Implementación de machine learning para modelado y caracterización de muestras complejas de hidrocarburos a partir de técnicas de espectroscopia

Sebastián Cárdenas Acevedo

Trabajo de Grado para Optar al Título de Ingeniero de Sistemas.

Director

Enrique Mejía Ospino

(PhD) Escuela Química. UIS

Codirector

Yesid Paul Goyes Peñafiel

PhD(c) en Ciencias de la Computación.

Escuela de Ingeniería de Sistemas e Informática. UIS

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Programa Académico

Bucaramanga

2024

Dedicatoria

Dedico este proyecto a Dios y a mi familia, quienes han estado presentes en cada paso que doy y en cada decisión que tomo; siempre los llevo en mi mente y en mi corazón.

A mi mamá, la persona más paciente que conozco, cuyas palabras de aliento siempre me ayudaron a esclarecer mis objetivos y mantenerme firme en mis decisiones. Ella me demuestra que, a través de sus ojos, siempre voy a ser suficiente independientemente de mis defectos.

A mi padre, que con su tenacidad diaria me enseña que la persistencia y la resiliencia son las virtudes más importantes para triunfar ante cualquier situación.

Y a mi hermano, la persona que más apoyo me ha brindado, no solo en la realización de este trabajo de grado sino en la vida, quien es mi mejor amigo y me ayuda e impulsa diariamente a superar mis limitaciones, quien me explica las cosas con amor y a su estilo, aunque yo no siempre sea la persona más receptiva y quien sin importar las situaciones adversas me ha demostrado que siempre puedo contar con él. Gracias Nico

Agradecimientos

Agradezco profundamente a la Universidad Industrial de Santander por formarme como líder profesional y darme las herramientas para afrontar nuevos retos con confianza y entusiasmo.

A mi director de proyecto Enrique Mejía por su acompañamiento y apoyo en el desarrollo del proyecto.

A mi codirector Paul Goyes, por sus consejos y recomendaciones en el ámbito computacional e investigativo.

A mi padre Fernando, inquebrantable, persistente y optimista, que con sus palabras de aliento siempre me anima a nunca bajar los brazos.

A mi madre Luz Carmen, que me demuestra diariamente su amor más puro e incondicional y que, ante los momentos negativos, siempre me recuerda mantener el foco en los aspectos más importantes de mi vida. Definitivamente eres Luz.

Y a mi hermano Nicolás, mi mejor amigo en el mundo, mi maestro de vida y la persona que más admiro.

.

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Resumen

Título: Implementación de machine learning para modelado y caracterización de muestras complejas de hidrocarburos a partir de técnicas de espectroscopia[*]

Autor: Sebastián Cárdenas Acevedo[**]

Palabras Clave: Aprendizaje automatizado, Carbono Conradson Residual, Espectroscopia, Fraccionamiento S.A.R.A, Índice de estabilidad de asfáltenos, Muestras complejas.

Descripción: Una correcta caracterización fisicoquímica del petróleo es crucial para optimizar procesos como la producción, el transporte y la refinación en la industria de los hidrocarburos. Entre las pruebas de caracterización de crudos pesados destacan el análisis SARA, el porcentaje de carbono Conradson residual (%CCR) y el índice de clase de estabilidad de asfáltenos (ASCI). Aunque efectivos, estos métodos presentan limitaciones como altos costos, largos tiempos de respuesta y el uso de solventes peligrosos. Para superar estas dificultades, este estudio explora el uso de modelos de machine learning (ML) aplicados a datos de espectrometría de masas de alta resolución (HR-MS) y espectroscopía infrarroja por transformada de Fourier (MIR-FTIR). El objetivo principal es desarrollar modelos predictivos que estimen estas propiedades del petróleo a partir de datos espectrales. La metodología desarrollada incluye la construcción de una base de datos robusta, procesada y normalizada, utilizando técnicas como el análisis de componentes principales (PCA) para mejorar la exactitud y reducir la dimensionalidad de los datos. Se evaluaron diferentes algoritmos de ML para clasificación y regresión, tales como SVC, LDA, SVR, PLS y redes neuronales, con el fin de identificar los modelos más adecuados. Los resultados muestran que los modelos basados en ML, en particular SVC y LDA, mejoran significativamente la exactitud y eficiencia en la predicción de propiedades fisicoquímicas del petróleo, superando los métodos tradicionales. Se concluye que la integración de machine learning con técnicas espectroscópicas ofrece una alternativa más rápida y de menor impacto ambiental para caracterizar grandes volúmenes de muestras complejas, reduciendo riesgos y mejorando la reproducibilidad en comparación con las pruebas convencionales.

---

[*] Trabajo de Grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas. Programa académico. Director: Enrique Mejía Ospino. Doctor en Ciencias Químicas. Codirector: Yesid Paul Goyes Peñafiel. Doctor (c) en Ciencias de la Computación.

Abstract

Title: Implementation of Machine Learning for Modeling and Characterization of Complex Hydrocarbon Samples Using Spectroscopic Techniques [*]

Author(s): Sebastián Cárdenas Acevedo[**]

Key Words: Complex Samples, Conradson Carbon Residue, Index of Asphaltene Stability, Machine Learning, SARA Analysis, Spectroscopy

Description: A proper physicochemical characterization of crude oil is crucial for optimizing processes such as production, transportation, and refining in the hydrocarbon industry. Among the characterization tests for heavy crudes, the SARA analysis, the Conradson Carbon Residue percentage (%CCR), and the Asphaltene Stability Class Index (ASCI) stand out. Although effective, these methods present limitations such as high costs, long response times, and the use of hazardous solvents. To overcome these challenges, this study explores the use of machine learning (ML) models applied to high-resolution mass spectrometry (HR-MS) and mid-infrared Fourier-transform infrared spectroscopy (MIR-FTIR) data. The main objective is to develop predictive models to estimate these crude oil properties from spectral data. The developed methodology includes building a robust, processed, and normalized database using techniques such as Principal Component Analysis (PCA) to improve accuracy and reduce data dimensionality. Various ML algorithms for classification and regression, such as SVC, LDA, SVR, PLS, and neural networks, were evaluated to identify the most suitable models. The results show that ML-based models, particularly SVC and LDA, significantly improve accuracy and efficiency in predicting the physicochemical properties of crude oil, surpassing traditional methods. It is concluded that integrating machine learning with spectroscopic techniques offers a faster and more environmentally friendly alternative to characterize large volumes of complex samples, reducing risks and improving reproducibility compared to conventional tests.

---

[*] Degree Work

[**] Faculty of Physical-Mechanical Engineering. School of Systems and Computer Engineering. Director: Enrique Mejía Ospino. PhD Chemical Science. Co-director: Yesid Paul Goyes Peñafiel. PhD (c) Computer Science.

Introducción

El estudio de las propiedades fisicoquímicas de muestras de crudo es de gran importancia para la industria petrolera, ya que permite establecer precios de ventas y plantear estrategias para su producción, transporte, refinación y comercialización, entre otras. Por ejemplo, permite prevenir o corregir problemas como el aseguramiento de flujo, la corrosión de equipos o incluso, accidentes durante el proceso de destilación por presencia de elementos no deseados (Flórez et al., 2017). Por lo tanto, una rápida y adecuada determinación de las propiedades fisicoquímicas del petróleo resulta importante, incluso para incrementar la productividad de un país (Filgueiras et al., 2014).

Dentro de los principales estudios fisicoquímicos realizados en los crudos, se encuentra el análisis SARA, el cual es un fraccionamiento que permite identificar el contenido de hidrocarburos saturados, aromáticos, resinas y asfaltenos en la muestra de petróleo; de ahí sus siglas S.A.R.A. La predicción temprana de la estabilidad de asfaltenos en el crudo mediante la utilización del análisis SARA permite una buena toma de decisión en cuanto a la implementación de métodos de prevención y un manejo apropiado de las precipitaciones de las partículas del crudo, además de reducir dificultades que se presentan en los procesos ya mencionados (Antonio et al., 2010). No obstante, es una prueba de alto costo por el uso de múltiples solventes, que además son potencialmente peligrosos para el medio ambiente si no se disponen adecuadamente. Del mismo modo, las pruebas toman en promedio dos días por muestra, son poco automatizables y tienen baja repetibilidad y reproducibilidad (Flórez et al., 2017).

Otro análisis importante en la industria de los hidrocarburos es el índice de clase de estabilidad de asfaltenos (ASCI, por sus siglas en inglés) el cual permite determinar la estabilidad

de los asfaltenos dentro de la mezcla de hidrocarburos y otros compuestos que forman al petróleo, conocer esta propiedad se vuelve útil para clasificar la calidad de una muestra de crudo (Niño et al., 2019). Sin embargo, al igual que el análisis SARA, conlleva largos tiempos de respuesta y el uso extendido de solventes potencialmente peligrosos. (Niño et al., 2019).

Otro análisis fisicoquímico que permite caracterizar el crudo es el porcentaje de Carbono Conradson Residual (%CCR), el cual indica la cantidad de coque que una muestra de crudo formará al final de su proceso de destilación en refinería. Esto permite conocer sus materiales asfalticos útiles para determinar otras propiedades como la viscosidad y calidad de la muestra. Esta prueba también presenta diversas desventajas, ya que es lenta, carece de precisión debido al ajuste de variables, además de presentar inconvenientes para la salud y el medio ambiente por la emanación de gases durante la combustión de la muestra (Noel, 1983).

En este contexto, diversas pruebas espectroscópicas se han utilizado como alternativa para caracterizar las muestras de crudo sin recurrir a las pruebas experimentales anteriormente mencionadas, entre otras (Li & Dai, 2012). Entre ellas, encontramos la técnica de espectroscopia de infrarrojo de transformada de Fourier en el infrarrojo medio (MIR-FTIR) una técnica no destructiva, sin etiquetas, altamente sensible y específica que proporciona información completa sobre la composición química de muestras biológicas. La técnica puede ofrecer información estructural fundamental y servir como una herramienta de análisis cuantitativo (De Bruyne et al., 2018). Asimismo, la Espectrometría de masas de alta resolución (HR-MS) se ha consolidado como una herramienta clave para la identificación precisa de compuestos moleculares en mezclas complejas, como el petróleo, permitiendo un análisis detallado y cuantitativo de su composición química (Hsu & Shi, 2013). Un ejemplo de aplicación de estas técnicas es el estudio de Wilt et al., (1998) quienes determinaron cuantitativamente el contenido de asfaltenos de 42 muestras de crudo

empleando espectroscopía infrarroja. Dicha predicción obtuvo un R^2 de 0.95. Las ventajas que ofrecen algunas pruebas espectroscópicas son que no requiere de preparación de muestras y el análisis se realiza en un tiempo más corto y empleando poca cantidad de crudo.

Sumado a estos antecedentes y aprovechando las capacidades del machine learning (ML), este estudio plantea implementar modelos de ML que permitan la caracterización de propiedades fisicoquímicas críticas (ASCI, SARA Y %CCR) de muestras de crudo a partir de datos de HR-MS y MIR-FTIR, al interior de la Universidad Industrial de Santander.

2. Objetivos

2.1. Objetivo General

Implementar modelos de *machine learning* que permitan la caracterización de propiedades fisicoquímicas de muestras complejas (petróleo) a partir de datos espectrométricos.

2.2. Objetivos Específicos

Construir una base de datos a partir de las pruebas espectroscópicas y análisis fisicoquímicos de muestras complejas haciendo uso de técnicas de *big data* y análisis de datos para depurar y limpiar la información.

Implementar modelos de machine learning que clasifiquen las muestras de petróleo entre crudo, fondo y gas, y estimen las propiedades fisicoquímicas del petróleo como el porcentaje carbono Conradson (%CCR), índice S.A.R.A y ASCI

Validar los modelos basados en machine learning a partir de métricas para las tareas de regresión y clasificación sobre pruebas en un conjunto de datos de laboratorio

3. Marco teórico

3.1. Caracterización de crudos con espectrometría de masas y espectroscopia de infrarrojo medio

La caracterización de crudos resulta importante en la industria petrolera en los procesos de transporte, producción, refinamiento y comercialización, por ejemplo, el conocimiento de parámetros físicos como la viscosidad, la liquidez y la presencia de partículas sólidas es crucial para prevenir problemas como la obstrucción o solidificación durante el transporte y procesamiento, otro ejemplo es la concentración de azufre en el petróleo el cual si llega a encontrarse en cantidades considerables puede llegar a corroer una planta de producción u obstruir sus tuberías.

En la actualidad, la mayoría de estas pruebas se realizan mediante métodos estándar desarrollados por la Sociedad Estadounidense de Pruebas y Materiales Internacionales (ASTM International y el Instituto de la Energía, anteriormente conocido como el Instituto del Petróleo (IP) Sin embargo, la mayoría de estos métodos son bastante lentos, elaborados y costosos, lo que requiere grandes cantidades de muestra de los crudos que se envían al laboratorio, como alternativa surgen las pruebas espectroscópicas (1995), ya que los espectros correspondientes reflejan la composición molecular completa de un crudo (Peinder, 2009).

3.1.1. Espectroscopía

La espectroscopia es el estudio de la absorción, emisión o dispersión de radiación electromagnética por átomos o moléculas, radiación que puede abarcar los rangos desde ondas de radio hasta rayos gamma, y moléculas que pueden encontrarse en fase gaseosa, liquida o sólida. Dentro de esta amplia rama de la ciencia destaca la técnica de espectroscopia de masas y de infrarrojos por su relación y eficiencia con las muestras relacionadas de petróleo crudo (Holla, 2013).

3.1.2. Espectrometría de masas

La espectrometría de masas se ha demostrado como la técnica más poderosa para la caracterización detallada a nivel molecular de mezclas complejas, lo que lleva a una comprensión de la química detrás de los procesos y la determinación de la composición molecular de los productos (Hsu & Shi, 2013).

Un espectrómetro de masas típicamente consta de tres etapas: una fuente de iones en la cual la muestra convertida se vaporiza y ioniza para hacerla analizable. Luego un analizador de masas que separa los componentes de la muestra según su masa. Finalmente, un detector, donde los iones separados se detectan y se registra la intensidad de la señal en función de la relación masa/carga (Morgan et al., 2010).

Existen diferentes métodos para ionizar una muestra, Las consideraciones más importantes son la energía interna transferida durante el proceso de ionización y las propiedades fisicoquímicas de la muestra que pueden ser ionizadas, dentro de ellos destacan la ionización química, la ionización de campo y la ionización por electrones.

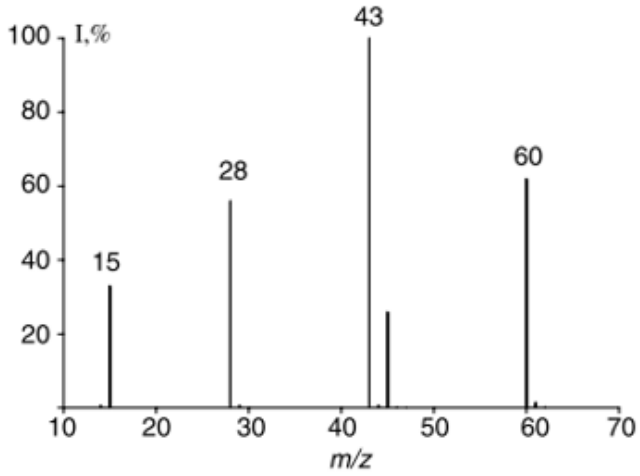
Una vez detectados los iones son transformados en una señal utilizable por un detector. Los detectores, al recibir los iones incidentes, pueden generar una corriente eléctrica proporcional a su abundancia. Actualmente, existen varios detectores que emplean diversas aproximaciones para detectar iones, según el diseño del instrumento y en las aplicaciones analíticas específicas. Esta elección de detectores se realiza considerando aspectos como la carga, la masa o la velocidad de los iones.

Como resultado se obtiene el espectro de masas el cual, contiene información sobre la masa molecular de la muestra y las masas de sus fragmentos estructurales. Los espectros de masas se representan en formato gráfico o tabular donde la abscisa representa la masa de los iones o su

relación masa-carga, m/z, mientras que la ordenada representa la intensidad relativa de los picos de estos iones, dichos picos se representan en porcentaje relativo al pico base en el espectro o a la abundancia total de todos los iones en los espectros (Ekman et al., 2002).

Figura 1.

Ejemplo de un espectro de masas de un compuesto (C2H4O2) en formato de tabla y gráfico

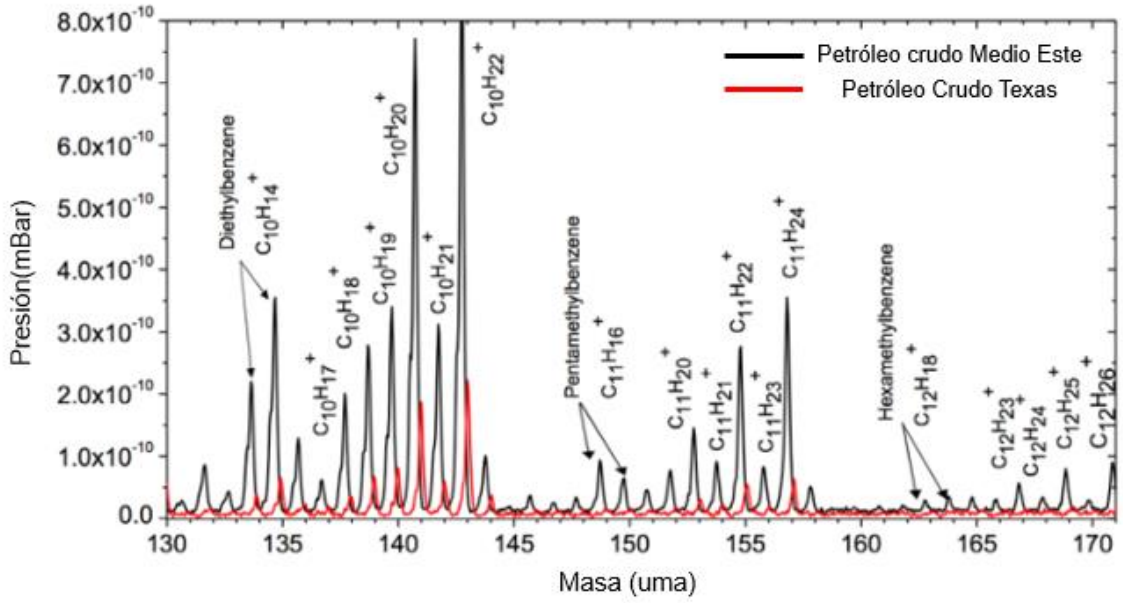


m/z	Intensidad (%)
14	0.7
15	33.0
28	56.0
29	0.65
30	0.11
43	100
44	2.2
45	26.0
46	0.31
47	0.10
60	62.0
61	1.48
62	0.26

Con esta información podemos caracterizar propiedades de diferentes gases y fluidos y sobre todo de muestras de petróleo para obtener información fisicoquímica. Por ejemplo, se ha utilizado en muestras de proteína en productos derivados del petróleo para obtener sus diferentes enzimas (dsz-A, dsz-B, y Frd-A), las cuales resultan útiles para poder detectar productos sulfurados en el petróleo y desarrollar herramientas de desulfuración biológica (Wolf et al., 1998).

Figura 2.

Análisis de petróleo crudo



También se desarrollaron análisis en petróleo crudo donde, según la cantidad de unidades de masa atómica encontradas en la muestra, se pudieron identificar y caracterizar las parafinas, naftenos, aromáticos, bencenos, toluenos y xilenos, de la muestra (Kaya et al., 2017), del mismo modo en la universidad industrial de Santander se han desarrollado pruebas de espectroscopia de masas de ultra alta resolución sobre muestras de petróleo crudo que serán objeto de este trabajo de grado para caracterizar propiedades fisicoquímicas.

