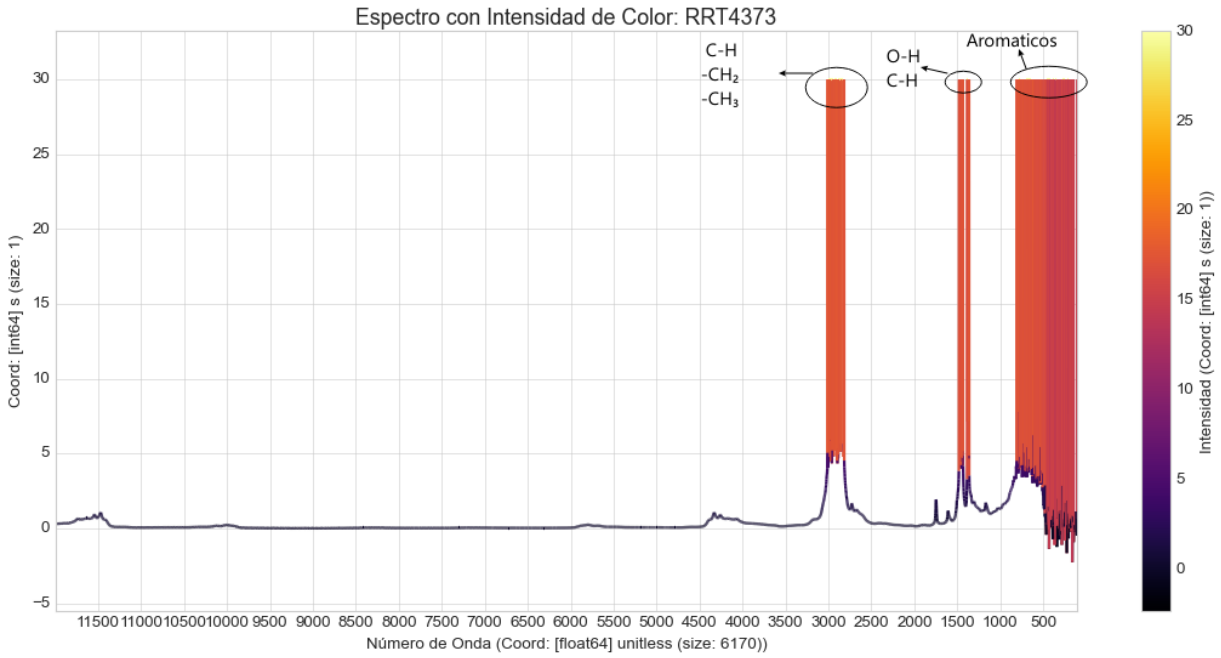
4.1. Análisis espectroquimico

Con el objetivo de garantizar la calidad y consistencia de los datos, se implementó un riguroso pipeline de preprocesamiento sobre los espectros NIR crudos recopilados de las muestras de diésel (Ver apéndices A y B). El análisis inicial de los espectros originales reveló la presencia

de la huella molecular característica del diésel (Figura 2), permitiendo la identificación de grupos funcionales claves asociados a su calidad fisicoquímica (Smith, 2011). Sin embargo, también se observaron variaciones de línea base y ruido instrumental que requerían corrección para un modelado robusto.

Figura 2

Análisis espectroquimico de las muestras



Nota. Espectro IR representativo del diésel analizado. Se resaltan las regiones clave correspondientes a compuestos aromáticos (UV-Vis), agua y sobretonos C-H (NIR), y las vibraciones fundamentales de alcanos (MIR) que definen la calidad del combustible.

4.1.1. Región UV-Visible y NIR Inicial (100-900 nm)

En esta zona, la absorbancia es dominada por las transiciones electrónicas de compuestos aromáticos. Una mayor intensidad aquí es un indicador directo de una alta concentración

de estructuras anilladas, las cuales son conocidas por reducir el índice de cetano y aumentar la producción de hollín. Por tanto, esta región nos permite cuantificar los componentes que degradan la calidad del combustible.

