Universidad
del Cauca

EPOXIDACIÓN CATALÍTICA DEL CICLOHEXENO MEDIANTE EL USO DE *MACHINE LEARNING*

FABIÁN RAMÍREZ CHAVARRO

Grupo de Investigación en Catálisis

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Programa de Química

Popayán, Colombia

2024

EPOXIDACIÓN CATALÍTICA DEL CICLOHEXENO MEDIANTE EL USO DE *MACHINE LEARNING*

Trabajo de grado en la modalidad de investigación

Presentado como requisito para optar al título de Químico.

Director

Ph. D. Cristian David Miranda Muñoz (Universidad del Cauca, Grupo de Investigación en Catálisis)

Codirectores

Ph. D. Alfonso Enrique Ramírez Sanabria (Universidad del Cauca, Grupo de Investigación en Catálisis)

Ph. D. Julián Fernando Muñoz Ordóñez (Corporación Universitaria Autónoma Comfacauca, Grupo de Investigación en Computación e Informática Aplicada, MIND)

Línea de investigación

Inteligencia Artificial en Catálisis

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Programa de Química

Popayán, Colombia

2024

Dedicatoria

Dedicatoria especial y póstuma a María José Rozo Lugo, con quien, aunque corto, fue suficiente el tiempo que compartimos para tenerla en mis pensamientos...

Dedicado con mucho cariño a lo mejor que le regaló mi etapa universitaria: mis grandes amigos, y por encima de todo, al gran amor de mi vida, Laura...

Agradecimientos

A todos los profesores que dejaron una indudable huella en mi formación personal, mucho más allá de lo académico, y que me permiten recordarlos de forma especial y por encima de su papel como docentes, como maestros.

A personajes con suma relevancia en el departamento y que fueron parte de mi proceso de formación, y que en algunas etapas pudieron compartir más que una charla académica, una conversación amistosa, como don Óscar y doña Graciela, a quienes debo más de lo que aparenta, por su paciencia y ayuda.

A mis directores de trabajo de grado, Alfonso Ramírez, Cristian Miranda y Julián Ordóñez, quienes estuvieron al cuidado de mi progreso en cada etapa nueva que enfrentaba, y no tuvieron nada más que paciencia y esfuerzo para conmigo y enfrentamos juntos cada una de las dificultades que se nos presentaron, y sin quienes esto nunca hubiese sido posible.

A Lucho, el profesor Lucho, a quien quiero agradecer de manera particular por su gran aporte a mi formación como profesional y como persona, quien me enseñó que esta profesión no se trata de acumular datos en un cuaderno y replicar respuestas en los exámenes, y a quien debo una gran parte de mi forma de ser y de aceptar la vida como es, por eso y más, gracias infinitas, maestro.

A mis mejores amigos, que no son pocos a mi consideración, y que para mí representan una parte indispensable de mi vida y la mi etapa en la universidad me ha permitido forjar estos lazos que fueron parte de mi motor diario para terminar este desafío personal y profesional. Aunque tal vez no los mencione a todos, confío plenamente en que todos mis queridos amigos saben cuánto los admiro, aprecio y quiero. Gracias a todos mis muchachos de Catálisis, que aportaron, en cualquier medida, a que pudiera terminar este arduo y extenso proceso de trabajo de grado, y que siempre cuidaron de mí mejor de lo que yo traté de cuidar de ellos; a Nathalia por siempre ayudarme y ayudarnos con todo lo que fue posible para ella; a Paula por ser el pilar consistente sobre el que pude apoyarme cuando lo necesité y estuvo allí desde el inicio; a Jancarlo, Carolina, Karol y Deicy por apoyarme siempre a pesar de no conocernos por mucho tiempo. Gracias a mis preciadas niñas Shary, Lizett y Stefanía, quienes cuidaron de mí cada vez que me vi en situaciones comprometidas y que me han apoyado desde que iniciamos esta carrera, sin fallarme ni una vez. Gracias a Laura Isabel y Juan Camilo, cuya presencia en mi vida estuvo llena de turbulencias, dimes y diretes y episodios no muy agradables, y que por azares del universo impredecible en el que vivimos, pudimos superar todo ello sabiendo sanar cualquier rencor remanente y haciendo de nuestra amistad una de las que más profundamente siento y aprecio en el corazón, por lo difícil que fue conseguirla. Gracias a Ana, Leidy, Barlahan, Donato, y cualquier otro miembro restante de la promoción 2016, porque sin su presencia en este largo camino este logro no tendría el gran significado que tiene ahora para mí. Gracias a mis compañeros ya graduados y que tanto me apoyaron durante la realización de mi proceso de investigación, Luisa, Carolina, Jean Pierre y Pacheco. Gracias a Keidy, Ronal, Burgos, Reynaldo, Juan Carlos, Daniela, William, Edward, Leo, y todos aquellos quienes, más temprano o tarde en mi carrera, de otro grupo, semestre o programa, pudieron aceptarme y ayudarme a crecer como persona, a formar estos preciosos vínculos que tan bien guardados me llevo de aquí y de por vida.