3.1.3. Espectroscopia de infrarrojos

Para iniciar hablando de la espectroscopia de infrarrojos se debe iniciar hablando de la luz, esa onda electromagnética que se mueve en dos planos ortogonales de electricidad y magnetismo cuya unidad se distingue como el fotón y cuya energía se define como la relación de la frecuencia por la longitud de su onda.

De acuerdo con estas propiedades de energía, frecuencia y longitud podemos definir el tipo de onda electromagnética; también conocido como espectro electromagnético (Chu et al., 2022).

Figura 3.

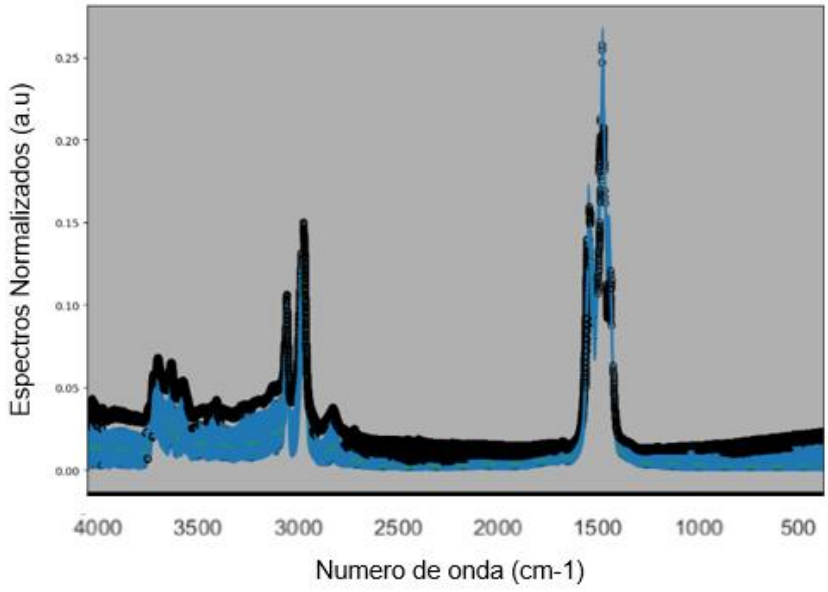
Parámetros relacionados con las ondas electromagnéticas utilizadas para el análisis espectral

E/eV	ν/Hz	λ	Tipo de Onda Electromagnética	Tipo de Transición
$> 2.5 \times 10^{5}$	$> 6.0 \times 10^{19}$	$< 0.005\,\text{nm}$	Región de rayos γ	Nivel nuclear
$2.5 \times 10^{5} \sim 1.2 \times 10^{2}$	$6.0 \times 10^{19} \sim 3.0 \times 10^{16}$	$0.005 \sim 10\,\text{nm}$	Región de rayos X	Niveles electrónicos K, L
$2 \sim 6.2$	$3.0 \times 10^{16} \sim 1.5 \times 10^{15}$	$10 \sim 200\,\text{nm}$	Región de ultravioleta lejano	Niveles de energía de electrones externos
$6.2 \sim 3.1$	$1.5 \times 10^{15} \sim 7.5 \times 10^{14}$	$200 \sim 400\,\text{nm}$	Región de ultravioleta cercano	Región de luz visible
$3.1 \sim 1.6$	$7.5 \times 10^{14} \sim 3.8 \times 10^{14}$	$400 \sim 800\,\text{nm}$	Región de luz visible	Región de luz visible
$1.6 \sim 0.50$	$3.8 \times 10^{14} \sim 1.2 \times 10^{14}$	$0.8 \sim 2.5\,\mu\text{m}$	Región de infrarrojo cercano	Niveles vibratorios moleculares
$0.50 \sim 2.5 \times 10^{-2}$	$1.2 \times 10^{14} \sim 6.0 \times 10^{12}$	$2.5 \sim 50\,\mu\text{m}$	Región de infrarrojo medio	Niveles vibratorios moleculares
$2.5 \times 10^{-2} \sim 1.2 \times 10^{-3}$	$6.0 \times 10^{12} \sim 3.0 \times 10^{11}$	$50 \sim 1000\,\mu\text{m}$	Región de infrarrojo lejano	Niveles rotacionales moleculares
$1.2 \times 10^{-3} \sim 4.1 \times 10^{-6}$	$3.0 \times 10^{11} \sim 1.0 \times 10^{9}$	$1 \sim 300\,\text{mm}$	Región de microondas	Región de microondas
$< 4.1 \times 10^{-6}$	$< 1.0 \times 10^{9}$	$> 300\,\text{mm}$	Región de ondas de radio	Espines de electrones y núcleos

Dentro de todos estos tipos de ondas destaca la región de infrarrojo medio (MIR por sus siglas en inglés) porque proporcionan información rica sobre estructura y composición, lo que es muy adecuado para la medición de parámetros fisicoquímicos de sustancias orgánicas que contienen hidrógeno, como productos agrícolas, productos petroquímicos y medicamentos (Chu et al., 2022).

Figura 4.

Diagrama espectroscopia NIR



Para ilustrar lo anterior, tenemos la ilustración 7, la cual muestra una salida típica de una espectroscopia de infrarrojos en muestras del petróleo. Típicamente utiliza ondas entre 400 y 4000 cm^{-1}, y detalla la intensidad (picos de absorbancia) frente al espectro de luz emitido (Samanta et al., 2011)

3.2. Caracterización de propiedades fisicoquímicas de muestras complejas

El petróleo se define como una muestra compleja de hidrocarburos que se halla en las rocas. Dicha muestra puede encontrarse en estado, sólido, liquido o gaseoso; además es común encontrar impurezas como el azufre o el nitrógeno (Schlumberger, n.d.-b).

También se pueden identificar diferentes propiedades físicas y químicas del petróleo para ayudar a conocer su comportamiento. Por ejemplo, el análisis SARA (Saturados, Aromáticos, Resinas y Asfaltenos), el índice ASCI (Índice de Clase de Estabilidad de Asfáltenos) y el %CCR

(porcentaje de Carbono Conradson Residual) permiten evaluar la solubilidad, estabilidad y otras propiedades del petróleo, proporcionando información valiosa para su clasificación y aplicación.

3.2.1. Análisis SARA

El análisis S.A.R.A es un método de caracterización para muestras de petróleo que consiste en separar la muestra en 4 componentes de acuerdo con su solubilidad en diferentes solventes. Tales fracciones son: saturados, aromáticos, resinas y asfáltenos (Schlumberger, n.d.-a). Dicha prueba se encuentra estandarizada por las normas ASTM D4124-09 y ASTM D2007-03.

Este análisis permite determinar una gran variedad de compuestos orgánicos dentro de las muestras de crudo. Sin embargo, esta prueba, a pesar de demostrar gran fiabilidad en resultados, ha demostrado ser costosa en tiempo y dinero (Mohammadi et al., 2021).

3.2.2. Índice ASCI

El índice de clase de estabilidad de asfáltenos, ASCI por sus siglas en inglés, es una medida utilizada para evaluar la estabilidad de los asfáltenos en petróleo crudo. Se basa en el inicio de precipitación de los asfáltenos en la muestra de crudo con diferentes proporciones de soluciones de n-heptano/tolueno y sus valores rondan entre 0 y 20 valores continuos, donde, cuanto mayor sea dicho valor, mayor será su estabilidad (Lamus et al., 2011).

3.2.3. %CCR

El porcentaje de Carbono Conradson Residual (CCR) es un parámetro utilizado, al igual que los anteriores, para conocer el crudo y poder clasificarlo. Este se caracteriza por determinar la cantidad de residuos de carbono o coque que un crudo puede formar bajo condiciones de degradación térmica, sin embargo, hallar esta propiedad carece de precisión, es lento y requiere cantidades grandes para poder llevarse a cabo (Noel, 1983).

La siguiente tabla sintetiza los principales laboratorios y sus características.

Tabla 1.

Resumen de laboratorios del estado del arte: análisis de ventajas y desventajas por autor

Parámetro	Norma estándar	Ventajas	Desventajas
Viscosidad	ASTM D445	-Medición de líquidos transparentes y opacos -Alta precisión	-Baja repetibilidad -No garantiza seguridad del usuario
S.A.R. A	ASTM 2007	-Determina varios compuestos orgánicos de la muestra	-Alto costo tiempo -Alto costo monetario
ASCI	ASTM E1655	-Rápido y Eficaz -Evaluación estabilidad de asfaltenos	-Restricciones en condiciones de muestras aceptadas -Equipo especializado
%CCR	ASTM D189	-Medida cuantitativa -Indicativo estabilidad térmica	-Baja Precisión -Lento -Equipo especializado -No amigable ambiente

Ante estas limitaciones de las pruebas tradicionales y una base de datos establecida surge naturalmente la idea de optar por soluciones actuales con la ayuda de tecnologías como el machine learning, haciendo uso de sus técnicas para maximizar la precisión de los resultados y minimizar los impactos negativos de la prueba en el ambiente y en la eficiencia del proceso.

Dentro de la literatura ya se registran el usó de *machine learning* para la caracterización de propiedades del petróleo, donde se evidencia el uso satisfactorio de modelos de regresión y clasificación, por ejemplo, el uso de técnicas como Regresión de vectores de soporte (SVR), la regresión por mínimos cuadrados parciales (PLS-R) y bosques de decisión (RF) para la caracterización automatizada de mezclas de ceras de petróleo (Barea-Sepúlveda et al., 2024), además se han desarrollado modelos mediante regresión PLS con base en datos de espectroscopia infrarrojo para la determinación del índice ASCI (Niño et al., 2019).

3.3. Estado del arte en el uso de machine learning en la estimación de propiedades fisicoquímicas a partir de ensayos de espectrometría de masas y espectroscopia de infrarrojo medio

3.3.1. Machine Learning

Machine Learning (ML) o aprendizaje automático en español, se puede definir como el proceso de enseñar a una máquina a "pensar" como un ser humano para realizar una tarea determinada, sin estar explícitamente programada. Este concepto ha tomado cada vez más fuerza y relevancia en la sociedad actual, gracias a la gran cantidad de data disponible. El uso de esta información para fines predictivos puede ayudarse a tomar decisiones; tanto así que varias áreas de la investigación y el desarrollo implementan cada vez más este concepto en sus labores. Dependiendo de la manera en que los datos se encuentren organizados y la cantidad de información utilizada para entrenar estos sistemas, el *machine learning* se puede dividir en 4 grandes grupos,

aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo y aprendizaje semi supervisado (Farnham et al., 2019).

3.3.2. PCA

En cualquier grupo de aprendizaje en el que se decida trabajar es común encontrarse con problemas de dimensionalidad; estos refieren a una gran cantidad de datos que hacen difícil de generar un modelo correcto o ralentizar su procesamiento. Este tipo de problemas son en realidad algo común en la actualidad, como se evidenciará en el capítulo 5.1 al trabajar con el primer conjunto de datos, debido a la escalabilidad que tiene la información. Los modelos que se desarrollan a lo largo de este proyecto no son la excepción; Sin embargo, existen diversos métodos para abordar este tipo de problemas. Entre ellos destacan los algoritmos de reducción de dimensionalidad, como el PCA, que además garantiza la eliminación de correlaciones entre las variables, y el LDA (Farnham et al., 2019). Estos resultan realmente útiles en las tareas de clasificación, donde es posible escoger un grupo acotado de predicciones y desarrollar el algoritmo con ayuda de estos valores.

3.3.3. Aprendizaje supervisado

Si se conocen los datos objetivo del sistema es posible etiquetarlas en el entrenamiento del modelo, de esta forma el modelo de *machine learning* se convierte en aprendizaje supervisado.

Esta clase de modelos se vuelven realmente útiles en tareas de clasificación. Por ejemplo, para la creación de un filtro de un correo a spam, el cual identifica posibles correos no deseados y los clasifica como tal (Farnham et al., 2019). También, como en este estudio, se puede aprovechar la data obtenida con espectrometría de masas de muestras de hidrocarburos para etiquetarlos como crudo, gas o fondo de vacío.

Otra gran utilidad en etiquetar los datos de solución se encuentra en predecir un valor numérico. Esta labor que desempeñan estos predictores tiene el nombre de regresión. Un ejemplo de regresiones es encontrar el precio de un auto de acuerdo con sus características (Farnham et al., 2019) y para el particular de este trabajo de grado, se encontrarán predicciones de características fisicoquímicas del petróleo como el análisis de saturados, aromáticos, resinas y asfáltenos (S.A.R.A), el porcentaje de carbono Conradson (%CCR) y el índice de clases de solubilidad de asfáltenos (ASCI), a partir de pruebas de espectrometría de masas y espectros infrarrojos, conceptos explicados más adelante.

Dentro de los algoritmos más destacados del aprendizaje supervisado se encuentra las máquinas de soporte vectorial (SVM), los árboles y bosques de decisión (DT y RF), los clasificadores de vectores de soporte (SVC), los regresores de vectores de soporte (SVR), la regresión por mínimos cuadrados parciales (PLS-R) y en ciertos casos las redes neuronales.

3.3.4. Redes Neuronales

Las redes neuronales artificiales es una rama del aprendizaje por refuerzo es un concepto en la computación donde se intenta replicar el funcionamiento de una red neuronal biológica en una computadora. Es un concepto que nace desde 1939 con Alan Turing, estudiando el cerebro humano para fines de computación. Gracias al poder de cómputo y la gran escalabilidad de los datos de la era actual, cobran una mayor fuerza y relevancia

La red neuronal artificial se compone de las siguientes características; tienen unas neuronas de entrada que son las que tomaran los datos de entrada, posteriormente tiene unas neuronas de procesamiento las cuales se encargan de clasificar y manipular la información. Estas neuronas puedes llevar diferentes niveles, los cuales se conocen como capas. Finalmente, están las neuronas de salida, que proporcionan la información de salida según la tarea especifica. Adicional a las

neuronas, hay más conceptos a destacar de las redes neuronales como las funciones de activación, los perceptrones, el algoritmo de *backpropagation*, conceptos que al ser utilizados permiten desarrollar una red neuronal ideal para unos datos específicos (Farnham et al., n.d.).

Las redes neuronales, los algoritmos de clasificación y regresión ya se han implementado en la industria del petróleo para predecir propiedades importantes (Camargo Reina & Hernandez Oviedo, 2014) es por eso por lo que se busca aprovechar su utilidad con librerías que le saquen el máximo provecho como *keras* o *scikit-learn.*

Por ejemplo (Raljević et al., 2021) utilizaron los datos obtenidos tras desarrollar una espectroscopia de resonancia magnética nuclear (NMR) en muestras de crudo para prever su estabilidad. El enfoque se centró en propiedades como el contenido de asfaltenos y la gravedad API, y los resultados se derivaron de pruebas estándar de laboratorio. Utilizaron un modelo de regresión lineal que arrojó un coeficiente de determinación R^2 satisfactorio de hasta 0.9892. Sin embargo, es importante señalar que este modelo podría presentar limitaciones debido al número reducido de pruebas de laboratorio realizadas, lo que se traduce en una menor cantidad de propiedades evaluadas y, por ende, en una caracterización más limitada.

Además de la cantidad de asfaltenos y la gravedad api, se han realizado modelos para la caracterización de mezclas de ceras de petróleo importantes en el proceso de refinación del petróleo, (Barea-Sepúlveda et al., 2024) utilizaron 72 muestras de crudo, clasificadas según su concentración de mezcla en su relación microcera-parafina (12.5, 25, 50 y 75%) tras esto se realizó un modelo Vis-NIR para obtener un análisis espectroscópico de cada muestra, para posteriormente aplicar modelos como PLSR, SVR, RF, con sus respectivas métricas como la raíz de error cuadrado medio (RMSE) y el coeficiente de determinación R^2 para determinar la cantidad de concentración en cada uno de las muestras , como resultado se identificó que el modelo de PLS obtuvo mejores

resultados sobre sus pares y se incita a replicar sus resultados en otras propiedades importantes del petróleo.

El *machine learning* sirve como un método de caracterización que además de preciso ha demostrado ser un método más rápido y eficiente que la prueba estándar de laboratorio, un claro ejemplo es la caracterización de biodiesel como se muestra en el estudio de Chen et al. (2023). En este estudio se tomaron como punto de partida y datos de salida el contenido de grupos insaturados, el contenido de oxígeno (O), y los contenidos de cuatro ésteres representativos para 71 muestras. Estas muestras fueron comparadas con los datos de entrada obtenidos a partir de un análisis de espectroscopia infrarroja (FTIR) en cada muestra, con el objetivo de predecir las propiedades de salida mencionadas, utilizando métodos del *machine learning* como SVM, Redes neuronales (NN), y bosques aleatorios (RF). Los resultados destacaron que el método de clasificación PCA, juega un rol clave para reducir la dimensionalidad de los datos espectroscopios (intensidad - absorbancia) y por consiguiente mejorar la precisión de predicción, además dentro de los modelos probados las redes neuronales demostraron mejor ajuste en la clasificación de datos, mientras que, en la regresión de estos, el método de bosques aleatorios obtuvo una mejor respuesta

A continuación, se presenta una tabla sintetizando estos estudios y sus principales características.

Tabla 2.

Resumen de resultados del estado del arte: análisis de ventajas y desventajas por autor

AUTOR	**DESCRIPCIÓN**	**VENTAJAS/DESVENTAJAS**
Raljević et al., (2021)	Predicción de gravedad api, cantidad asfaltenos a partir de	Este autor uso solamente regresión logística y utilizo

	intensidad de onda usando NMR en muestras de crudos.	pocas propiedades de caracterización.
Barea-Sepúlveda et al., (2024)	Predicción de mezclas de ceras de petróleo a partir de intensidad de onda usando Vis-NIR en muestras de crudos.	El modelo PLS demostró ser altamente efectivo en propiedades del petróleo
Chen et al., (2023)	Predicción de contenido grupos insaturados, contenido Oxigeno, contenido ésteres, a partir de intensidad de onda usando FTIR en muestras de crudos.	-PCA demostró ser muy importante en reducir la dimensionalidad de pruebas FTIR -Las redes neuronales demostraron un gran ajuste en métodos de clasificación mientras que los bosques aleatorios mostraron

4. Metodología

4.1. Construcción de una base de datos

A partir de las pruebas espectroscópicas haciendo uso de técnicas de *big data* y análisis de datos para depurar y limpiar la información

4.1.1. Obtención de datos:

Los datos son suministrados por la Universidad y pueden provenir de pruebas de espectroscopía de masas o infrarroja realizadas a muestras de petróleo o crudo; *así como información de las propiedades fisicoquímicas (%CCR, S.A.R.A o ASCI).*

4.1.2. Modificación y manipulación de datos:

Utilizando librerías de *Python* como *Pandas* y *NumPy* se aplican las funciones y técnicas correspondientes para eliminar redundancia y datos innecesarios, así como aplicación de la normalización de datos.

4.2. Implementación de modelos

En esta etapa se realiza la ejecución de los modelos de machine learning que clasifiquen las muestras de petróleo entre crudo, fondo y gas, y predigan propiedades fisicoquímicas del petróleo como el porcentaje carbono Conradson (%CCR), índice S.A.R.A y ASCI

4.2.1. Creación de modelos de machine learning:

Con apoyo de librerías como *Scikit-learn* y *Keras* se prueban diferentes modelos, de clasificación como un clasificador de soporte vectorial (*SVC*) o análisis discriminante lineal (*LDA*); de regresión como un regresor de vectores de soporte (*SVR*) o el método de mínimos cuadrados parciales (*PLS*), y redes neuronales probando diferentes funciones de activación y optimización como la unidad literal rectificada (*RELU*), *softmax*, *sigmoide*, descenso del gradiente estocástico (*SGD*), estimación adaptativa del momento (*ADAM*) y propagación de la raíz media cuadrática (*RMSprop*) que minimice la función de costo.