4.1.2. Región del Infrarrojo Cercano (NIR) (1300-1500 nm)

Esta ventana es crucial para el monitoreo de contaminantes y de la composición general. La banda ancha alrededor de 1450 nm corresponde al primer sobretono del enlace O-H, señalando la presencia de trazas de agua. Adyacente a esta, las señales en ~1400 nm se asocian a combinaciones de enlaces C-H de alcanos. La relación entre estas señales permite evaluar tanto la pureza como la composición alifática del diésel.

4.1.3. Región del Infrarrojo Medio (MIR) (2800-3100 nm)

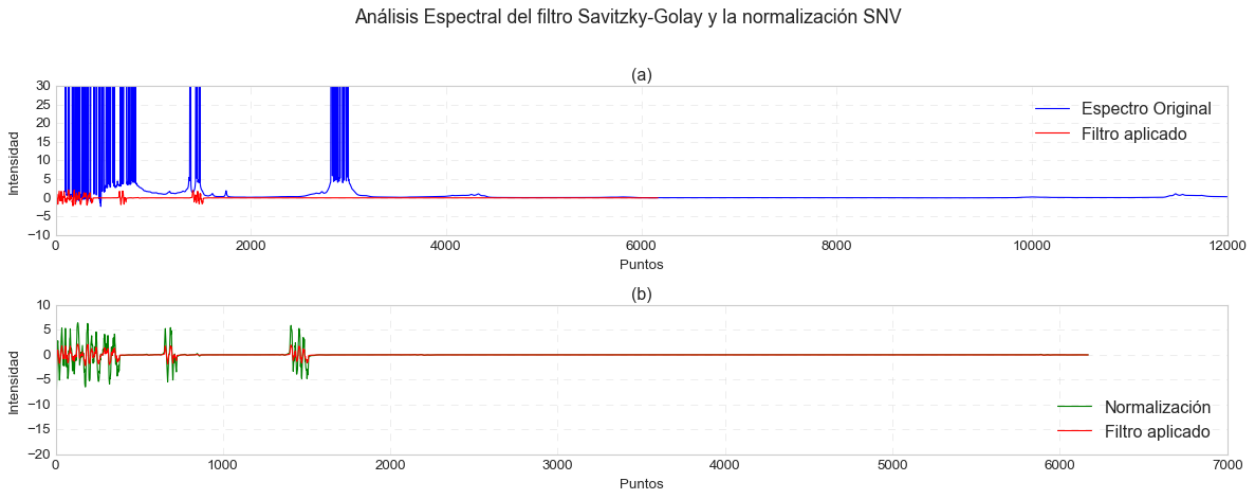
Esta es la región más informativa sobre la estructura de los hidrocarburos, ya que contiene las vibraciones de estiramiento fundamentales de los enlaces C-H. Específicamente, el área entre 2850 y 2960 nm es dominada por las señales de grupos metilo ($-CH_3$) y metileno ($-CH_2$), componentes básicos de las parafinas. La intensidad integrada en esta zona es el proxy más directo del contenido parafínico del combustible; una señal fuerte y definida aquí es el principal indicador de un alto índice de cetano y una buena calidad de ignición.

Se aplicó un filtro Savitzky-Golay para reducir el ruido de alta frecuencia y, simultáneamente, calcular la primera derivada del espectro. Se utilizaron los parámetros de polyorder=2, window_length=20 y deriv=1. El uso de la primera derivada fue crucial para eliminar los desplazamientos aditivos de la línea base y para acentuar las regiones de cambio correspondientes a los picos de absorción relevantes. (Figura 3)

Posteriormente, se aplicó la técnica SNV a los espectros derivados. Este paso normalizó cada espectro de forma individual, corrigiendo eficazmente los efectos de dispersión multiplicativos causados por variaciones físicas entre las muestras. (Figura 3)

Figura 3

Comparación de filtro y normalización aplicados.



Nota. (a) Comparación entre la muestra original, y el filtro Savitzky-Golay en donde se observa la reducción de la dimensión de la muestra. (b) Comparación de la escala entre el filtro aplicado y la normalización SNV, se observa el aumento de escala de la muestra.

4.2. Selección del Modelo Óptimo

Para cuantificar la relación entre los espectros NIR preprocesados y la variable objetivo, se desarrolló un modelo de regresión basado en la técnica de Mínimos Cuadrados Parciales (PLS). Se seleccionó este método por su probada eficacia en el manejo de datos con alta colinealidad, como es el caso de los datos espectrales.