Finalmente, y obviamente más importante, gracias a la persona más importante en mi vida, mi primer amor verdadero y mi soporte y apoyo firme e incondicional, quien me permitió seguir con

una actitud tenaz ante cada una de las adversidades que se me presentaron en estos ocho años, y que fueron muchas; me permitió darme cuenta que las dificultades son parte del camino y más aún parte de este camino que elegimos; fue mi compañía en los momentos más difíciles, me alimentó cuando no tuve cómo, me cuidó del agua y del frío cuando me faltó abrigo, fue apoyo para mi familia y cada vez que aportaba un grano de arena a mi vida que por momentos se tornada gris y turbia, lo hacía con toda la alegría que podía sentir, porque me amaba y me sigue amando de verdad, y ha hecho de mi carrera como químico haya valido cada segundo de sufrimiento y dudas que haya podido experimentar; gracias por y para siempre, a ti, mi amor, Laura Camila Maca Castro, gracias de verdad.

Nota de aceptación

Director

Ph.D. Cristian David Miranda Muñoz

Jurado

Ph.D. Richard Fernando D'Vries Arturo

Jurado

Ph.D. Danny Alejandro Arteaga Fuertes

Fecha de sustentación: 08 de abril del 2024

Colombia-Cauca-Popayán-Universidad del Cauca

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ESQUEMAS

1 RESUMEN

La Inteligencia Artificial (IA) se ha convertido en una herramienta esencial para la resolución de cuestiones científicas, reduciendo de manera consecuente y significativa las demandas diarias en la mayoría de los sectores de la sociedad. Esto se refleja en áreas tan diversas como el consumo de plataformas de *streaming*, las recomendaciones en buscadores en línea, el mercado bursátil y el reconocimiento de voz. Desde su primera mención en 1956, la simplificación de procesos operacionales ha sido uno de los mayores avances logrados por la IA. Su desarrollo ha sido notable; sin embargo, su aplicación no se limita solo a tareas cotidianas; también ha experimentado un auge significativo en campos como el desarrollo científico e investigativo, contribuyendo enormemente al avance de la ingeniería y la medicina, principalmente. En las ciencias naturales, su uso también ha sido fundamental, logrando hitos importantes como la invención de robots capaces de realizar procesos de laboratorio repetitivos pero imprescindibles.

En el campo de la química, los logros impulsados por la IA no solo se circunscriben a la invención de maquinaria operativa, sino que también abarcan la mejora en la eficiencia de las reacciones químicas. Mediante una de sus ramas más relevantes, el Aprendizaje Automático o *Machine Learning* (ML), ha sembrado la semilla de la Catálisis Digital. Este proceso implica el uso de herramientas de ML para identificar patrones que ayuden a los modelos matemáticos a predecir una variable seleccionada.

Este trabajo tiene como objetivo demostrar una aplicación práctica del uso de herramientas computacionales de IA y ML en el modelado y la predicción del rendimiento de la reacción de epoxidación catalítica del ciclohexeno, el cual fue de **85,54%**. Esto se logró extrayendo decenas de datos experimentales de la bibliografía publicada y organizándolos en un conjunto de datos (*dataset*), que se utilizó para alimentar distintos modelos de ML y de Aprendizaje Profundo o *Deep Learning* (DL, un subconjunto de ML). El propósito fue entrenar estos modelos para predecir el porcentaje de rendimie*nt*o de la reacción, evaluando posteriormente la precisión de las predicciones mediante la replicación de la reacción en el laboratorio. Los descriptores de entrada varían desde características del catalizador hasta condiciones de reacción, todos con una relativa facilidad de obtención, mientras que la salida es el valor del porcentaje de rendimiento.

Los resultados del entrenamiento sugirieron un modelo con una buena correlación relativa entre las variables independientes y la dependiente (**|R^2|=0,7928**). Algunos aspectos relacionados con la descripción del catalizador podrían ser claves para refinar el modelo y mejorar la precisión en las predicciones, que con un porcentaje de error del **94,39%** entre los datos experimentales y los predichos, demuestra que aún tiene potencial de mejorar.