4.2.2. Predicción de datos:

Utilizando el módulo *predict* de las librerías implementadas, se evaluarán los modelos a partir del rendimiento en las predicciones y se realizará una matriz de comparación para determinar el modelo que mejor resuelve la tarea de predicción y clasificación de muestras complejas.

4.3. Validación de los modelos basados en *machine learning*

Se realizó la validación de los modelos de machine learning utilizando métricas específicas para las tareas de regresión y clasificación. Este proceso se llevó a cabo con un conjunto de datos de laboratorio, incorporando soporte gráfico para visualizar los resultados obtenidos.

4.2.2. Medición y validación de predicciones:

A través de herramientas estadísticas y funciones de perdida como el error de la raíz media cuadrática (RMSE), el Error cuadrático medio (MSE) o el coeficiente de determinación R^2 de las librerías establecidas se evalúa numéricamente la calidad del modelo y se realizan ajustes de ser necesario.

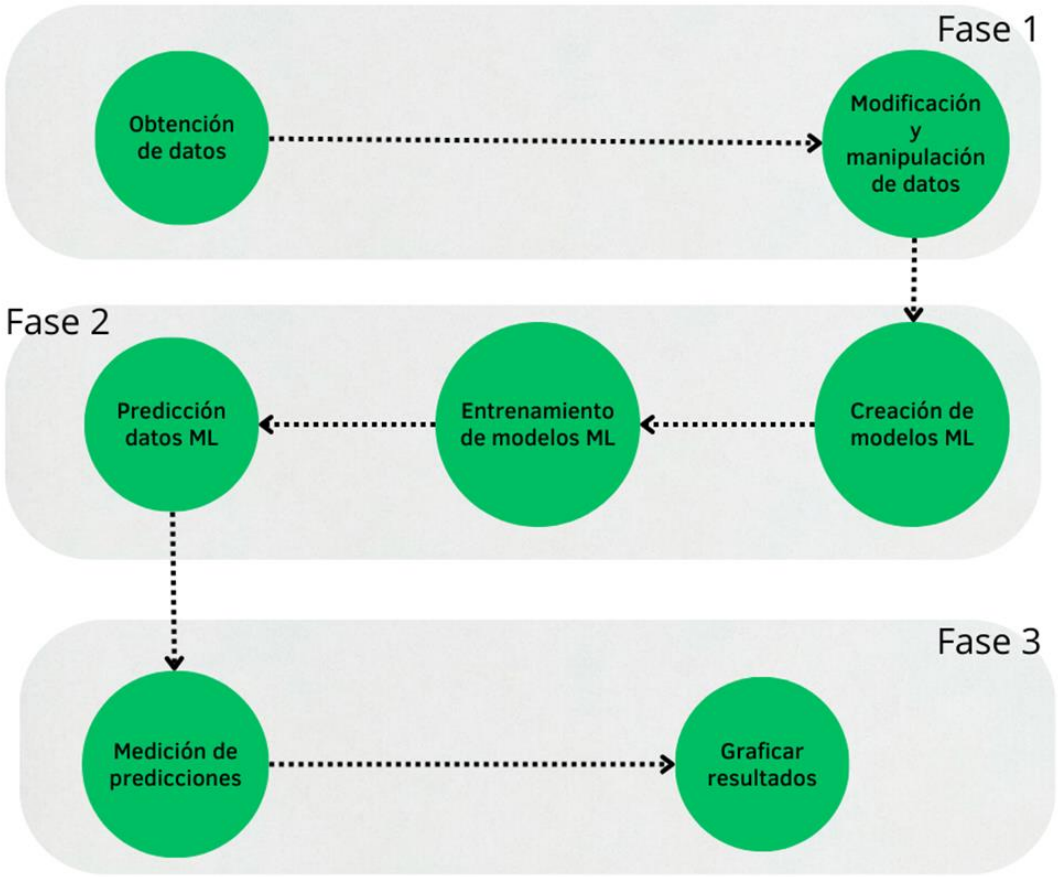
4.2.2. Creación y ejecución de una interfaz gráfica:

Se creará una interfaz gráfica con apoyo de la librería *Tkinter* de Python de libre uso y fácil acceso, dentro de su desarrollo tendrá precargado los modelos, gracias a la librería *Joblib*, con los mejores resultados obtenidos. Al usarlo permitirá escoger entre las propiedades fisicoquímicas y, tras subir la información de la muestra de crudo, mostrará en pantalla la predicción realizada.

A continuación, se presenta un apoyo grafico de la anterior metodología descrita.

Figura 5.

Diagrama de la metodología descrita



5. Resultados

5.1. Resultados de la construcción de una base de datos y análisis con técnicas de *big data*

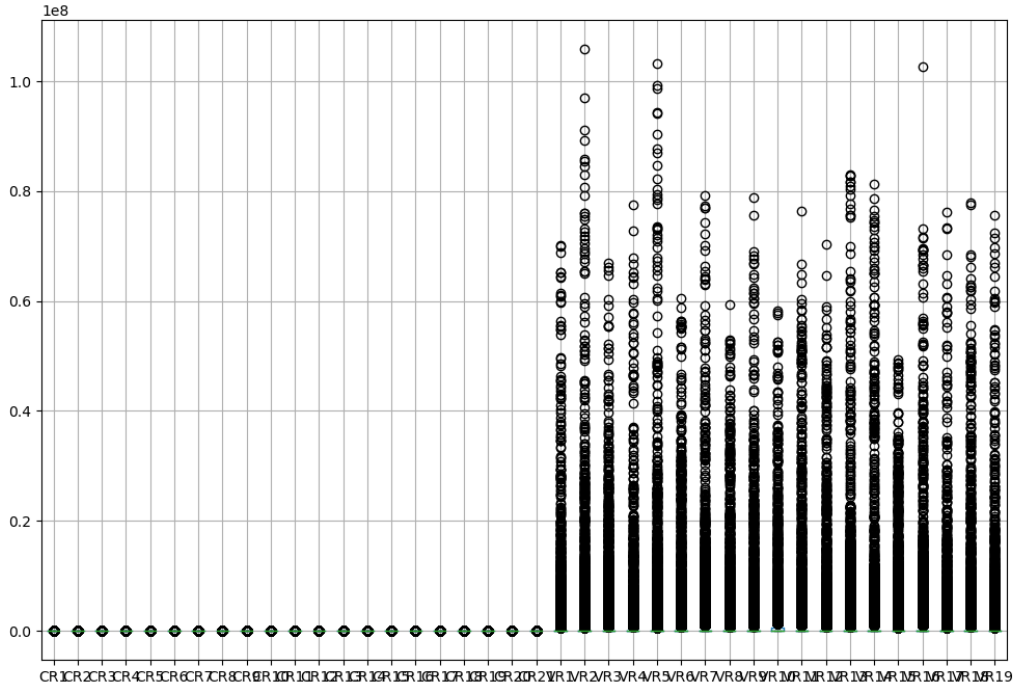
Los datos de espectroscopia de las muestras de petróleo fueron suministrados por la Universidad Industrial de Santander, resultado de pruebas de espectroscopía de masas de ultra alta resolución y espectroscopía infrarroja por transformada de Fourier en la región del infrarrojo medio (MIR-FTIR), en formato *csv*. Cada modelo que se realizó consta de diferentes técnicas de *big data* ajustadas para la obtención de modelos de clasificación y regresión acordes a los datos

ingresados. A continuación, se presentan la obtención y manipulación de datos para cada uno de los modelos de machine learning desarrollados.

El primer conjunto de datos obtenido es resultado de la realización de pruebas de espectroscopia. Dicho conjunto contiene 40 muestras de petróleo etiquetadas en crudos (CR) o fondos de vacío (VR). El espectro de masas resultante muestra picos que corresponden a los iones formados. Con esta información en consideración, inicialmente se realiza un diagrama de caja para identificar la distribución de los datos:

Figura 6.

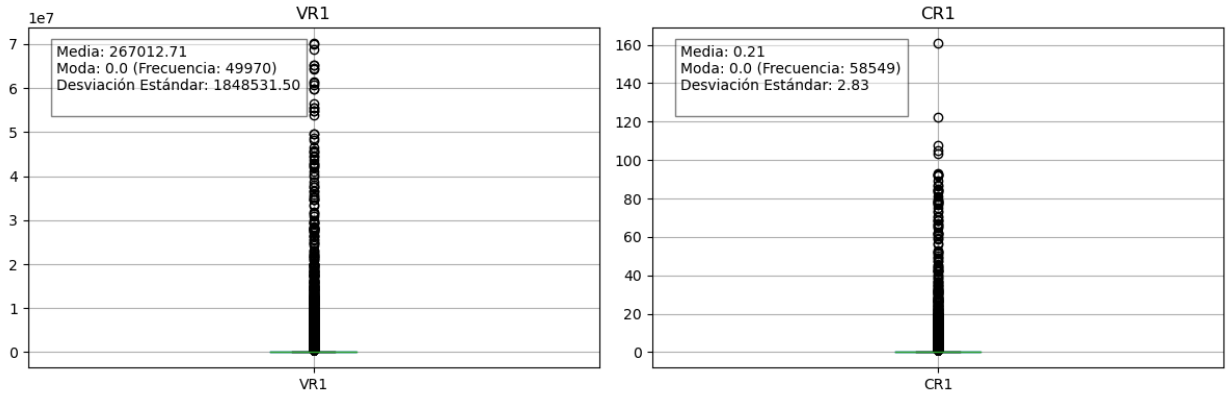
Diagrama de caja del primer conjunto de datos



Analizando la gráfica generada se hace evidente una distribución dispar de datos, entre las muestras clasificadas como crudos y aquellas clasificadas como fondo de vacío, por lo que la siguiente ruta a toma es separar los datos entre estas categorías.

Figura 7.

Diagrama de cajas de la primera muestra de fondos y primera muestra de crudos



Una vez realizado el diagrama de caja de cada uno de los grupos, se pudo identificar que, inicialmente, los fondos de vacío tienen una mayor abundancia o concentración de las moléculas o compuestos que se están midiendo en comparación con los crudos. Esto se hace evidente al observar que la media de los datos en las muestras de vacío es significativamente mayor que en las muestras de crudo. Tal diferencia facilitaría la creación de un modelo de clasificación que prediga si una muestra pertenece a un crudo o a un fondo, lo cual podría funcionar como una forma de validación de datos de espectrometría de masas.

Una segunda observación es la elevada cantidad de ceros en las muestras, lo que podría derivar de sustancias que realmente no están aportando información útil. Además, es necesario realizar una normalización de los datos antes de aplicar técnicas de análisis, como PCA. La normalización ajusta los valores de las características a una escala común, lo que ayuda a evitar que las características con rangos más amplios dominen el análisis. Esto es especialmente relevante en este caso, ya que las diferencias en la escala entre crudos y fondos podrían afectar la precisión del modelo.

El desafío, entonces, es que el número de moléculas es muy grande en relación con el número de muestras, lo que crea un problema de dimensionalidad alta. Esto puede llevar a dificultades para analizar todos los datos debido a la complejidad computacional y a la obtención de modelos robustos, si no es posible usar todas las variables. Por lo tanto, tomar un enfoque que incluya la normalización de nuestros datos de entrada es de gran importancia.

La normalización se puede expresar matemáticamente como:

$$X_{normalizado} = \frac{X}{||X||_p}$$

Donde $||X||_p$ es la norma p de los vectores de entrada, y el tipo de norma se puede especificar según el contexto (por ejemplo, norma L2 o L1). La normalización L2 utilizada en este conjunto de datos ajusta los vectores de características para que la suma de los cuadrados de sus elementos sea igual a 1. Este proceso implica dividir cada elemento del vector por la norma L2 del mismo, que se calcula como la raíz cuadrada de la suma de los cuadrados de los elementos:

$$||X||_2 = \sqrt{\sum_{i=1}^{n} {x_i}^2}$$

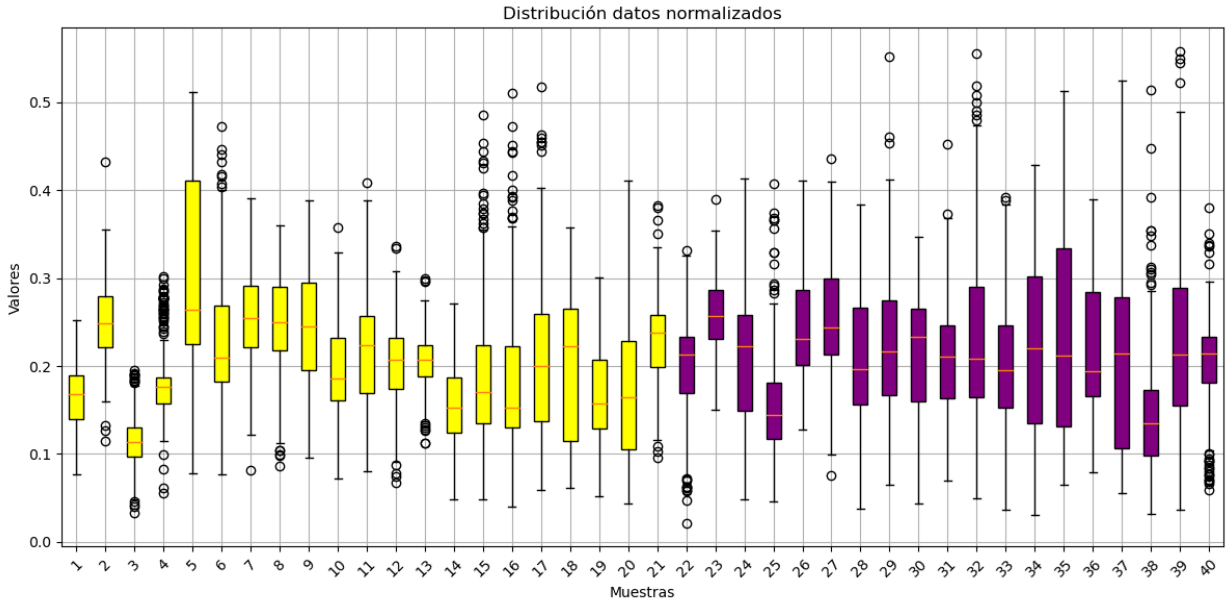
En ese mismo sentido, el uso de PCA (Análisis de Componentes Principales) también podría ser beneficioso. PCA es una técnica de reducción de dimensionalidad que transforma los datos a un nuevo sistema de coordenadas donde las primeras componentes principales capturan la mayor parte de la variabilidad de los datos. La transformación se puede expresar matemáticamente como:

$$Z = XW$$

donde Z son los datos transformados, X son los datos originales, y W es la matriz de componentes principales. Esto no solo ayuda a simplificar el modelo, sino que también puede mejorar su rendimiento al eliminar ruido y redundancia.

Figura 8.

Diagrama de caja del set de datos normalizado y separados por crudos y fondos de vacío



Sin embargo,hay que recordar que existen desventajas significativas al aplicar PCA. Por ejemplo, el problema "*curse of dimensionality*" puede hacer que PCA no funcione de manera eficiente. Cuando se tiene un número tan alto de características en comparación con las muestras, el análisis puede colapsar debido a la falta de suficientes muestras para representar bien la varianza en tantas dimensiones

Posteriormente tras eliminar los ceros que no aportan información, se logró reducir la cantidad de datos a analizar de 59.692 a 55.154, es decir una diferencia de 4538 filas innecesarias, que al normalizar generan mayor precisión y certeza en los modelos al momento de crearlos. Estas

4538 filas pueden ser también fácilmente extraíbles y marcadas con la molécula correspondiente, para encontrar relación entré estos datos eliminados.

Una vez analizado el conjunto de datos de muestras se pueden observar los datos de los valores a predecir con los modelos de *machine learning.* Para ello con ayuda de la librería pandas se hace una revisión del archivo Supplementary_CCR_properties.csv y Crudos_Fondos_Propiedades_NombresyEtiquetas.csv, los cuales contienen la información de las propiedades fisicoquímicas de interés (%CCR y análisis S.A.R.A) para cada una de las muestras recopiladas.

Tabla 3.

Visualización de una porción de los datos %CCR

Muestra	**CCR (%)**
CR1	12,94
CR2	4,34
CR3	9,67
CR4	7,03
CR5	17,84
VR1	35,5
VR2	13,5
VR3	21,6
VR4	24,8
VR5	19,8

De la información del CCR, podemos calcular la media de los valores, su valor mayor y su valor menor, para de esta manera saber qué valores esperar al momento de realizar modelos de regresión. Del mismo modo es posible dividir estas operaciones en 3 grupos: todo el conjunto de los datos del %CCR, datos etiquetados como crudos (CR) y datos etiquetados como fondo de vacío (VR). A continuación, se presentan los resultados de estos 3 grupos en una tabla.

Tabla 4.

Análisis %CCR

	Total	**Grupo CR**	**Grupo VR**
Media	16.7240	9.1028	25.14736
Varianza	96.6342	18.8949	45.7226
Moda	3.38	3.38	13.5
Máximo	37.6	17.84	37.6
Mínimo	3.38	3.38	13.5
Percentil 25 (Q1)	8.2775	5.34	20.5
Percentil 75 (Q3)	22.4	11.8	30.35

Observando la información obtenida es posible concluir que para los datos suministrados es factible esperar valores de entre 3.38 y 37.6 de %CRR, además los valores del grupo correspondiente a fondos de vacío tienen por naturaleza valores mayores que el grupo de los crudos. Asimismo, ante una alta variabilidad en los datos respecto a la media es posible optar por modelos como el *SVR* y *PLS* caracterizados por su eficacia ante valores atípicos (*outliers*) en un conjunto de datos.

Por su parte, el conjunto de datos del cual es posible recopilar el análisis S.A.R.A, cuenta con una cantidad considerable de valores nulos, no etiquetados o no asociados a las muestras analizadas. Es por eso por lo que con ayuda de la librería pandas se eliminan estos datos y se organizan de una manera que facilite su visualización como se presenta a continuación.

Tabla 5.