La robustez y capacidad de generalización del modelo se evaluaron rigurosamente mediante un procedimiento de K-Fold. Los datos de entrenamiento se dividieron aleatoriamente

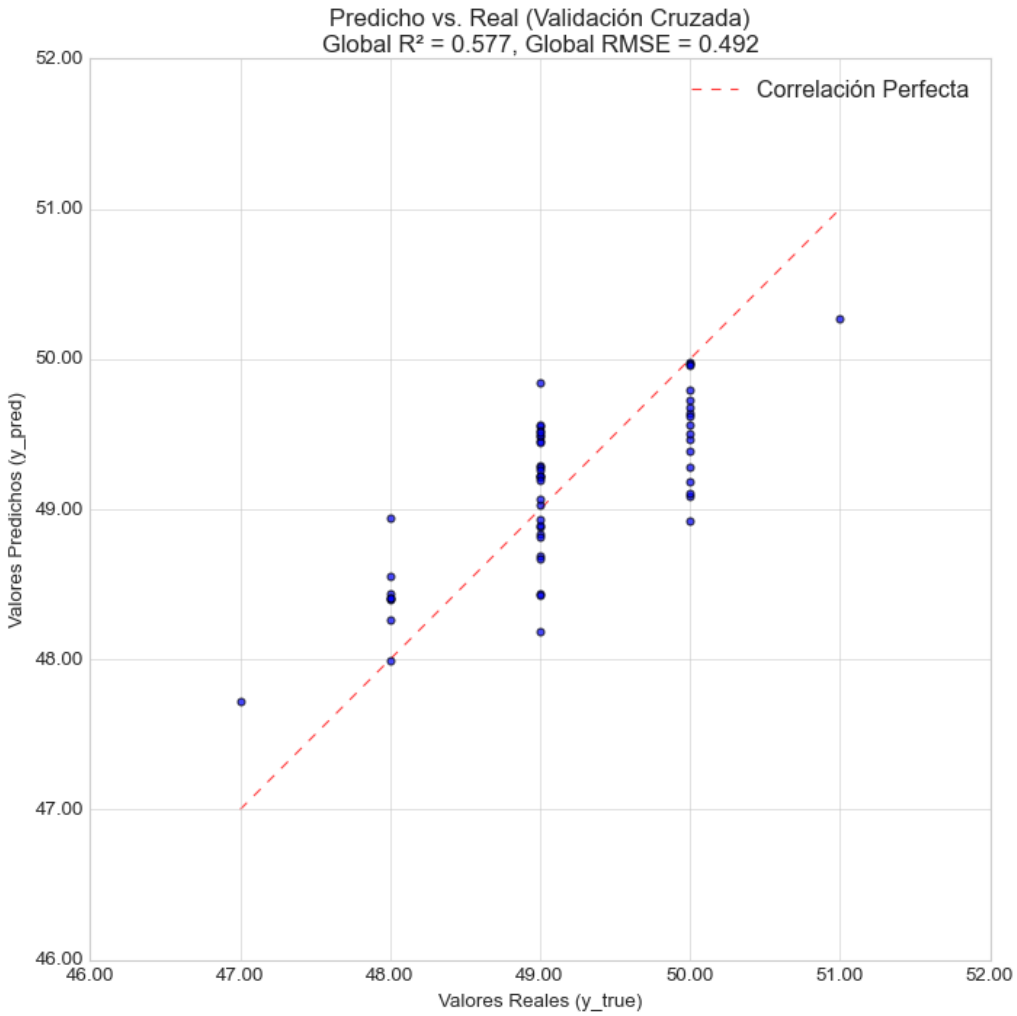
(shuffle=True, random_state=60) en múltiples pliegues para asegurar que la evaluación no dependiera de una partición de datos específica.

Para la arquitectura del modelo, se optimizó el número de componentes latentes, estableciendo un valor final de 7 componentes (n_components=7) para capturar la máxima covarianza predictiva entre los espectros y la variable de interés. El rendimiento del modelo se cuantificó a través del R^2 y el RMSE, promediados a lo largo de todos los pliegues de la validación.