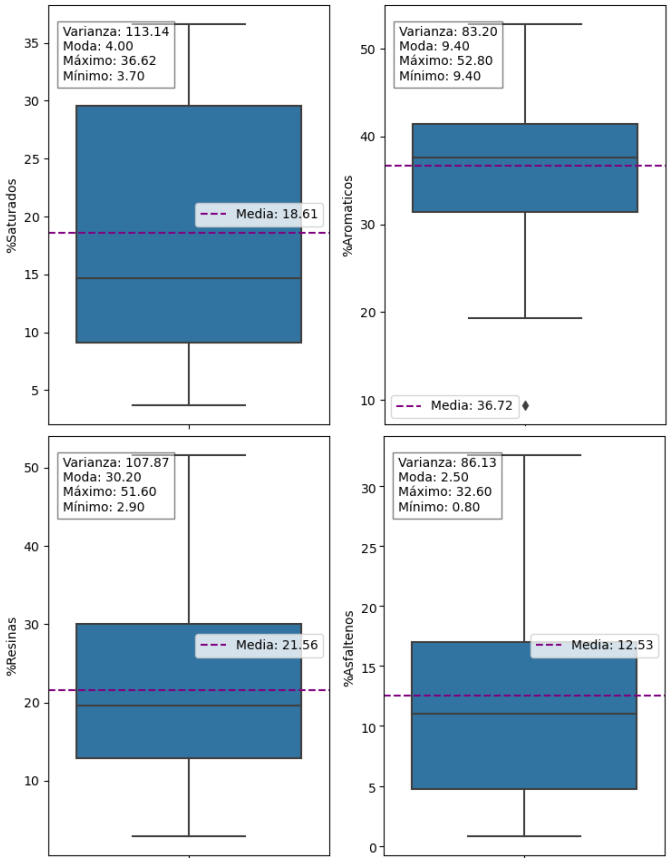
Visualización de una porción de los datos S.A.R.A

	%Saturados	%Aromáticos	%Resinas	%Asfáltenos
S1	33.88	34.12	8.78	11.59
S2	31.10	35.7	13.4	0.8
S4	25.0	39.7	21.5	2.5
S5	12.9	40.9	19.4	17.3
FV1	11.72	41.58	18.68	28.02
FV2	14.5	49.1	30.7	5.8
FV4	5.3	36.2	51.6	6.8
FV5	4.0	33.3	33.2	29.5

Para poder desarrollar un análisis efectivo se separó cada una de las clases de solubilidad y se trataron por separado, para ello se llevó a cabo un diagrama de cajas con ayuda de la librería *matplotlib* de *Python* que permita ver la distribución de los datos, a su vez se halló información importante que pueda aportar información a la comprensión intrínseca de los datos, como la media, la varianza, la moda y los valores máximos y mínimos sintetizada en el siguiente gráfico.

Figura 9.

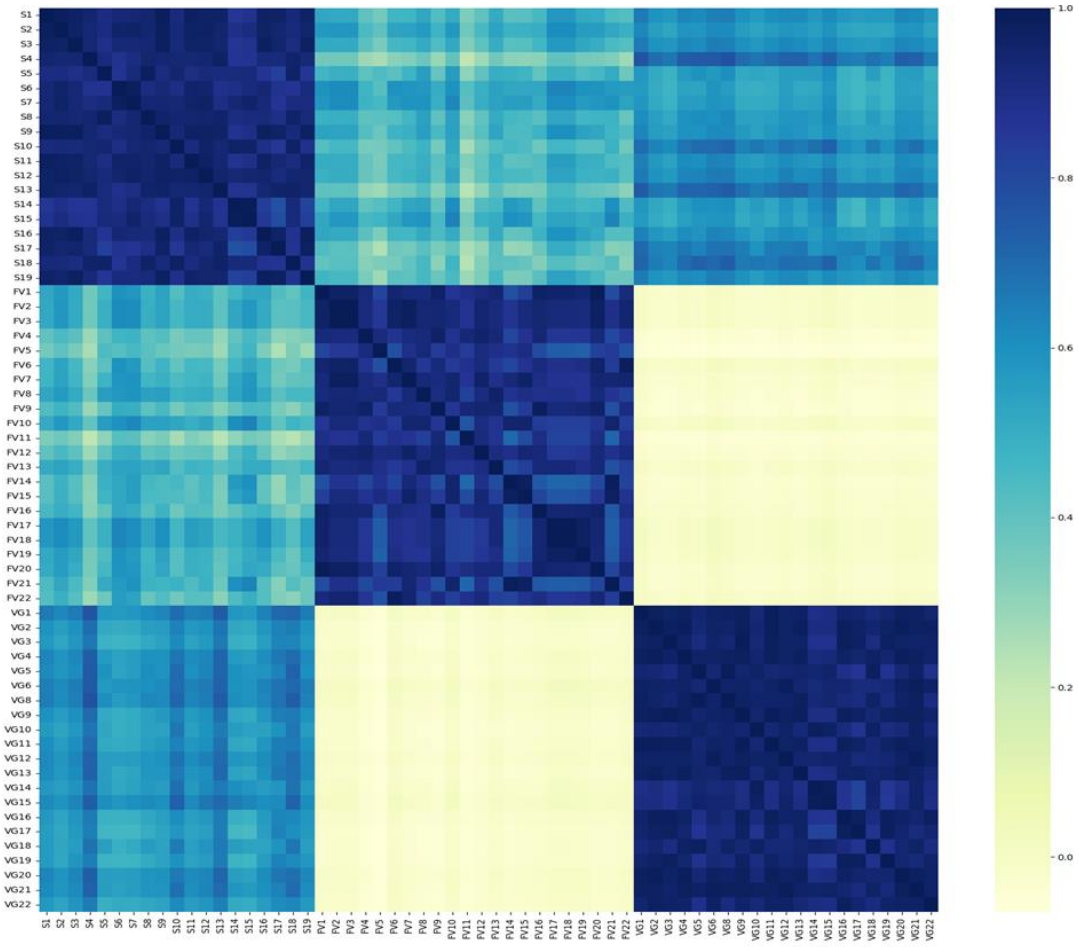
Diagrama de cajas S.A.R.A



El segundo conjunto de datos lleva por nombre ListaComp_appi_CRFVVG.csv, y consta de los resultados de realizar una espectroscopia de masas a diferentes muestras del petróleo. Esta vez se analizaron 62 muestras entre, fondo de vacío, crudo y gas. A continuación, se presenta el mapa de calor de la matriz de correlación.

Figura 10.

Mapa de calor, matriz correlación del segundo conjunto de datos



El uso del mapa de calor permite visualizar claramente cómo los datos etiquetados se separan entre sí al calcular sus correlaciones. En este caso, se observa un bajo índice de correlación entre los datos etiquetados como fondo de vacío y aquellos etiquetados como gas (FV-VG). La relación entre fondo de vacío y crudo (FV-S) es algo mayor, mientras que la relación más estrecha se encuentra entre los datos de crudo y gas (S-VG). Al identificar estas diferencias, podemos optar por un modelo de clasificación que distinga entre crudo, fondo de vacío y gas, y que funcione como validación frente a datos de espectrometría de masas. Esto nos da confianza en que el modelo

podrá identificar correctamente a cuál de los tres grupos pertenecen los datos según las pruebas de espectroscopía de masas.

El tercer conjunto de datos son los datos ASCI_IR, los cuales contienen la información de 82 pruebas de espectroscopia de infrarrojos a muestras del petróleo relacionadas cada uno con su índice ASCI. Sabiendo esta información y conociendo que el valor del índice ASCI es un numero continuo entre 0 y 20, surge la idea de realizar un modelo de clasificación, que a partir de los datos recopilados por la prueba de espectroscopia de infrarrojos logre predecir 1 entre los 21 valores posibles. Para ello se presenta una visualización de la distribución de los datos infrarrojos de las 82 muestras y una tabla sintetizando su contenido.

Figura 11.

Diagrama de caja de las 82 muestras (columnas)

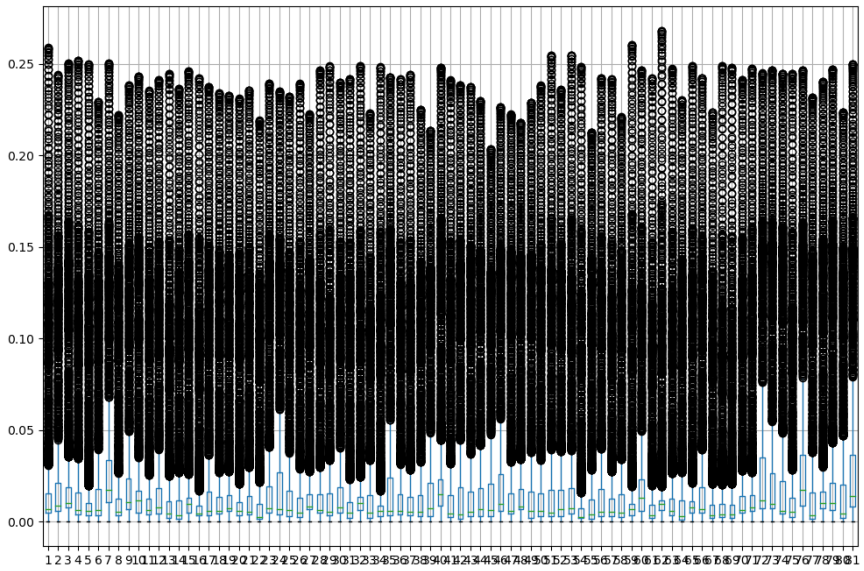


Tabla 6.

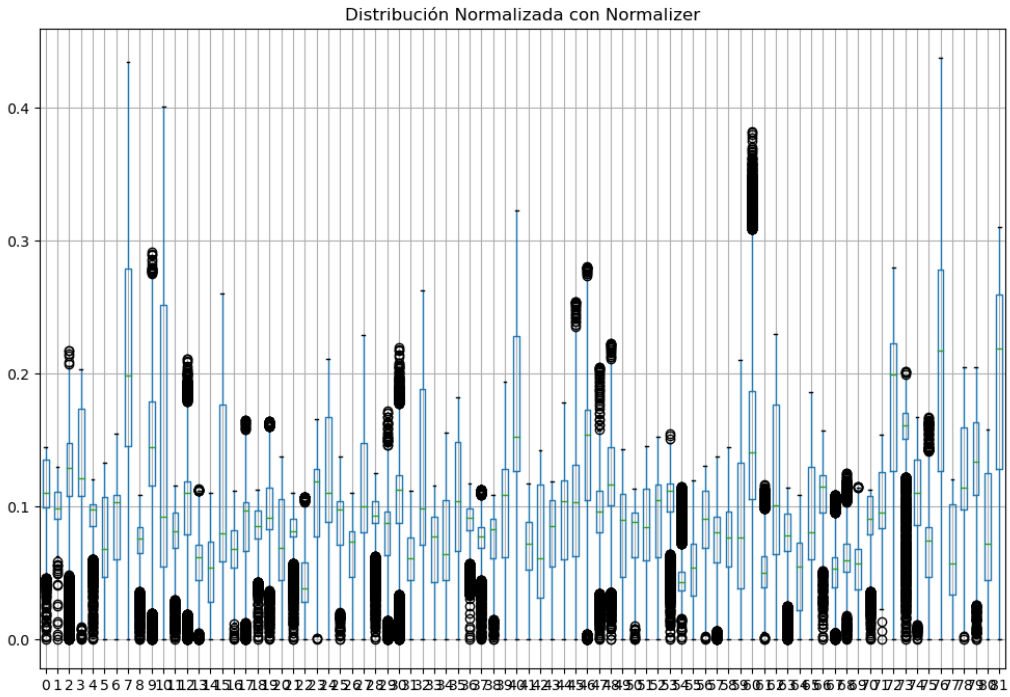
Datos importantes de algunas columnas

	Máximo	Mínimo	Promedio	Varianza
Columna1	0.237002	0.0	0.01651	0.0007886667
Columna2	0.24389401	0.0	0.01816	0.00082
Columna39	0.21353	0.0	0.01690	0.00078
Columna80	0.22372	0.00	0.01694	0.00088
Columna81	0.24980	0.0	0.02687	0.00109
Total	0.26799	0.0	0.01576	0.00081

Se observan ciertas características comunes y variaciones específicas que proporcionan una visión general de la distribución de los datos. Por ejemplo, se aprecia que la salida producida por una prueba de espectroscopía de infrarrojos contiene valores comúnmente bajos y con una varianza baja entre ellos, lo que puede dificultar el proceso de clasificación. Además, al manejar valores de los datos en el rango de las milésimas, sería útil realizar un proceso de normalización para escalar adecuadamente los datos y facilitar su interpretación. Por ello, se aplicó una normalización con norma L2 a través de *normalizer* de *sklearn*, que ajusta los valores para que cada muestra tenga una longitud unitaria.

Figura 12

Diagrama de caja del set de datos normalizado



A continuación, se realizó un análisis de varianza acumulada con un umbral del 80%, lo cual resultó en la selección de 2 componentes principales para llevar a cabo un Análisis de Componentes Principales (PCA). Esta técnica se utilizó para reducir la dimensionalidad de los datos a 2 componentes principales. La reducción dimensional obtenida permite visualizar de manera más clara las relaciones entre las muestras y explorar posibles agrupamientos o patrones dentro de los datos, facilitando así la interpretación y la posterior construcción del modelo de clasificación.

Figura 13

Diagrama de caja de los componentes principales (PCA)

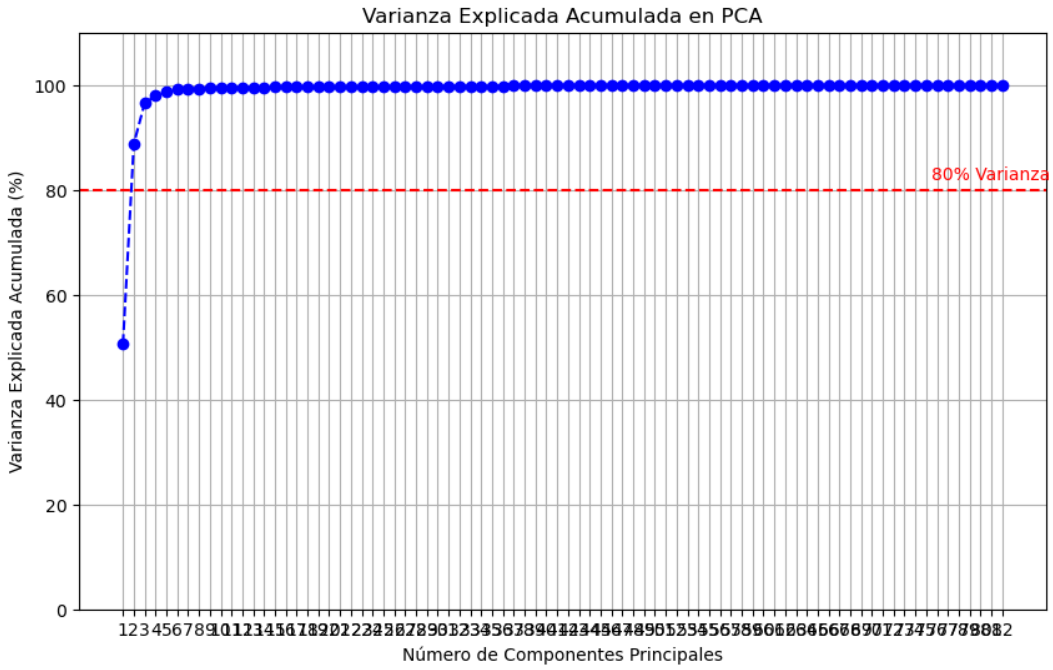


Figura 14

Resumen de datos analizados

Datos de Entrada	Análisis Realizado
Espectroscopia de masas 59692 características 40 muestras	Clasificador Crudo-Fondo (40 muestras) Regresión SARA (36 muestras) Regresión CCR (40 muestras)
Espectroscopia de masas 11917 características 60 muestras	Clasificador Crudo-Fondo-Gas (60 muestras)
Espectroscopia de infrarrojos 7468 características 82 muestras	Clasificador ASCI (0-20) (82 muestras)

5.2. Resultados de la implementación de modelos de machine learning

En total, se desarrollaron siete modelos de machine learning para clasificar y predecir propiedades fisicoquímicas de muestras de petróleo crudo. Estos modelos emplean técnicas como LDA, SVR, PLS y la generación de redes neuronales, aplicadas a datos obtenidos a partir de espectroscopía de masas e infrarrojos.

- **Primer modelo, Clasificador Crudos – Fondos:**

A partir de los datos obtenidos del análisis de espectroscopía de masas y el algoritmo *PCA*, estableciendo el número de componentes en dos, se implementó un modelo eficiente de clasificación entre crudos y fondos de vacío. Los datos obtenidos a través del *PCA* fueron utilizados como entrada, asignando valores numéricos a las etiquetas de crudo (1) y fondo de vacío (0) como datos de salida. El modelo se entrenó utilizando un 70% de los datos disponibles, empleando el método *train_test_split* de *sklearn*. Para la creación del modelo de clasificación, se

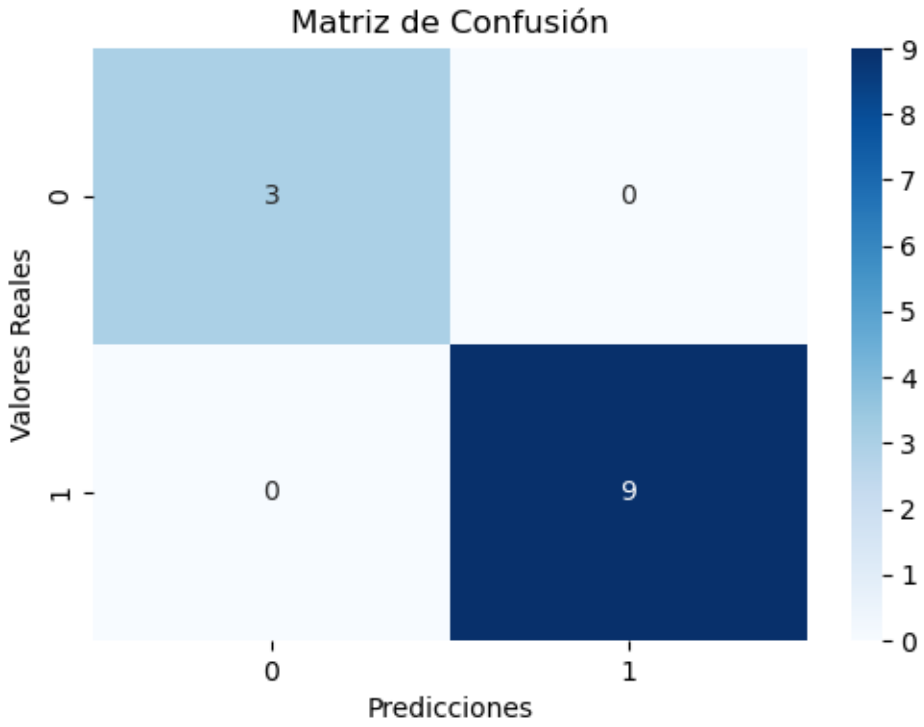
compararon dos algoritmos: LDA y SVC, este último con un parámetro regulador C=60 y un *kernel* lineal (Figura 12).

La utilización de LDA es especialmente valiosa por su capacidad para maximizar la separación entre las clases, optimizando la proyección de los datos en un espacio de menor dimensión, lo que resulta en una representación más clara y manejable de las muestras. Por otro lado, SVC, configurado con un valor alto de C, favorece la creación de un clasificador más preciso al reducir la tolerancia a errores de clasificación en el conjunto de entrenamiento. Esta combinación no solo mejora la precisión del modelo, sino que también proporciona interpretaciones claras de las relaciones entre las variables espectroscópicas, facilitando la identificación de patrones significativos en los datos analizados

Una vez ajustados los modelos a los datos de entrenamiento mediante el método *fit* de *sklearn*, se procedió a realizar predicciones sobre el conjunto de prueba (30% de los datos totales) utilizando el método *predict* de *sklearn* cbteniendo como resultado, se obtuvo un arreglo con las predicciones para los datos de prueba, clasificando entre crudo (1) y fondo de vacío (0).

Figura 15.

Matriz de confusión para las predicciones del modelo en el conjunto de prueba



- **Segundo y tercer modelo, Regresores %CCR:**

ara la generación de un modelo regresor de machine learning, se optó por utilizar un modelo de Máquinas de Vectores de Soporte para Regresión (SVR) con parámetros C=50, kernel polinómico y grado 2, así como un modelo de Regresión por Mínimos Cuadrados Parciales (PLS) con n_components=2. La elección de SVR se debe a su capacidad para manejar relaciones no lineales complejas entre las variables, lo que es especialmente relevante en el contexto de las características espectroscópicas. El uso de un *kernel* polinómico permite capturar la naturaleza no lineal de los datos, mientras que un valor de C moderado ayuda a equilibrar la complejidad del modelo y el ajuste a los datos. Por otro lado, el modelo PLS es eficaz para reducir la

dimensionalidad y extraer las características más relevantes que explican la variabilidad en los datos, lo cual es crucial cuando se trabaja con conjuntos de datos de alta dimensionalidad. Estos enfoques combinados permiten obtener una comprensión más precisa de la relación entre las características espectroscópicas y el índice objetivo.

Luego, se procedió a separar los datos en conjuntos de entrenamiento y prueba. En este caso, para las 40 muestras disponibles, se optó por utilizar un 80% de los datos para el entrenamiento del modelo. Se consideraron dos enfoques para evaluar el rendimiento de los modelos:

1. **Enfoque tradicional:** En este enfoque, se realizó una partición simple de los datos en conjuntos de entrenamiento y prueba, como se ilustra en la Figura 13. Este método sigue las prácticas estándar de machine learning.
2. **Enfoque iterativo**: En este enfoque, se implementó un ciclo *for* que iteró más de mil veces, particionando los datos en cada iteración y midiendo el rendimiento del modelo en cada ciclo. El objetivo fue identificar la partición que ofrecía el mejor rendimiento general del modelo, como se muestra en la Figura 14.

Figura 16.

Predicciones modelos frente a valores reales de prueba, enfoque tradicional %CCR

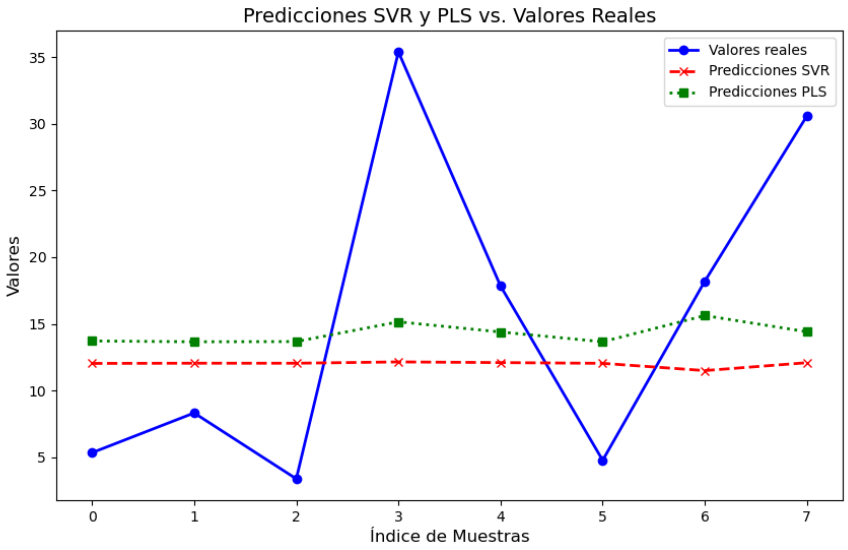
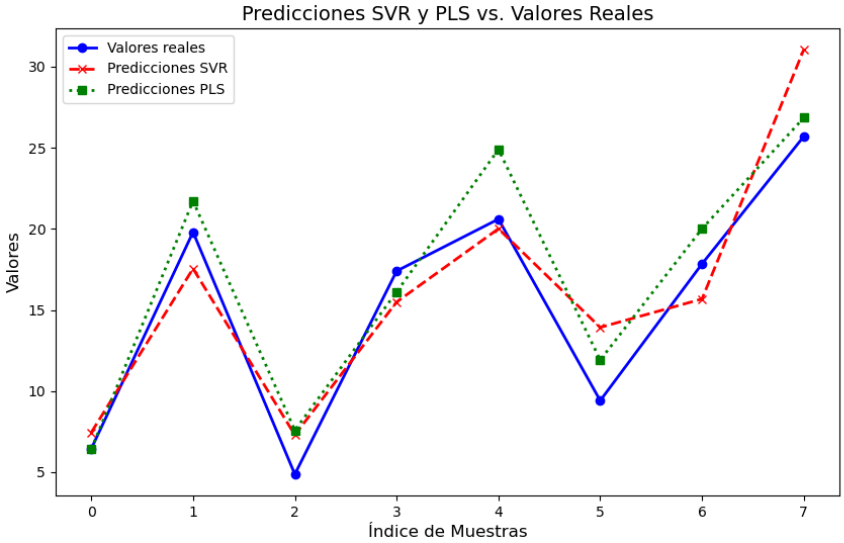


Figura 17.

Predicciones modelos frente a valores reales de prueba, enfoque iterativo %CCR



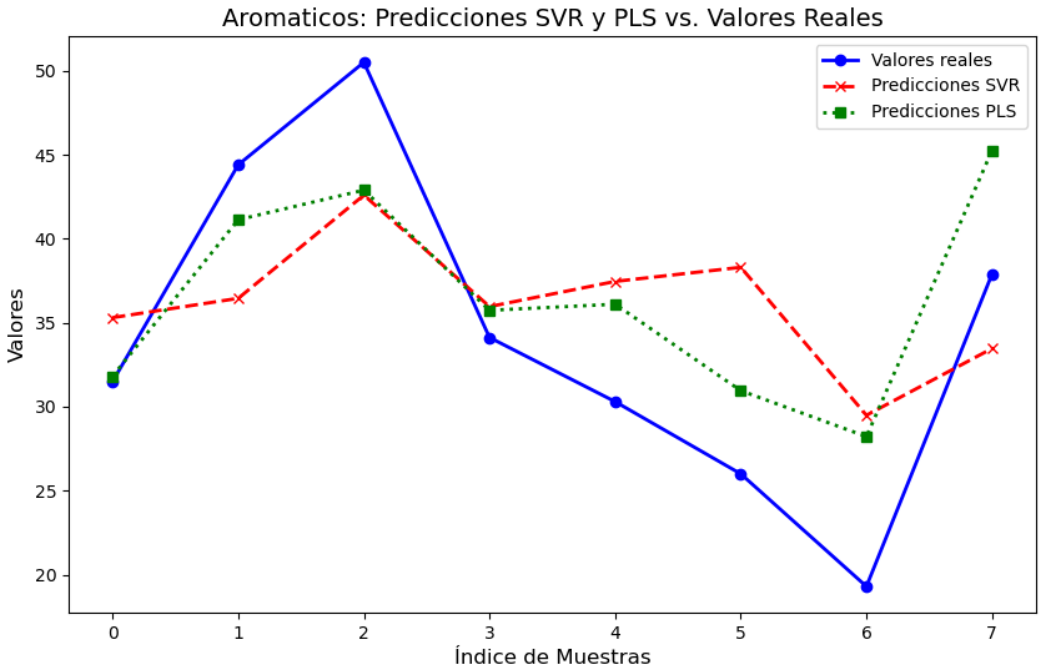
- **Cuarto modelo, Regresor S.A.R.A:**

Para la predicción del análisis SARA (Saturados, Aromáticos, Resinas y Asfaltenos), se implementó un modelo de regresión utilizando los algoritmos SVR y PLS. Siguiendo un enfoque iterativo similar al aplicado en modelos anteriores, se empleó un ciclo *for* para realizar múltiples particiones de los datos y seleccionar la partición que ofreciera el mejor rendimiento. Los datos se dividieron en un 80% para entrenamiento y un 20% para prueba.

Este proceso se repitió para cada una de las fracciones del análisis SARA, permitiendo ajustar el modelo de manera específica a las propiedades fisicoquímicas de cada fracción. El objetivo principal de este enfoque iterativo fue optimizar la precisión de las predicciones, asegurando que el modelo pudiera capturar adecuadamente las variaciones en las concentraciones de saturados, aromáticos, resinas y asfaltenos en las muestras de crudo.

Figura 18.

Predicciones modelos frente a valores reales de prueba de la fracción de Aromáticos en el Análisis SARA



- **Quinto modelo, Clasificador Crudos - Fondos – Gases:**

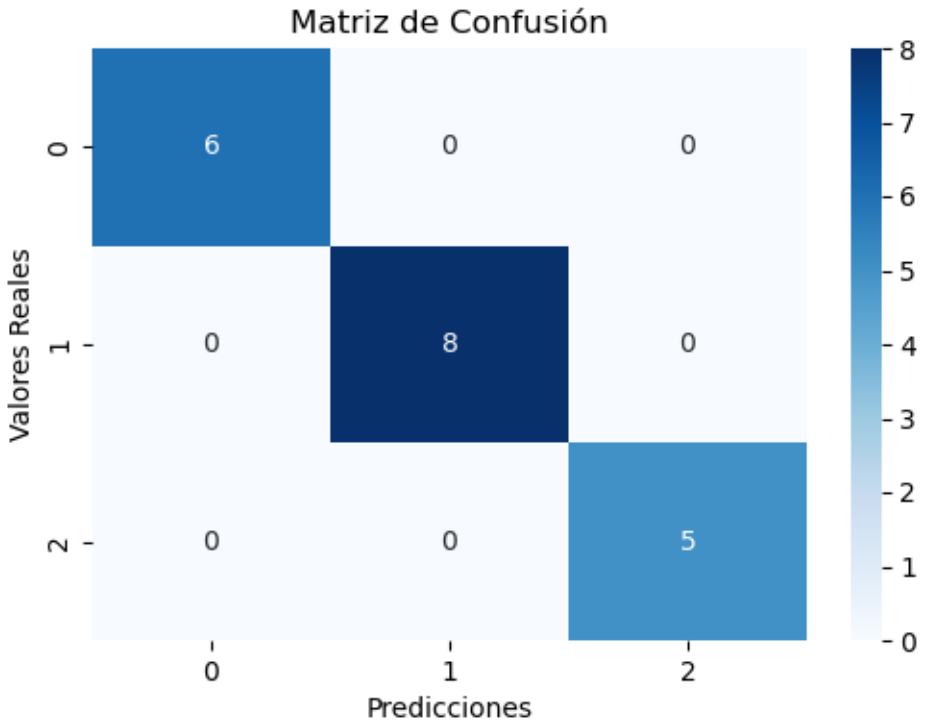
El quinto modelo desarrollado es un clasificador diseñado para distinguir entre tres categorías: crudo, fondo y gas, asignando a cada una de ellas los valores 2, 1 y 0, respectivamente. Para la creación de este modelo, se emplearon los algoritmos de *Support Vector Classification* (*SVC*) y *Linear Discriminant Analysis* (*LDA*), con el objetivo de identificar cuál de ellos ofrecía un mejor ajuste a los datos.

El proceso de desarrollo del modelo comenzó con el ajuste de los datos de entrenamiento mediante el método *fit* de la librería *sklearn*, permitiendo que el modelo aprenda las características

distintivas de cada clase (crudo, fondo y gas). Posteriormente, se utilizó el método *predict* para aplicar el modelo entrenado a los datos de prueba, generando así las predicciones correspondientes.

Figura 19.

Matriz de confusión para las predicciones del modelo en el conjunto de prueba

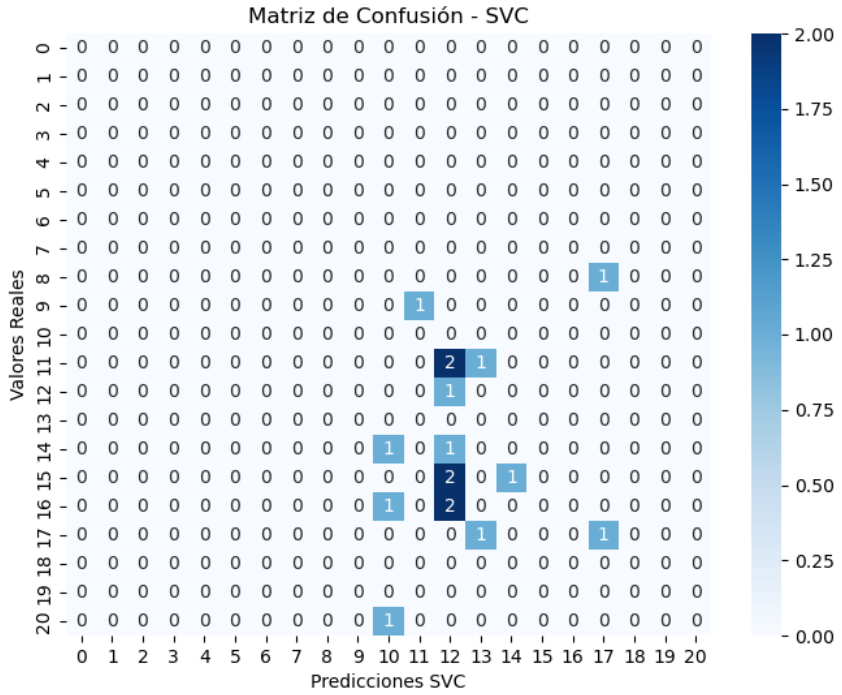


- **Sexto y séptimo modelo: Clasificador y Red Neuronal índice ASCI:**

El sexto modelo se centra en la predicción del índice ASCI utilizando un enfoque de clasificación. Para ello, se realizaron pruebas con diferentes clasificadores, entre ellos los modelos *SVC* y *LDA*, los cuales demostraron ser útiles en la clasificación de datos de espectroscopia de masas. Además, se utilizaron algoritmos ampliamente reconocidos en clasificación, como *Random Forest*, que es adaptable a datos con altos niveles de ruido y menos susceptible al sobreajuste (*overfitting*), como se observó en nuestro análisis de datos. También se empleó *Gradient Boosting*, que es efectivo en situaciones con relaciones no lineales, mostrando una alta precisión al manejar

valores continuos de entrada (entre 0 y 1) para predecir el valor entero del índice ASCI. Para llevar a cabo estas pruebas, se utilizó la librería *Scikit-learn* con los algoritmos *SVC*, *LDA*, *RandomForestClassifier* y *GradientBoostingClassifier*, implementando los métodos *fit* y *predict* para evaluar el rendimiento, como se muestra en la siguiente image. (ver Figura 18).

Figura 20. *Matriz de confusión para las predicciones del modelo en el conjunto de prueba*



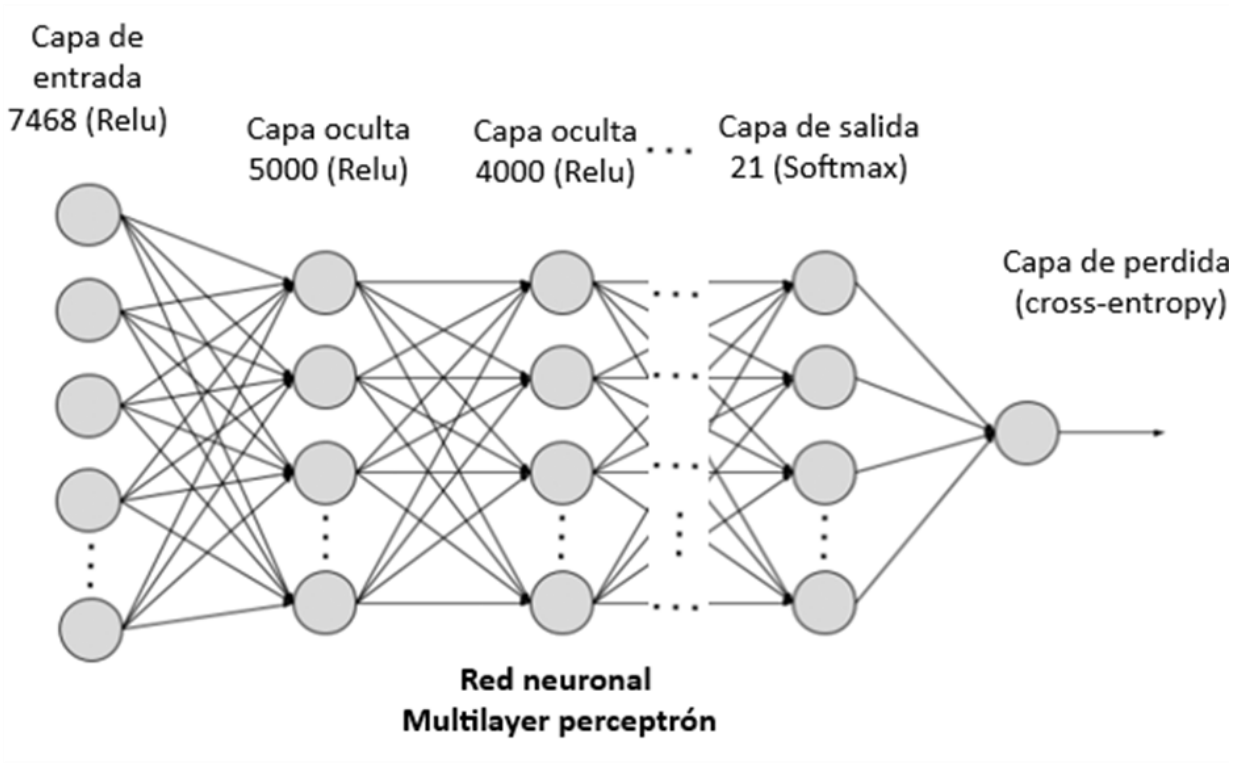
Por otro lado, el séptimo modelo es una red neuronal profunda diseñada específicamente para predecir el índice ASCI a partir de los datos de espectroscopia infrarroja. La red está compuesta por 15 capas de neuronas, cada una con la función de activación *ReLU*, seguida de una capa de salida con 21 neuronas referente a las 21 salidas posibles del índice ASCI con función de activación *softmax* para su clasificación.

La red neuronal fue entrenada utilizando el optimizador Adam, y se empleó la función de pérdida de entropía cruzada categórica (*categorical_crossentropy*) para ajustar los pesos del

modelo. Además, se utilizó la exactitud (*accuracy*) como métrica de evaluación. El modelo se + entrenó durante 30 épocas, lo que permite un entrenamiento más eficiente (Ver Figura 19).

Figura 21.

Estructura de la red neuronal profunda en la predicción del Índice ASCI



La siguiente tabla presenta un resumen detallado de los modelos desarrollados a lo largo de la investigación. En ella se especifican los métodos utilizados, el tipo de pruebas realizadas, la cantidad de datos empleados tanto para el entrenamiento como para los objetivos, así como las salidas obtenidas a partir de cada modelo. Este resumen ofrece una visión clara y estructurada de los enfoques y técnicas implementadas, facilitando la comprensión del proceso y los resultados obtenidos.

Tabla 7.

Resumen de Modelos de Machine Learning

Modelo	Método(s) Utilizado	Datos entrenamiento	Salida Modelo
Clasificador Crudos - Fondos	SVC - LDA	Espectrometría de masas 40 muestras	Crudo (0) - Fondo (1)
Regresor %CCR	SVR - PLS	Espectrometría de masas 40 muestras	Predicción del %CCR
Regresor %CCR	SVR - PLS Con iteraciones	Espectrometría de masas 40 muestras	Predicción del %CCR
Regresor SARA	SVR - PLS Con iteraciones	Espectrometría de masas 40 muestras	Predicción del fraccionamiento S.A.R.A
Clasificador Crudos - Fondos - Gases	PCA - LDA	Espectrometría de masas 62 muestras	Crudo (2) - Fondo (1) - Gas (0)
Regresor ASCI	SVC – LDA- RF- GB	Infrarrojos 82 muestras	Predicción índice ASCI (0-20)
Red Neuronal ASCI	Red Neuronal	Infrarrojos 82 muestras	Predicción índice ASCI (0-20)

5.3. Resultados de la validación de modelos basados en machine learning

- **Primer modelo: Clasificador Crudos - Fondos**

Para medir el rendimiento de los modelos clasificadores *Support Vector Classifier* (*SVC*) y el *Linear Discriminant Analysis* (*LDA*) se utilizó el conjunto de datos de prueba y se midieron sus desempeños mediante tres métricas: La exhaustividad o sensibilidad (recall), él área bajo la curva (AUC-ROC) y el puntaje de exactitud (*accuracy score*).