Los resultados de la validación cruzada arrojaron un R^2 promedio de 0.5061 (+/- 0.1387) y un RMSE promedio de 0.4906 (+/- 0.0462). Estos valores indican una prometedora capacidad del modelo para predecir la calidad del diésel a partir de su huella espectral. (Figura 4 y 5)

Figura 4

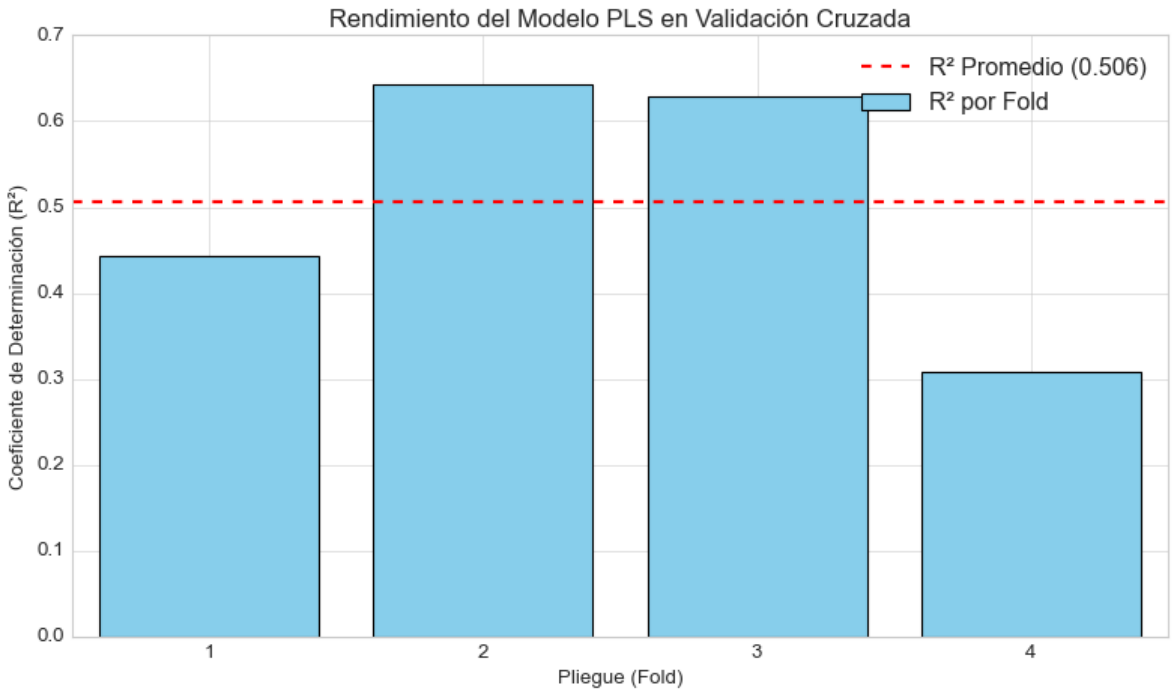
Correlación de modelos durante el entrenamiento.



Nota: Correlación de los modelos durante la fase de entrenamiento. Mostrando que los modelos se aproximaron a los valores de prueba de cada sesión.

Figura 5

Rendimiento de cada pliegue(fold), usado para entrenar el modelo.



4.3. Contraste de Rendimiento: La Robustez de PLS frente a la Inestabilidad de CNN

Tras la optimización del modelo, se procedió a una validación final y definitiva utilizando el conjunto de datos de prueba (hold-out set), el cual fue excluido de todas las etapas previas de entrenamiento y validación cruzada.

La evaluación del rendimiento en estos datos completamente nuevos arrojó un R^2 de -0.1763 y un RMSE de 0.2941. Estos resultados confirman la capacidad de generalización del modelo y representan la estimación final de su rendimiento predictivo en condiciones reales. (Figura 6)(Tabla 1)

Figura 6

Correlación entre los valores predichos de muestras no vistas por el modelo.