- **Sensibilidad (recall):**

El Recall es una métrica que mide la capacidad de un modelo para identificar correctamente los ejemplos positivos de cada clase. En este caso, tanto el modelo SVC

como el modelo LDA lograron un valor de 1.00 (o 100%), lo que indica que los modelos fueron capaces de capturar todos los ejemplos positivos del conjunto de prueba sin cometer errores. Un recall perfecto significa que no hubo falsos negativos, es decir, todas las instancias que pertenecen a una clase específica fueron correctamente clasificadas por los modelos.

- **AUC-ROC:**

El AUC-ROC (Área bajo la curva ROC) es una métrica que evalúa la capacidad de un modelo para distinguir entre diferentes clases. Un valor de 1.00 para ambos modelos significa que tanto el SVC como el LDA lograron una clasificación perfecta, asignando las probabilidades correctas a todas las clases sin confundir una con otra. Esto sugiere que los modelos no solo fueron precisos en las predicciones, sino que también tuvieron una excelente discriminación entre las clases, logrando una separación perfecta entre los ejemplos positivos y negativos.

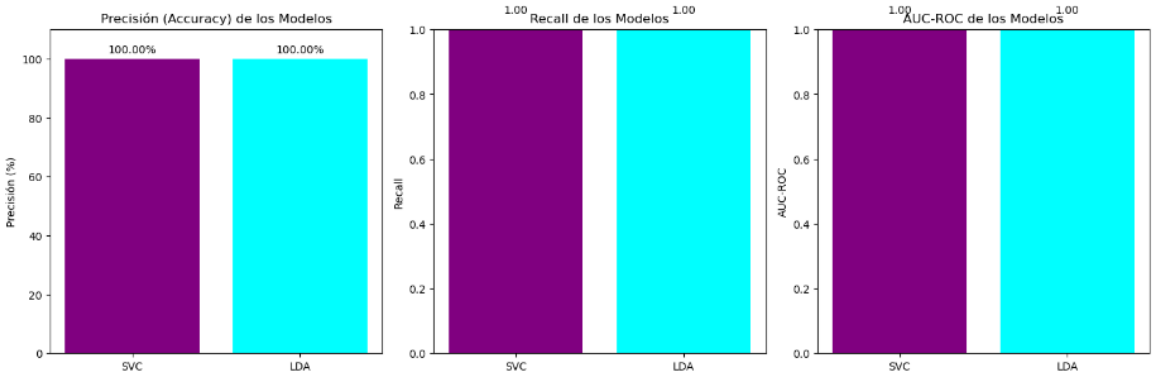
- **Puntaje de exactitud (*Accuracy score*):**

La exactitud, o *accuracy*, es una métrica muy común en modelos de clasificación. Representa la proporción de predicciones correctas sobre el total de predicciones realizadas. En nuestro desarrollo, los resultados de exactitud fueron del 100% lo cual indica que todas las instancias del conjunto de datos de prueba fueron correctamente clasificadas. Este nivel de exactitud implica que los modelos no cometieron errores en la clasificación y sugiere un rendimiento ideal bajo las condiciones de prueba.

A continuación, se presenta el resultado del *recall*, *AUC-ROC* y el *accuracy score* en los datos de prueba (y_te1) en los modelos *SVC* (predicciones1) y *LDA* (prediccionesLDA) con ayuda de la librería *matplotlib* de Python.

Figura 22

Resultados Accuracy Score, Recall y AUC-ROC primer modelo clasificación

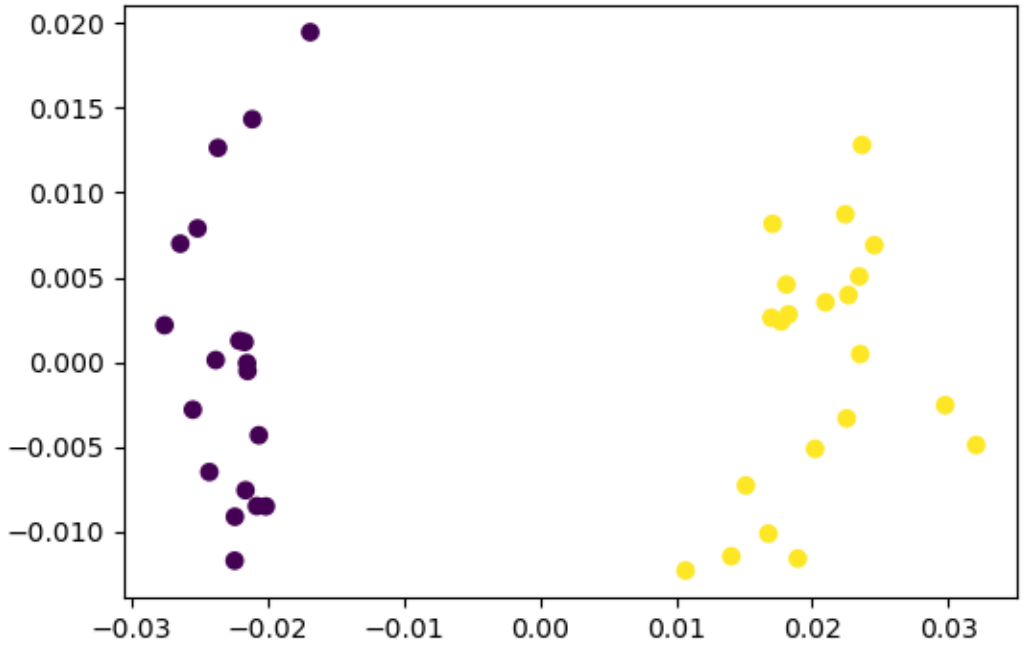


Estas mediciones dejan en evidencia el correcto acoplamiento del modelo a los datos brindados inicialmente, teniendo un acierto el 100% de las veces. Del mismo modo podemos graficar está información con ayudas de las librerías *seaborn* y *matplotlib*.

Por ejemplo, la siguiente grafica muestra la diferencia de datos y valores de entrada identificados como crudos (morado) y los valores identificados como fondos (amarillo).

Figura 23

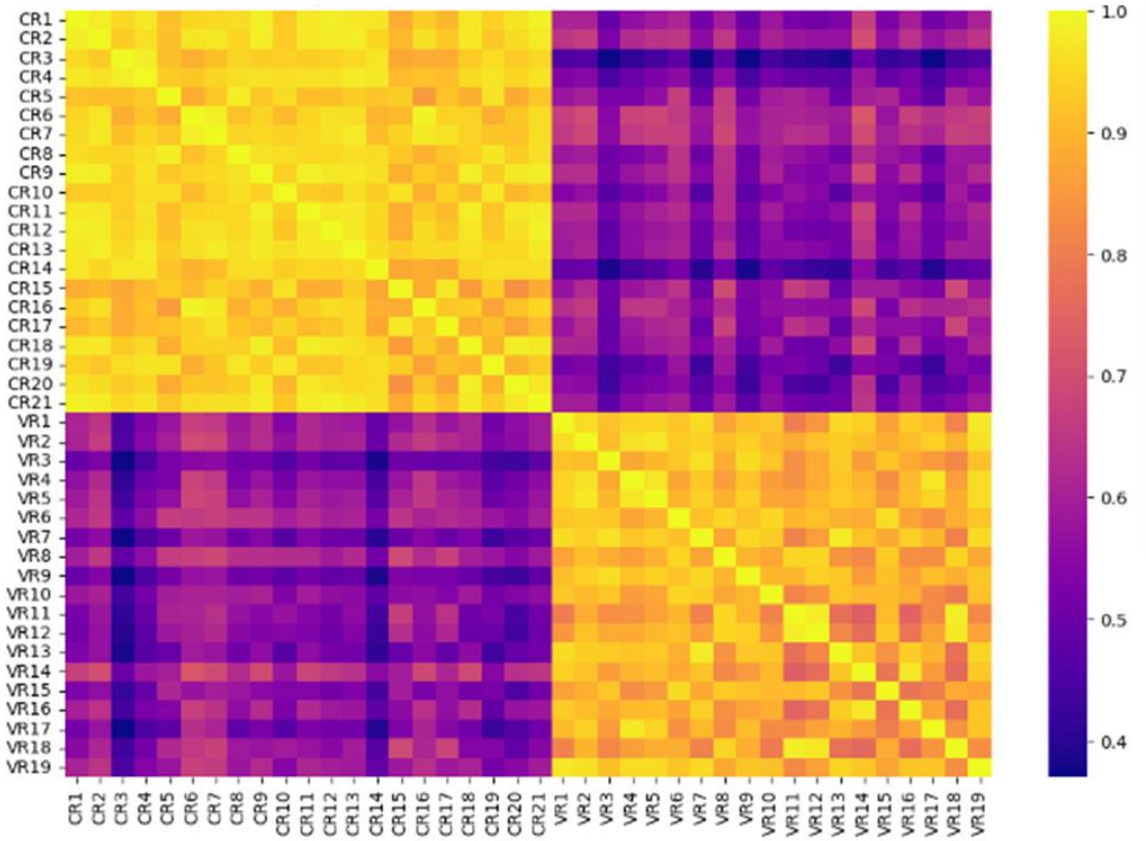
Separación de datos de manera grafica crudos (morado) y fondos (amarillo)



Otra grafica que nos permite observar una clara clasificación de los datos, es un mapa de calor que correlacione los datos, como vemos en la figura 22, esta grafica deja en evidencia como los datos clasificados como crudos (CR) y Fondos (VR) tienen una fuerte correlación al ser relacionados por sí mismos, del mismo modo se evidencia como al ser relacionados entre sí, esta correlación disminuye.

Figura 24

Mapa de calor, matriz correlación del primer modelo de clasificación



- **Segundo y tercer modelo: Regresores %CCR:**

Dado que en nuestro segundo y tercer modelos se emplearon regresores, *Support Vector Regression* (*SVR*) y *Partial Least Squares Regression* (*PLS*), se emplearon métricas pertinentes para estas condiciones de trabajo por ejemplo Error Medio Cuadrado (*MSE*) y el coeficiente de Determinación R^2.

- **Error medio cuadrado (MSE):**

El Error medio cuadrado (*MSE*) es una métrica que calcula la media de los cuadrados de los errores, es decir, las diferencias promedio al cuadrado entre los valores predichos y los valores verdaderos. Un *MSE* más bajo indica un mejor ajuste del modelo, con un

valor ideal de 0, lo que implicaría que el modelo predice perfectamente los valores del conjunto de prueba.

- **Coeficiente de determinación (R^2):**

El Coeficiente de Determinación (R^2) mide la proporción de la variabilidad en la variable dependiente que es explicada por el modelo. Un R^2 de 1 indica que el modelo explica perfectamente la variabilidad de los datos, mientras que un R^2 de 0 indica que el modelo no explica nada.

En nuestro conjunto de datos el modelo *SVR* mostró un *MSE* a penas más bajo (68.9) en comparación con el *PLS* (76.93), lo que indica que el modelo *SVR* tiene un mejor rendimiento al predecir los valores del conjunto de prueba, ya que sus predicciones están más cerca de los valores reales en comparación con las del modelo *PLS*.

En nuestro modelo *SVR* obtuvo un R^2 de 0.47, lo que sugiere que explica el 47% de la variabilidad en los datos del conjunto de prueba. (Figura 25). En contraste, el modelo *PLS* que obtuvo un R^2 de 0.41, lo que indica que solo explica el 41% de la variabilidad en los datos.

Por su parte los modelos SVR y PLS realizados con iteraciones en busca de la partición de datos con el mejor rendimiento, obtuvo un aumento significativo en su rendimiento. El *MSE* del modelo SVR pasó de 68.9 a 10.35 mientras que en el PLS pasó de 76.93 a 24.36 a su vez R^2 aumentó de 0.47 a 0.8 en nuestra máquina de soporte vectorial lo que significa que, tras las iteraciones, el modelo SVR fue capaz de explicar un 80% de la variabilidad en los datos del conjunto de prueba (Figura 24) y de 0.33 a 0.48 en nuestra regresión de mínimos cuadrados parciales. Esto demuestra lo beneficioso que puede ser la implementación de iteraciones en los modelos y en este caso en particular lo beneficioso que ha resultado para el modelo SVR, evidenciando lo útil que puede llegar a ser para medir datos provenientes de la espectroscopía de

masas. La siguientes graficas muestra la relación entre los valores reales y los valores predichos por el modelo para los datos de entrenamiento (azul) y los datos de prueba (rojo) siendo su diagonal la referencia ideal, donde los valores predichos serían exactamente iguales a los valores medidos.

Figura 25

Relación entre valores medidos y predichos modelo SVR Enfoque Tradicional

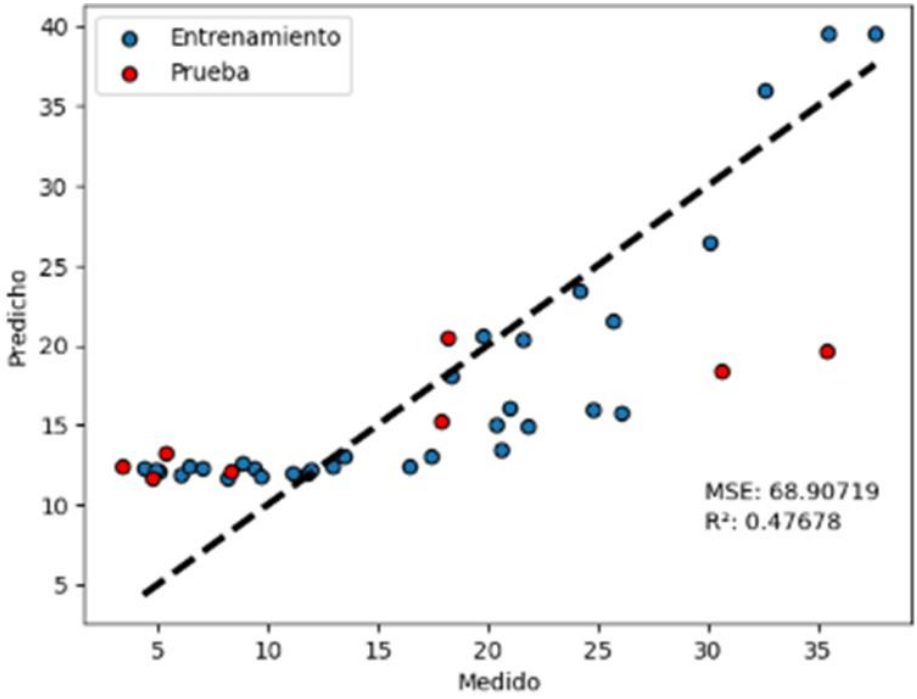
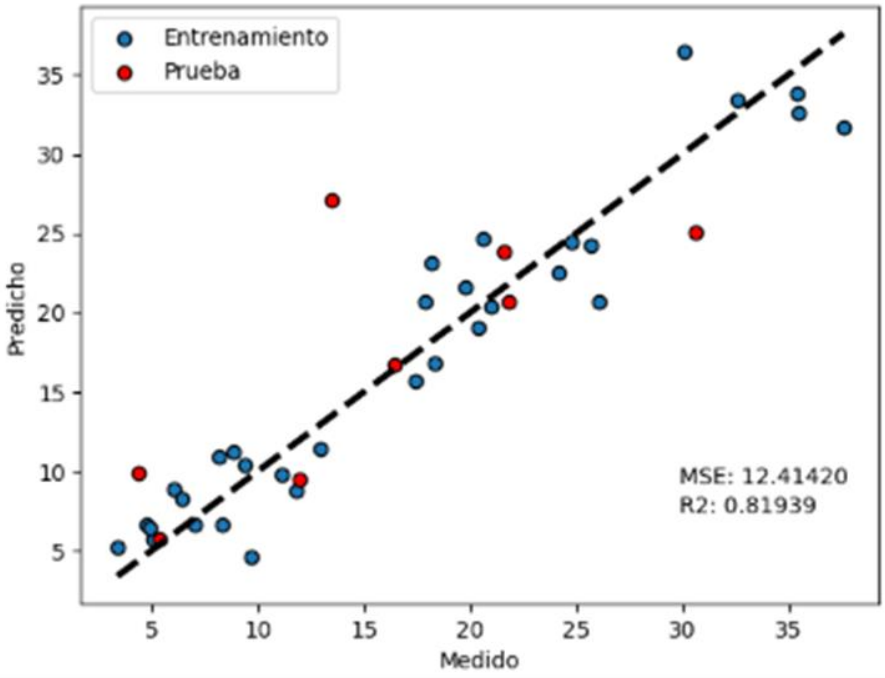


Figura 26

Relación entre valores medidos y predichos modelo SVR Enfoque Iterativo



Al realizar esta comparación y observar visualmente qué tan bien se ajusta el modelo a los datos de entrenamiento y prueba, notamos un desajuste en los datos con un %CCR inferior a 10. Esto se debe a que los crudos livianos generan una menor cantidad de %CCR, lo que los hace difíciles de caracterizar en este tipo de modelos.

- **Cuarto modelo: Regresor S.A.R.A:**

Para el cuarto modelo al tratarse de nuevo de la elaboración de un modelo de regresión por medio de SVR y PLS para cada una de las propiedades SARA, se evaluaron los modelos utilizando las métricas MSE y R^2.

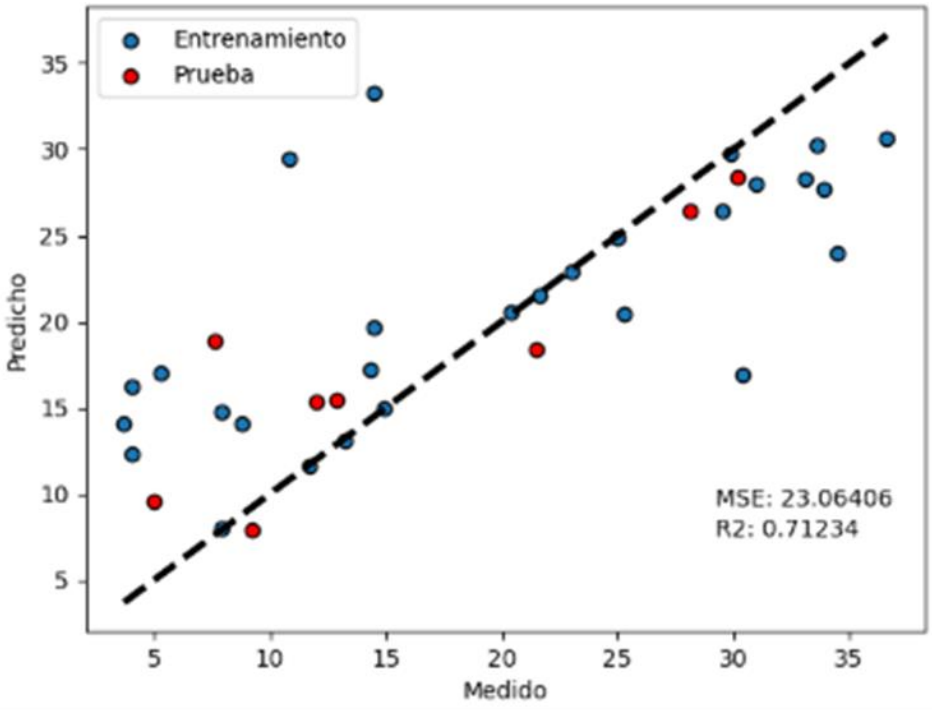
En la predicción de los saturados el modelo SVR demostró ser más preciso en la predicción de la fracción de saturados, con un MSE un poco más bajo (23.07) en comparación con el modelo PLS (29.82). El R^2 de 0.71 para el SVR sugiere que este modelo explica el 71% de la variabilidad

en la fracción de saturados, lo que indica un rendimiento robusto. En contraste, el modelo PLS, aunque capaz de explicar el 62% de la variabilidad (R^2 de 0.62), muestra un mayor error en sus predicciones (Figura 25).