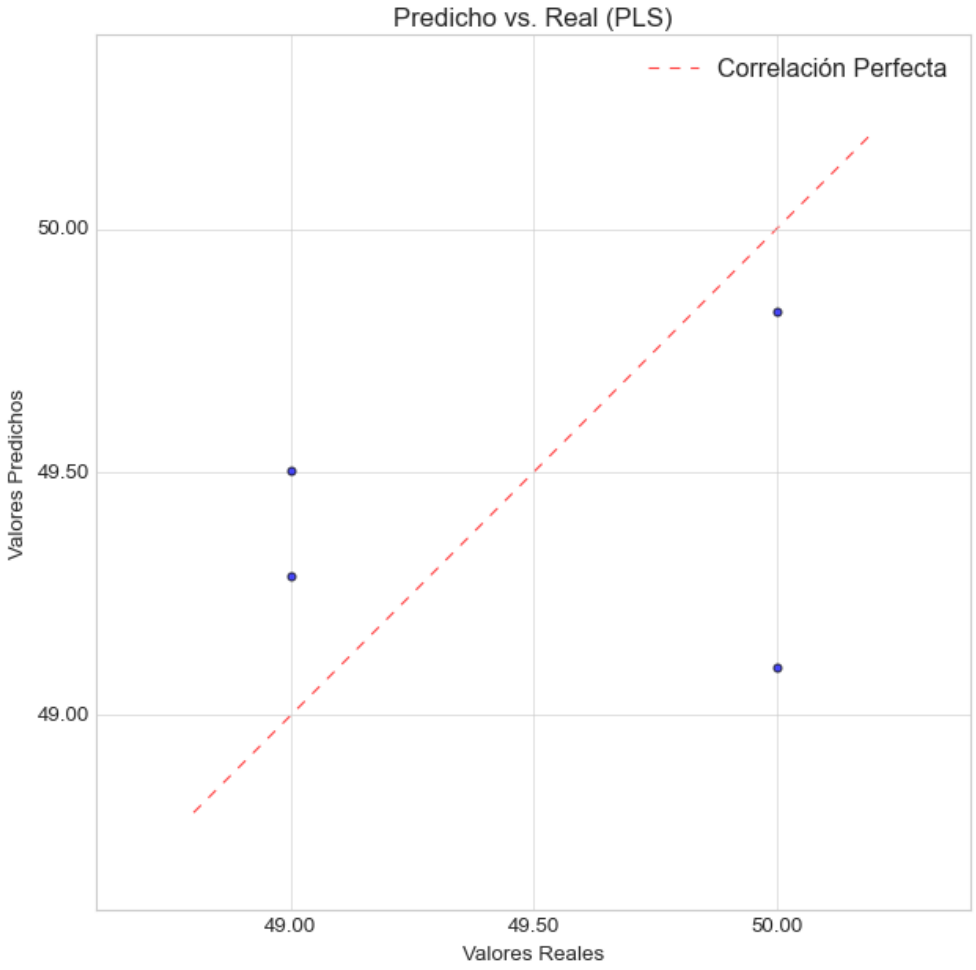


Tabla 1

Errores de truncamiento en la predicción del modelo

Muestra	***Valor real***	***Valor predicho***	***Error truncamiento [%]***
RRT4368	49	49.5	1.02%
RRT4369	49	49.29	0.59%
RRT4371	50	49.1	1.80%
RRT4372	50	49.83	0.34%

5. Conclusiones

- Se consolidó un conjunto de datos de calidad, aplicando las correcciones y normalizaciones necesarias para asegurar la integridad de la información utilizada en la etapa de modelado, lo cual representa un activo para futuras investigaciones.
- La implementación y entrenamiento de las arquitecturas CNN incurrieron en un severo sobreajuste (*overfitting*), atribuible a la alta complejidad de la red en relación con el tamaño del conjunto de datos. Esto impidió que los modelos aprendieran las relaciones fisicoquímicas fundamentales, resultando en su fallo predictivo.
- Se cuantificó el rendimiento de cada arquitectura, evidenciando que tanto los modelos CNN como el modelo PLS optimizado no son viables. A pesar de que el modelo PLS (R^2: de 0.5061 +/- 0.1387; RMSE: 0.4906 +/- 0.0462) fue numéricamente superior, su bajo poder predictivo lo descarta para fines prácticos.
- El desarrollo de un modelo predictivo viable basado en CNN para la determinación del índice de cetano no se alcanzó en su totalidad. Aunque se implementaron y evaluaron diversas arquitecturas, estas demostraron una incapacidad de generalización, resultando en un rendimiento predictivo nulo ($R^2 < 0$) (Ver apéndices C y D).