Los saturados encuentran valores típicos entre un 30% y 60% los valores fuera de estas muestras denotan una mayor presencia de las demás fracciones, lo cual puede presentar retos en la tarea de selección en el modelo para valores fuera de estos rangos

Figura 27

Relación entre valores medidos y predichos modelo SVR propiedad ***Saturados***

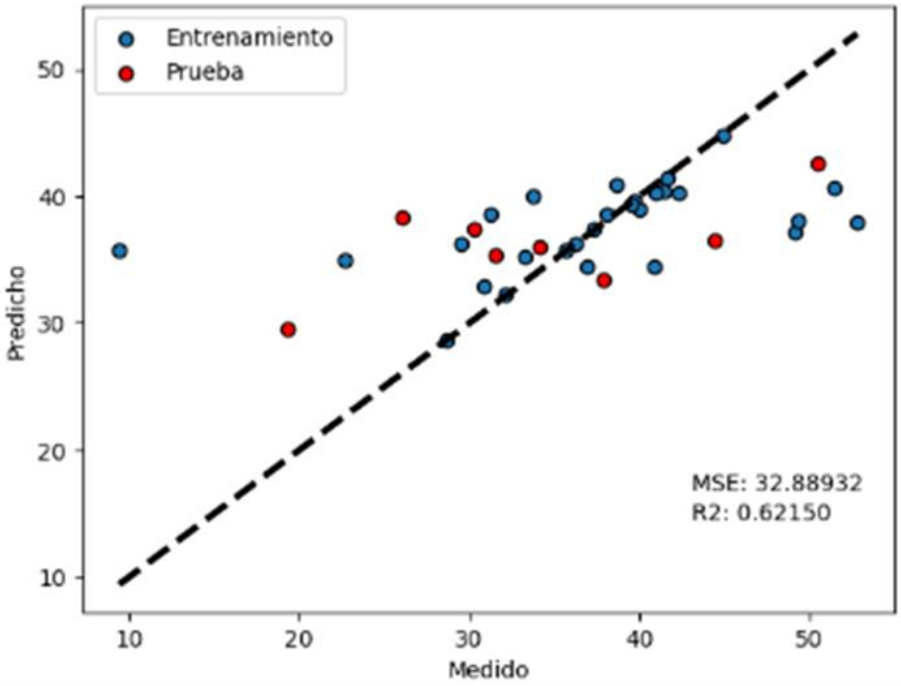


En el caso de la predicción de aromáticos, el modelo PLS superó al SVR de manera significativa. El PLS obtuvo un MSE de 58.69 y un R^2 de 0.62, lo que indica que este modelo explica el 62% de la variabilidad en los datos. Por otro lado, el modelo SVR mostró un rendimiento deficiente con un R^2 de 0.32, lo que indica que explica el 32% de la variabilidad en los datos (Figura 26).

Al encontrarse la mayoría de sus valores en los rangos entre 35% y 45%, da muestras de una salida típica de aromáticos la cual indica una presencia moderada de hidrocarburos cíclicos. Esto sugiere que el crudo tiene una composición balanceada, con una tendencia razonable a ser denso, pero no demasiado pesado.

Figura 28

Relación entre valores medidos y predichos modelo SVR propiedad ***Aromáticos***



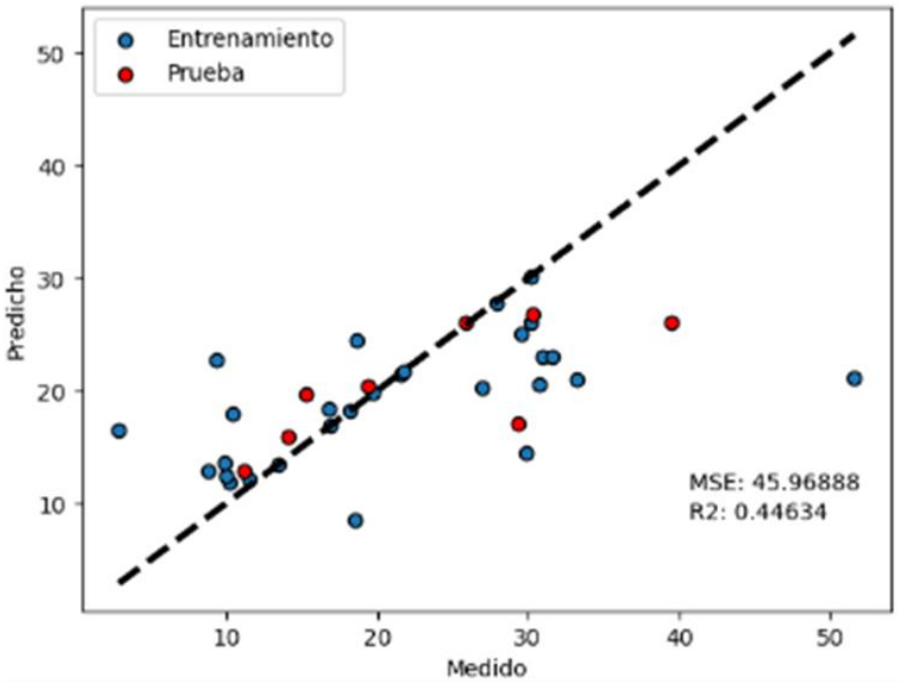
En la predicción de resinas, el PLS se impuso con mejores resultados (Figura 27), con un MSE de 35.0 y un R^2 de 0.57, el SVR, aunque con resultados similares, obtuvo un R^2 de 0.44, y presenta un mayor error (MSE de 45.96), lo que sugiere que el PLS es mejor para modelar esta propiedad específica.

Los valores de las resinas indica que el crudo tiene una moderada cantidad de compuestos polares con heteroátomos (azufre, nitrógeno, oxígeno) al encontrarse en un rango común entré

10% y 20%, las muestras por encima de esté valor enfrentan retos de clasificación al aumentar la cantidad de heteroátomos, lo que puede aumentar su capacidad de procesamiento.

Figura 29

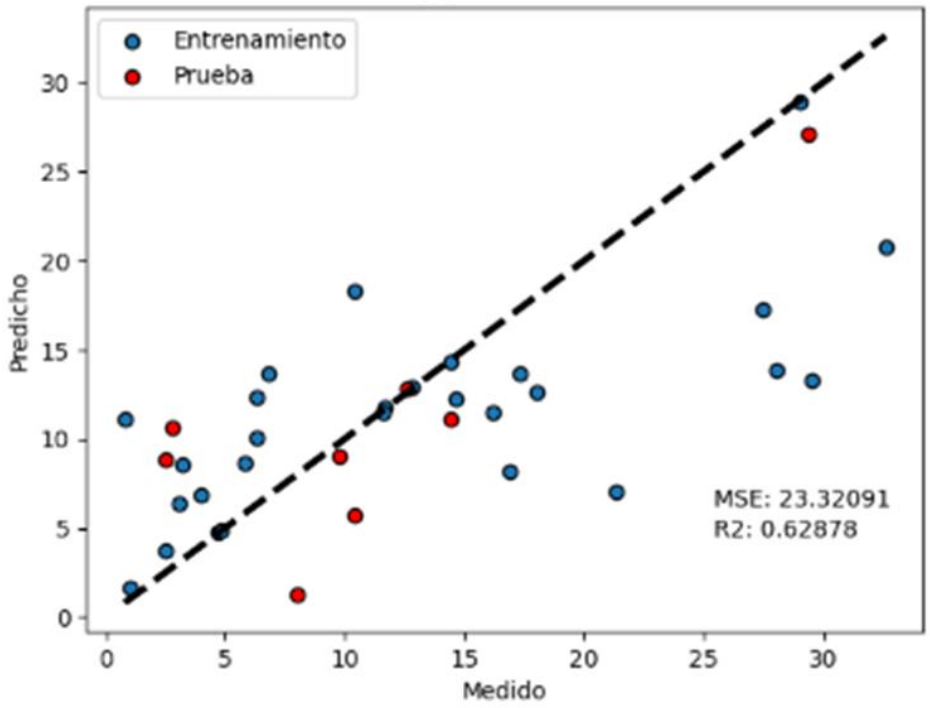
Relación entre valores medidos y predichos modelo SVR propiedad ***Resinas***



Para la predicción de asfaltenos, el SVR y el PLS se encontraron en rendimientos muy similares, esté primero con un MSE de 23.32 y un R^2 de 0.62 (Figura 28). El PLS mostró un rendimiento un poco menor con un MSE de 24.04 y un R^2 de 0.61.

Figura 30

*Relación entre valores medidos y predichos modelo SVR propiedad **Asfaltenos***



Estos resultados sugieren que mientras el SVR es un modelo más robusto para la mayoría de las propiedades SARA, el PLS puede ser preferible para la predicción de aromáticos. En general, la elección del modelo debe basarse en la propiedad específica que se está prediciendo y la importancia relativa de minimizar el error o maximizar la explicación de la variabilidad, a continuación, se presenta una tabla que resume los resultados obtenidos en el quinto modelo de machine learning desarrollado.

Tabla 8

Resumen de resultados métricas del fraccionamiento S.A.R.A

Fraccionamiento SARA	**Modelo**	**Error Medio Cuadrado (MSE)**	**Coeficiente de determinación (R^2)**

Saturados	SVR	23.07	0.71
	PLS	29.82	0.62
Aromáticos	SVR	58.69	0.32
	PLS	32.89	0.62
Resinas	SVR	45.96	0.44
	PLS	35.00	0.57
Asfaltenos	SVR	23.32	0.62
	PLS	24.04	0.61

- **Quinto modelo: Clasificador Crudos - Fondos - Gases:**

Para el quinto modelo desarrollado cuyo objetivo es categorizar muestras de espectroscopía de masas en crudo (2), fondo (1) o gas (0), se usan de nuevo los modelos clasificadores Support Vector Classifier (SVC) y Linear Discriminant Analysis (LDA). Se hace uso de las métricas de puntaje de exactitud (accuracy score), recall y AUC-ROC. Ambos modelos tanto el SVC como el LDA alcanzaron una exactitud del 100% en el conjunto de prueba, demostrando que ambos modelos fueron capaces de clasificar correctamente todas las muestras de espectroscopía de masas en sus respectivas categorías sin cometer errores.

En términos de recall, ambos modelos también lograron un 100%, lo que indica que todas las muestras de cada categoría fueron correctamente identificadas. Sin embargo, en la métrica de AUC-ROC, el modelo LDA alcanzó un 100%, mientras que el modelo SVC obtuvo un valor ligeramente inferior, aunque aún muy alto como se evidencia en la figura 31.

Figura 31

Resultados Accuracy Score, Recall y AUC-ROC modelo clasificación Crudo-Fondo-Gas

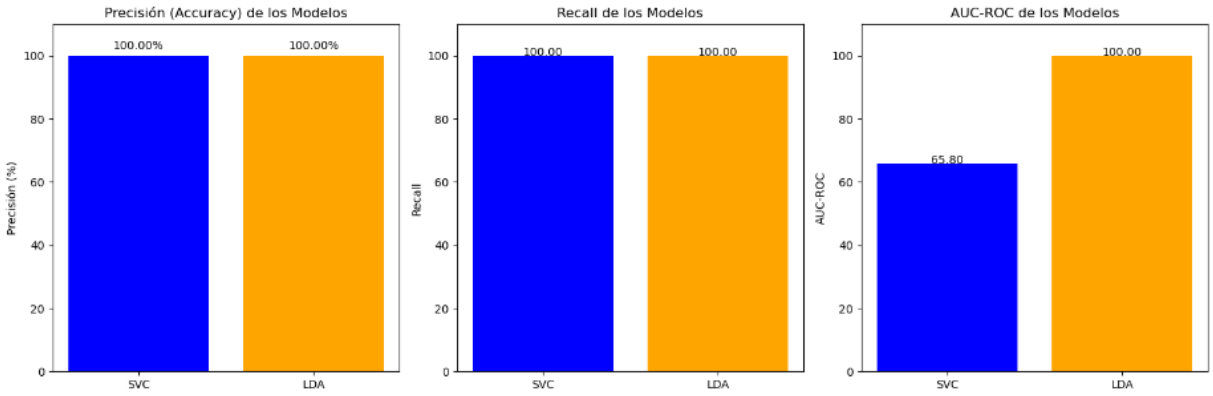
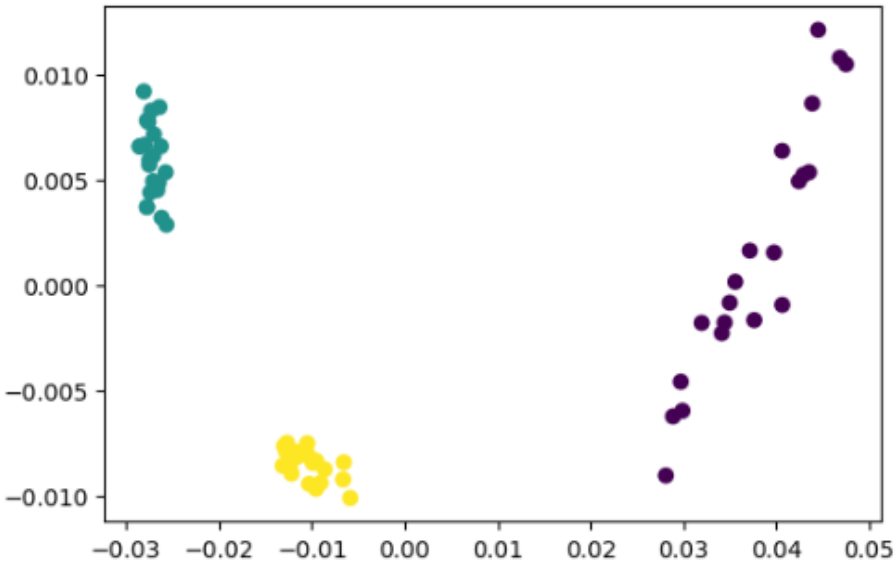


Figura 32

Separación de datos de manera grafica Crudos – Fondo - Gas



- **Sexto y séptimo modelo: Regresor y Red Neuronal índice ASCI:**

Para la medición de los modelos de clasificación del índice ASCI, se utilizaron los clasificadores SVC, LDA, *Random Forest* y *Gradient Boosting*. Las métricas evaluadas fueron en la exactitud (accuracy) en el conjunto de prueba.

En el sexto modelo (Figura 31) la exactitud (accuracy), que refleja el porcentaje de clasificaciones correctas realizadas por el modelo, el SVC alcanzó la mayor precisión con un 17.65%. Los modelos LDA, Random Forest y Gradient Boosting tuvieron precisiones casi idénticas de 5.88%, lo que indica una baja capacidad para clasificar correctamente (Figura 32) los datos la cual puede ser atribuida a la disparidad entre el número de muestras y las características disponibles.

Figura 33

Resultados Accuracy Score modelo clasificación Índice ASCI

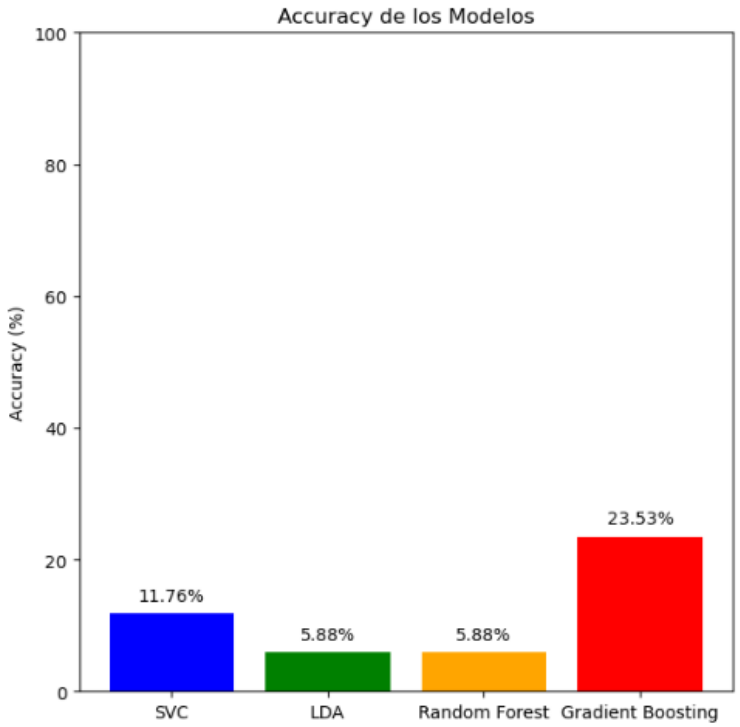
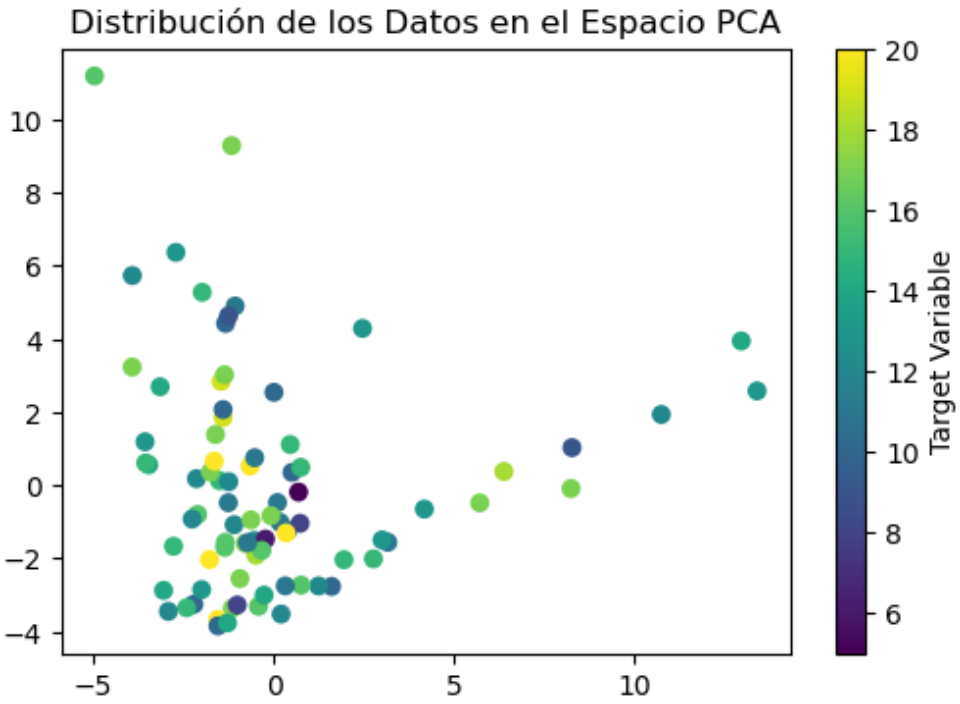


Figura 34

Separación de las 82 muestras entre los 20 grupos posibles de datos ASCI

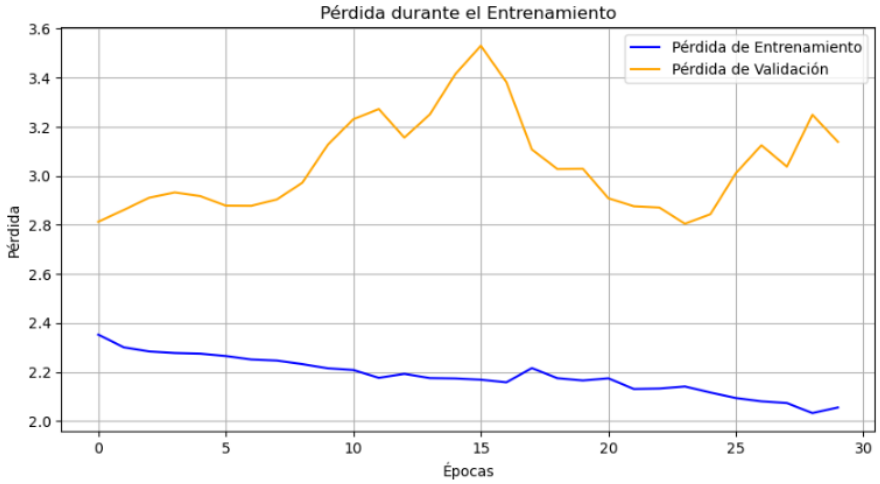


Para el séptimo modelo, se empleó una red neuronal con el optimizador Adam y una tasa de aprendizaje de 0.001. La función de pérdida utilizada fue *categorical_crossentropy*, y las métricas de evaluación incluyeron la exactitud (accuracy). El modelo fue entrenado durante 30 épocas.