  El modelo de PLS se presentó como la mejor alternativa encontrada, al ser optimizado se determinaron los valores que ofrecían el mejor rendimiento con el conjunto de datos. Sin embargo, a pesar de los esfuerzos, este modelo alcanzó un rendimiento limitado. La capacidad predictiva es insuficiente para su aplicación práctica y fiable en el sector industrial.

  El hallazgo principal del proyecto es, por tanto, la demostración de la alta complejidad que presenta la predicción de este parámetro bajo las condiciones de datos estudiadas.

- Se concluye adicionalmente que el número limitado de muestras disponibles fue un factor crítico que afectó negativamente el rendimiento de todos los modelos. Para arquitecturas complejas como las CNN, que son inherentemente dependientes de grandes volúmenes de datos, un conjunto de datos pequeño impide una generalización adecuada. De igual forma, aunque el modelo PLS es más robusto en estas condiciones, su capacidad para construir una calibración fiable también se ve comprometida si las muestras no abarcan toda la variabilidad química posible del diésel. Por lo tanto, se considera que la expansión significativa del conjunto de datos es un requisito indispensable para futuros intentos de desarrollar un modelo predictivo robusto para este parámetro.

6. Recomendaciones

- Los modelos de CNN son también conocidos como “Data hungry”, esto significa que requieren una cantidad de datos significativa para realizar el proceso de aprendizaje profundo. Por ello es necesario obtener más muestras o tener un conjunto de datos mayor desde el cual el modelo pueda generalizar las características de estos.
- Si bien los modelos de aprendizaje profundo son complejos, esta complejidad aporta a la generación del sobre ajuste. Debido a que tienen tantos hiperparametros (tasa de aprendizaje, número de capas, filtros, etc.), que se requiere un estudio con un costo computacional alto para determinar los valores óptimos de estos.

Referencias Bibliográficas

ASTM International. (2021). Standard Test Method for Calculated Cetane Index by Four Variable Equation. https://www.astm.org/d4737-21.html

ASTM International. (2023). Standard Test Method for Cetane Number of Diesel Fuel Oil. https://www.astm.org/standards/d613?lang=es-ES

Bishop, C. M. (1994). Neural networks and their applications. Review of Scientific Instruments, 65(6), 1803–1832. https://doi.org/10.1063/1.1144830

Chollet, F. (2015). Keras [Software]. GitHub. https://github.com/keras-team/keras

Díaz Serrano, J. A. (2015). Aplicación de la regresión lineal múltiple y las redes neuronales artificiales para la predicción del índice de cetano en el diésel utilizando propiedades macroscópicas y espectros NIR [Tesis de pregrado]. Universidad Industrial de Santander.

Domínguez, L. F., Torres, J. A., & Ramírez, C. (2020). Deep learning applied to fuel property prediction: A case study with diesel samples. Energy & Fuels, 34(3), 2305–2314. https://doi.org/10.1021/acs.energyfuels.9b03152

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Herrera Susa, D. A., Bermúdez Santaella, J. R., & Castilla Álvarez, C. E. (2020). Análisis del desempeño de la potencia y el torque de un motor diésel operando con mezclas de biodiésel de palma. Ingeniería, 25(3), 250–263. https://doi.org/10.14483/23448393.15676

Hosseinpour, S., Aghbashlo, M., Tabatabaei, M., & Khalife, E. (2016). Exact estimation of biodiesel cetane number (CN) from its fatty acid methyl esters (FAMEs) profile using partial least square (PLS) adapted by artificial neural network (ANN). Energy Conversion and Management, 124, 389–398. https://doi.org/10.1016/j.enconman.2016.07.027

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137–1145. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

Ministerio de Ambiente y Desarrollo Sostenible. (2021). Resolución 40103 de 2021. https://www.minambiente.gov.co/documento-entidad/resolucion-40103-de-2021/

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings of the 21st International Conference on Machine Learning. https://doi.org/10.1145/1015330.1015435

Ponce, P. (2010). Inteligencia artificial con aplicaciones a la ingeniería (1st ed., Vol. 1). Alfaomega.