A continuación, se presentan los resultados de la función de pérdida del entrenamiento del modelo.

Figura 35

Desempeño red neuronal



El desempeño del modelo de red neuronal para la predicción del índice ASCI revela ciertos desafíos. A lo largo del entrenamiento, la precisión del modelo se mantiene baja, y la pérdida tanto en el conjunto de entrenamiento como en el de validación muestra una variabilidad significativa. Esto sugiere que el modelo podría estar enfrentando problemas de *overfitting* o dificultades para encontrar una buena representación de los datos. La pérdida de validación y la precisión de validación no muestran una mejora clara con el tiempo, lo que podría indicar que el modelo no está aprendiendo patrones útiles en los datos de entrada, reflejando que el modelo tiene problemas para hacer predicciones precisas del índice ASCI.

5.4. Otros resultados

Con el propósito de facilitar y promover el uso de los modelos desarrollados, se creó una interfaz gráfica (GUI) accesible y amigable para el usuario. Esta interfaz permite utilizar fácilmente los modelos, aprovechando la librería *joblib* para guardar los modelos con mejor ajuste

en las mediciones de las propiedades fisicoquímicas. Posteriormente, los modelos pueden ser nuevamente cargados utilizando la misma librería.

La interfaz gráfica fue construida con la librería *tkinter* de *Python*, una herramienta poderosa y versátil para el desarrollo de aplicaciones de escritorio. El objetivo principal de esta interfaz es simplificar la interacción con los modelos de clasificación y regresión, facilitando la predicción de las propiedades de diferentes tipos de crudo, tales como su clasificación en crudo, fondo o gas, así como la predicción de valores de CCR, SARA y ASCI.

5.4.1. Diseño y Arquitectura de la Interfaz

La interfaz gráfica está diseñada de forma modular y funcional, proporcionando una experiencia de usuario clara y eficiente. El diseño de la interfaz se desglosa en las siguientes ventanas:

La ventana principal (*root*) (Figura 35) referente a la ventana inicial que el usuario visualiza al abrir la aplicación. Aquí se presentan cuatro opciones, cada una correspondiente a un modelo diferente, clasificación (fondo-gas-crudo), predicción del %CCR, predicción del fraccionamiento SARA, predicción del índice ASCI. Cada una de estas opciones está vinculada a una ventana individual que contiene las herramientas necesarias para la interacción con el modelo seleccionado.

Figura 36

Ventana principal de la interfaz grafica

Una vez que el usuario selecciona uno de los cuatro modelos en la ventana principal, se abre una nueva ventana dedicada al uso de ese modelo en particular (Figura 35). Cada ventana tiene dos botones principales. Cargar archivos, el cual utiliza la funcionalidad *filedialog* de la librería *tkinter* para que el usuario seleccione el archivo con los datos necesarios (en formato CSV) y el botón predecir quien ejecuta el modelo correspondiente utilizando los datos ingresados y muestra los resultados de la predicción como se evidencia a continuación en la ventana dedicada a la predicción del fraccionamiento S.A.R.A.

Figura 37

Ventana dedicada a la predicción del fraccionamiento S.A.R.A

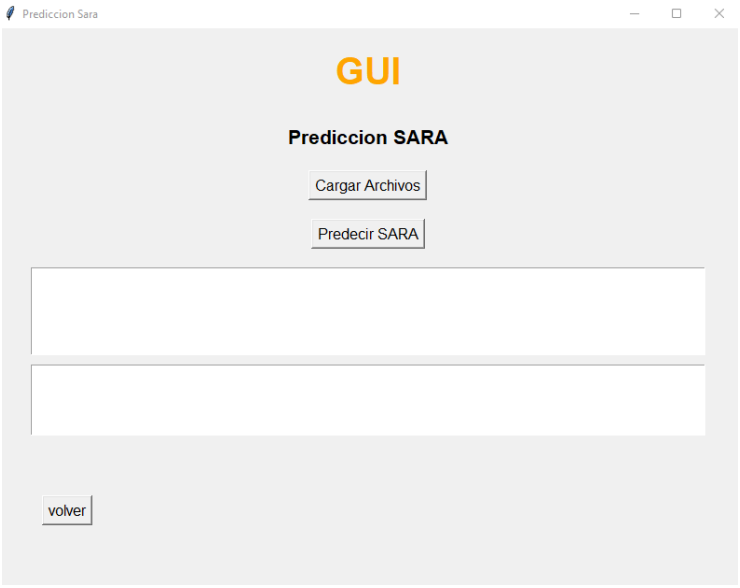


Figura 38

Selección de archivos (CSV) con filedialog de tkinter

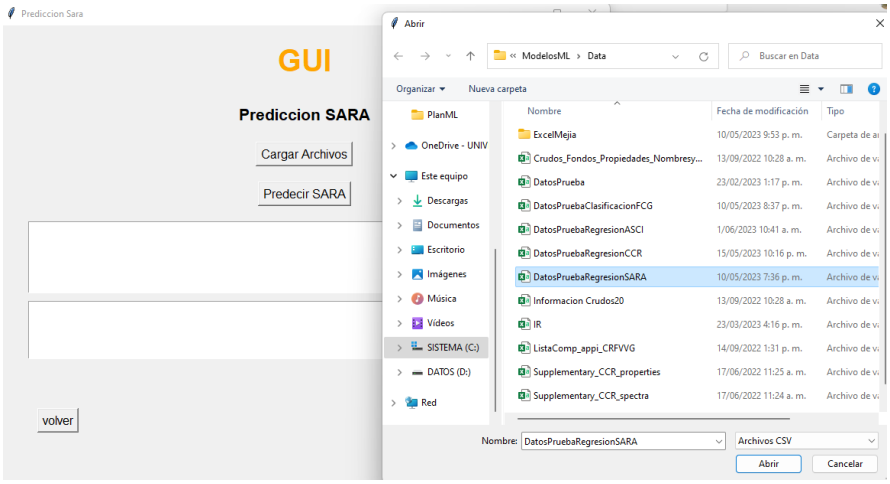
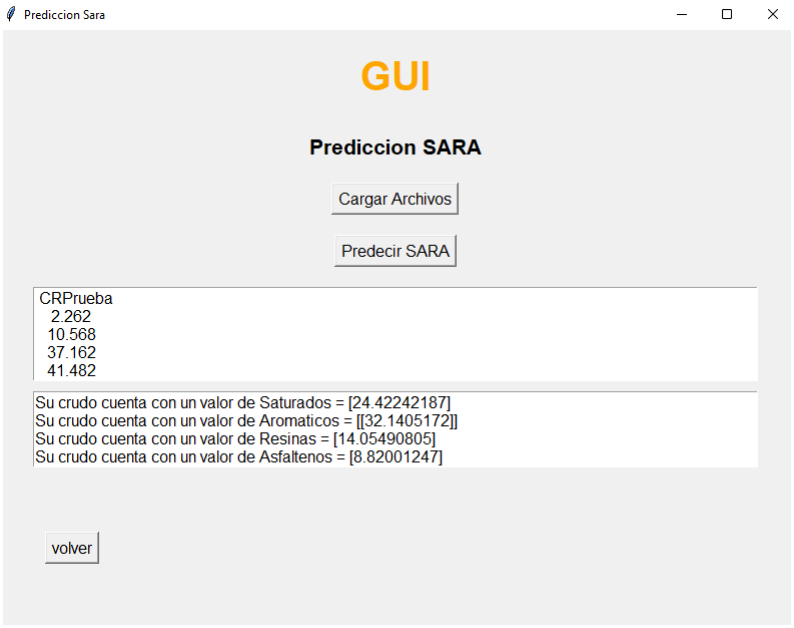


Figura 39

Visualización de los datos cargados y los datos predichos con nuestro modelo precargado



6. Conclusiones

- **Optimización de la calidad de los datos mediante preprocesamiento:**

Se logró construir una base de datos sólida y bien preprocesada mediante técnicas de limpieza, normalización y eliminación de redundancias. Esto resultó fundamental para asegurar que los modelos de machine learning recibieran información precisa y estructurada, lo cual mejoró la confiabilidad y eficiencia de las predicciones de las propiedades fisicoquímicas.

- **Manejo de la alta dimensionalidad mediante PCA:**

Un reto significativo fue la alta dimensionalidad de los datos de espectroscopia, lo que dificultó el rendimiento y la precisión de los modelos. Para abordar este desafío, se utilizó el PCA, que permitió reducir la cantidad de variables al identificar las componentes principales que

capturan la mayor variabilidad en los datos. Esta técnica no solo mejoró la escalabilidad de los modelos, sino que también preservó la información crítica de los datos espectroscópicos.

- **Clasificadores en la diferenciación de crudo, fondo y gas:**

Los clasificadores como LDA y SVC demostraron ser altamente efectivos para separar las muestras entre crudo, fondo y gas, aprovechando la reducción de dimensionalidad de PCA. Estos modelos proporcionaron resultados con alta precisión y tiempos de respuesta reducidos, especialmente en el manejo de datos espectroscópicos, donde las relaciones no lineales fueron capturadas con éxito

- **Regresores para estimar propiedades fisicoquímicas:**

Aunque los clasificadores mostraron un buen desempeño, los regresores, como SVR y PLS, presentaron un potencial significativo en la predicción de propiedades continuas como el %CCR y el análisis SARA. Estos modelos podrían mejorarse mediante ajustes adicionales en el manejo de la dimensionalidad y la incorporación de datos más específicos, lo que mejoraría aún más la precisión en la estimación de propiedades fisicoquímicas complejas

7. Recomendaciones

1. **Ampliar la base de datos de pruebas:**

Es fundamental incrementar la cantidad de datos utilizados para entrenar y validar los modelos. Una mayor cantidad de datos permitiría capturar mejor las variaciones y complejidades inherentes a las muestras de espectroscopía de masas e infrarrojos. Esto ayudará a mejorar la precisión de los modelos y a reducir el riesgo de sobreajuste, logrando un desempeño más robusto y generalizable en la clasificación y predicción de propiedades fisicoquímicas.

2. **Explorar nuevos algoritmos y métodos de modelado:**

Aunque los algoritmos desarrollados han demostrado un buen rendimiento, seria favorable explorar enfoques adicionales de machine learning. Estos métodos podrían ofrecer mejoras adicionales en la precisión y capacidad de generalización de los modelos, especialmente si se trabaja con datos más complejos y variados.

3. **Retroalimentar los modelos con datos reales:**

Es crucial retroalimentar los modelos de machine learning con los nuevos datos generados por los usuarios, ya que esto les permitirá adaptarse continuamente a nuevas condiciones y patrones. Este enfoque de aprendizaje continuo mejorará la precisión y relevancia de los modelos, además de ayudar a identificar posibles desviaciones en los datos o en las predicciones, asegurando que los modelos se mantengan actualizados y precisos.

Referencias Bibliográficas

Antonio, J., Gaona, S., Pablo, J., Manrique, B., & Medina Majé, Y. (2010). *Predicción de la Estabilidad de los Asfaltenos Mediante la Utilización del Análisis SARA para Petróleos Puros. Stability Prediction for Asphaltenes Using SARA Analysis for Pure Petroleum*.

Barea-Sepúlveda, M., Calle, J. L. P., Ferreiro-González, M., & Palma, M. (2024). Machine learning-based approaches to Vis-NIR data for the automated characterization of petroleum wax blends. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *310*, 123910. https://doi.org/10.1016/j.saa.2024.123910

Camargo Reina, K., & Hernandez Oviedo, S. (2014). *CONSTRUCCIÓN DE UN MODELO PREDICTIVO PARA LA DETERMINACIÓN*.

Chen, C., Liang, R., Xia, S., Hou, D., Abdoulaye, B., Tao, J., Yan, B., Cheng, Z., & Chen, G. (2023). Fast characterization of biodiesel via a combination of ATR-FTIR and machine learning models. *Fuel*, *332*(May 2022). https://doi.org/10.1016/j.fuel.2022.126177

Chu, X., Huang, Y., Yun, Y. H., & Bian, X. (2022). Chemometric Methods in Analytical Spectroscopy Technology. In *Chemometric Methods in Analytical Spectroscopy Technology*. https://doi.org/10.1007/978-981-19-1625-0

De Bruyne, S., Speeckaert, M. M., & Delanghe, J. R. (2018). Applications of mid-infrared spectroscopy in the clinical laboratory setting. In *Critical Reviews in Clinical Laboratory Sciences* (Vol. 55, Issue 1, pp. 1–20). Taylor and Francis Ltd. https://doi.org/10.1080/10408363.2017.1414142

Ekman, R., Silberring, J., Westman-Brinkmalm, A., & Kraj, A. (2002). *Mass Spectrometry - Instrumentation, Interpretation, and Applications*.

Farnham, B., Tokyo, S., Boston, B., Sebastopol, F., & Beijing, T. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION*.

Filgueiras, P. R., Sad, C. M. S., Loureiro, A. R., Santos, M. F. P., Castro, E. V. R., Dias, J. C. M., & Poppi, R. J. (2014). Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel*, *116*, 123–130. https://doi.org/10.1016/j.fuel.2013.07.122

Flórez, M. A., Guerrero, J. E., Cabanzo, R., & Mejía-Ospino, E. (2017). SARA analysis and Conradson carbon residue prediction of Colombian crude oils using PLSR and Raman spectroscopy. *Journal of Petroleum Science and Engineering*, *156*, 966–970. https://doi.org/10.1016/j.petrol.2017.06.007

Holla, J. M. (2013). *Modern SPECTROSCOPY FOURTH EDITION*. *332*(2), 13–17.

Hsu, C. S., & Shi, Q. (2013). Prospects for petroleum mass spectrometry and chromatography. *Science China Chemistry*, *56*(7), 833–839. https://doi.org/10.1007/s11426-013-4896-7

Kaya, G., Kaya, N., Amani, M., Rahman, A. H. M. J., Kolomenskii, A. A., & Schuessler, H. A. (2017). Direct Mass Spectroscopy Analysis and Comparison of Middle Eastern and Texas Crude Oils. *International Journal of Organic Chemistry*, *07*(04), 312–318. https://doi.org/10.4236/ijoc.2017.74025

Lamus, C., Guzman, A., Murcia, B., Cabanzo, R., & Mejía-Ospino, E. (2011). *Uso de Análisis Multivariado En La Determinación SARA De Crudos Por Espectroscopia NIR Multivariate Analysis To SARA Determination In Crudes By NIR Spectroscopy*.

Li, S., & Dai, L. K. (2012). Classification of gasoline brand and origin by Raman spectroscopy and a novel R-weighted LSSVM algorithm. *Fuel*, *96*, 146–152. https://doi.org/10.1016/j.fuel.2012.01.001

Meléndez, L. V., Lache, A., Orrego-Ruiz, J. A., Pachón, Z., & Mejía-Ospino, E. (2012). Prediction of the SARA analysis of Colombian crude oils using ATR-FTIR spectroscopy and chemometric methods. *Journal of Petroleum Science and Engineering*, *90–91*, 56–60. https://doi.org/10.1016/j.petrol.2012.04.016

Mohammadi, M., Khanmohammadi Khorrami, M., Vatani, A., Ghasemzadeh, H., Vatanparast, H., Bahramian, A., & Fallah, A. (2021). Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *245*. https://doi.org/10.1016/j.saa.2020.118945

Morgan, M., M., Christie, J., M., Steckler, T., Harrison, J., B., Pantelis, C., Baltes, Lader, …, & M. (2010). Mass Spectroscopy. Encyclopedia of Psychopharmacology. In *Encyclopedia of Psychopharmacology*. https://doi.org/10.1007/978-3-540-68706-1

Niño, A. R., Ramírez, C. X., Hernández, R. C., Picón, H., Guerrero, J. E., & Mejía-Ospino, E. (2019). FTIR-ATR Predictive Model for Determination of Asphaltene Solubility Class Index (ASCI) Based on Partial Least-Squares Regression (PLS-R). *Energy and Fuels*, *33*(12), 12213–12218. https://doi.org/10.1021/acs.energyfuels.9b02829

Noel, F. (1983). *Alternative to the Conradson test carbon residue*.

Peinder, P. De. (2009). Characterization and Classification of Crude Oils Using a Combination of Spectroscopy and Chemometrics. In *Solutions* (Issue december).

Pilařová, V., Plachká, K., Khalikova, M. A., Svec, F., & Nováková, L. (2019). Recent developments in supercritical fluid chromatography – mass spectrometry: Is it a viable option for analysis of complex samples? In *TrAC - Trends in Analytical Chemistry* (Vol. 112, pp. 212–225). Elsevier B.V. https://doi.org/10.1016/j.trac.2018.12.023

Raljević, D., Parlov Vuković, J., Smrečki, V., Marinić Pajc, L., Novak, P., Hrenar, T., Jednačak, T., Konjević, L., Pinević, B., & Gašparac, T. (2021). Machine learning approach for predicting crude oil stability based on NMR spectroscopy. *Fuel*, *305*(August). https://doi.org/10.1016/j.fuel.2021.121561

Samanta, A., Ojha, K., & Mandal, A. (2011). Interactions between acidic crude oil and alkali and their effects on enhanced oil recovery. *Energy and Fuels*, *25*(4), 1642–1649. https://doi.org/10.1021/ef101729f

Schlumberger. (n.d.-a). *Análisis SARA*. Retrieved March 5, 2024, from https://glossary.slb.com/en/terms/p/petroleum

Schlumberger. (n.d.-b). *Petroleum*. Retrieved March 5, 2024, from https://glossary.slb.com/en/terms/p/petroleum

Wilt, B. K., Welch, W. T., & Rankin, J. G. (1998). *Determination of Asphaltenes in Petroleum Crude Oils by Fourier Transform Infrared Spectroscopy*. https://pubs.acs.org/sharingguidelines

Wolf, B. P., Sumner, L. W., Shields, S. J., Nielsen, K., Gray, K. A., & Russell, D. H. (1998). *Characterization of Proteins Utilized in the Desulfurization of Petroleum Products by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry*.