Sánchez, Y., Piloto, R., Goyos, L., & Ferrer, N. (2012). Predicción del número de cetano de biocombustibles a partir de su composición de ácidos grasos. Tecnología Química, 32(2), 163–170.

Smith, B. C. (2011). Infrared spectral interpretation: A systematic approach (2nd ed.). CRC Press. https://doi.org/10.1201/b10776

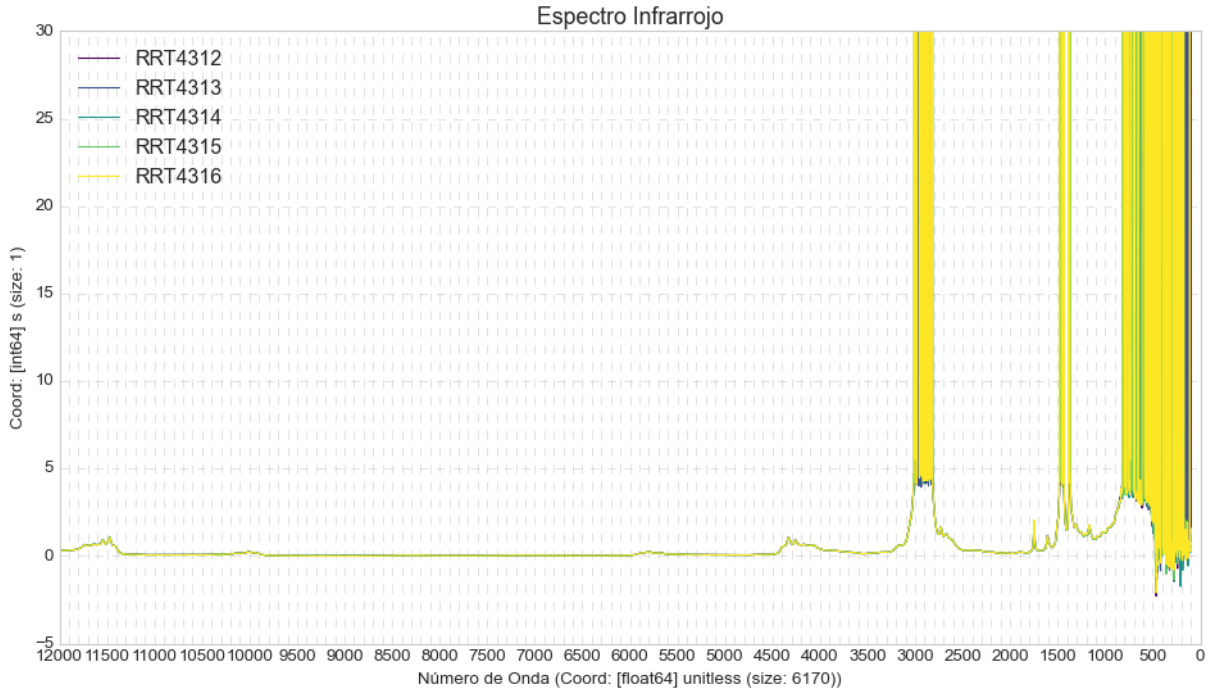
Apéndices

Apéndice A*. Base de datos de espectros NIR e índice de cetano de las muestras.*

Los apéndices están adjuntos y puede visualizarlos en la base de datos de la biblioteca UIS.

Apéndice B. *Superposición de espectros NIR de las muestras.*

Superposición de los espectros infrarrojos obtenidos para un conjunto representativo de las muestras de diésel. El eje Y representa la absorbancia, que es proporcional a la concentración de los componentes químicos, mientras que el eje X representa la longitud de onda en nanómetros (nm), que indica la región específica del espectro electromagnético. La variabilidad en la altura y forma de las curvas demuestra la diversidad en la composición molecular entre las diferentes muestras de combustible.

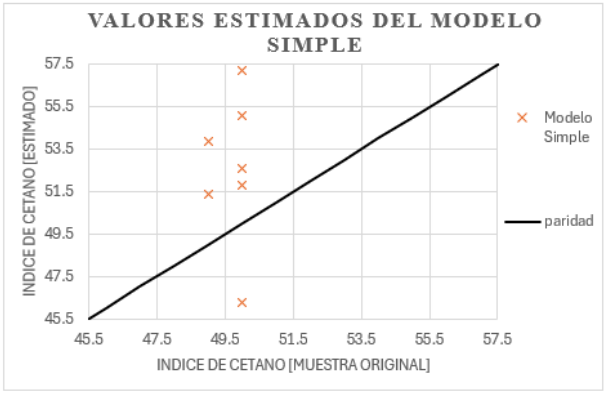


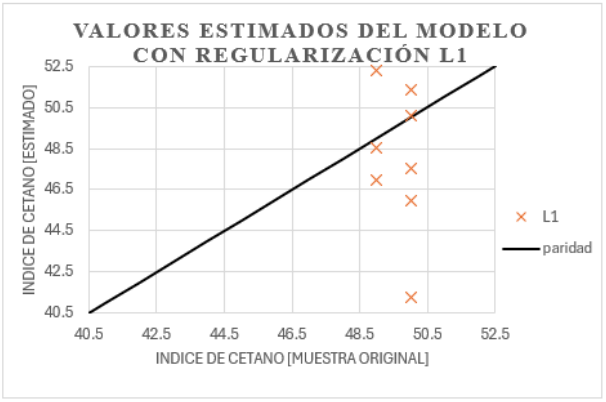
Apéndice C. *Resultados de red neuronal convencional.*

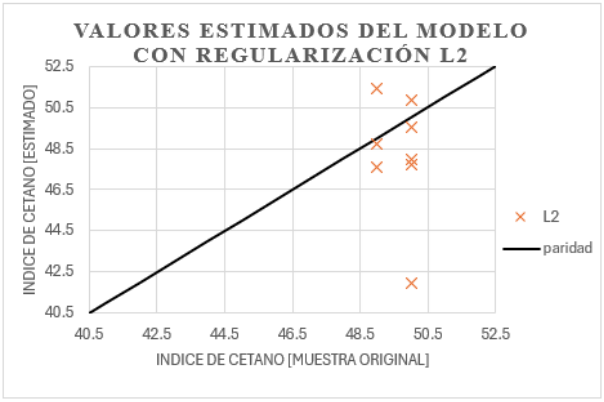
Resultados de una arquitectura diseñada con una red neuronal convencional.

Muestra	CI	CI Sin penalización	CI penalización L1	CI penalización L2
RRT4363	50	57.243378	51.33832	50.870632
RRT4364	49	57.61679	52.264442	51.448334
RRT4365	50	52.60494	47.511417	47.706203
RRT4366	50	55.073654	50.095634	49.549572
RRT4367	50	46.26327	41.274223	41.98351
RRT4368	49	51.397533	46.978783	47.623608
RRT4369	49	53.890743	48.557983	48.701252
RRT4371	50	51.82869	45.937378	47.96757
RRT4372	50	49.94590008	49.47934592	50.29274547

Correlación entre los valores de índice de cetano de referencia y los predichos por el modelo en el conjunto de prueba.







Apéndice D. *Resultados de modelos de CNN.*

Parámetros del modelo de CNN

- Capa de Entrada: Un vector con las 6170 absorbancias del espectro preprocesado.
- Primera Capa Convolucional (Conv1D): 16 filtros con un tamaño de kernel de 7 y función de activación ReLU.
- Primera Capa de Agrupación (MaxPooling1D): Con un tamaño de pool_size de 4 para reducir la dimensionalidad.
- Capa de Aplanado (Flatten): Para convertir los mapas de características 2D en un vector 1D.
- Capa Densa: Una capa totalmente conectada con 32 neuronas y activación ReLU.

- Capa de Salida: Una única neurona con activación lineal, que entrega el valor predicho del índice de cetano.

Modelos CNN probados y sus resultados. Se adjuntan solo los modelos con mejor rendimiento.

Arquitectura	**Arquitectura**	**R^2**	**RMSE**
CNN-A	1 capa conv. (16 filtros)	-1.0289 (+/- 0.7784)	1.1113 (+/- 0.0550)
CNN-B	2 capa conv. (16, 32 filtros)	-2.9113 (+/- 0.2709)	1.6240 (+/- 0.2609)

Correlación entre los valores de índice de cetano de referencia y los predichos por el modelo CNN en el conjunto de prueba.