2 INTRODUCCIÓN

2.1 Contextualización y establecimiento de la idea

Desde el año 2010, se han venido realizando diversos experimentos que comparan la capacidad humana frente a la de la inteligencia artificial (IA) y otras herramientas computacionales o robóticas en contextos de laboratorio. Un ejemplo notable es el de George Dahl, estudiante de doctorado en la Universidad de Toronto, quien en 2012 lideró al grupo ganador del *"Merck Drug Discovery Competition"*. A pesar de no tener conocimientos suficientes en química o biología para diseñar la síntesis de un fármaco, estos estudiantes recurrieron a la programación de computadoras para identificar propiedades inaccesibles para científicos especializados [1]. En 2018, el profesor Leroy Cronin de la Universidad de Glasgow demostró que los robots podían superar a los humanos en la predicción y cristalización de nuevos polioxometalatos [2].

Los anteriores son solamente dos de varios antecedentes que se tienen sobre la mejoría potencial de los recursos computacionales sobre los seres humanos en tareas de laboratorio, en la ejecución de la ciencia; hechos como estos empiezan a generar, incluso de forma involuntaria, cuestionamientos dentro y fuera de la comunidad científica que inevitablemente conllevan a comparar la capacidad humana y la robótica para seguir desarrollando la ciencia y la tecnología.

La mayoría de estas comparaciones, mejorías, críticas y aplicaciones que surgen en los campos tanto de la razón humana como de la eficiencia "sin límites" de las máquinas, provienen del auge de la Inteligencia Artificial (IA), que es un término general para referirse al área de la ciencia que usa ordenadores para replicar los comportamientos humanos inteligentes, y que es una combinación de las ciencias computacionales, naturales, políticas y muchas otras áreas de la ciencia, lo cual converge en un concepto estrechamente relacionado con temas más allá de la mera investigación científica, y cuyas aplicaciones han llegado a estar más impregnadas en las vidas cotidianas del humano de a pie de lo que se podría pensar [3].

La IA ha tenido un desarrollo vigoroso desde que fue presentada por primera vez como concepto en 1956 en la Conferencia de la Universidad de Dartmouth por John McCarthy [3]; pero el punto de quiebre que se considera fue el evento que inició la investigación científica en el campo de las redes neuronales artificiales (ANN, por su nombre en inglés *Artificial Neural Networks*, o simplemente NN) fue la presentación del modelo de neurona artificial, en 1943 [3]. Después de estos eventos, que se pueden considerar como el periodo de fundación de la IA, siguieron una serie de sucesos que impulsaron su desarrollo y que, hoy en día, y en gran parte gracias al desarrollo de Internet, la mantienen como uno de los campos de investigación de mayor relevancia y aplicabilidad en el mundo [3].

Esta evolución ha generado preguntas sobre el papel de los científicos, especialmente de los químicos, en un futuro dominado por la IA. ¿Podrán los robots y la IA replicar todas las tareas de laboratorio realizadas por humanos con mayor eficiencia? No hay una respuesta definitiva, pero esta incertidumbre ofrece la oportunidad de fusionar la tecnología más avanzada con las habilidades humanas, creando una armonía entre humanos y máquinas que beneficie la investigación científica. Un desafío significativo para los químicos es la falta de formación en programación, ciencia de datos

e IA en muchos currículos académicos, lo que suele llevar a profesionales de otras áreas, como ingenieros de sistemas, a implementar estos componentes informáticos en proyectos científicos [1].

A pesar de estos obstáculos, se ha comenzado a introducir la era de la "catálisis digital", empleando alternativas computacionales que superan el método tradicional de ensayo y error. Esto incluye desde el uso de la Mecánica Cuántica Computacional (QM) hasta la implementación del Aprendizaje Automático (*Machine Learning*, ML) en el desarrollo de la catálisis [4], [5], [6], [7]. Los estudios catalíticos frecuentemente combinan métodos mecano-cuánticos y la inteligencia artificial, aunque los costos económicos elevados de las técnicas computacionales de QM presentan limitaciones para su aplicación en laboratorios de investigación convencionales [8], [9].

El énfasis en el uso de herramientas computacionales en la catálisis moderna no busca otra cosa que evaluar la viabilidad y las formas de aprovechar estas herramientas para fortalecer, y no debilitar, la profesión química. Ante la hipótesis planteada, surge la principal motivación de los autores de este trabajo, que es ofrecer a la comunidad científica una visión sobre el potencial y prometedor uso de las herramientas computacionales, en especial la IA, en el desarrollo de las ciencias químicas y su posible rol como complemento en la profesión.

2.2 Fundamento teórico

Dado el enfoque del presente trabajo, se hace necesario presentar conceptos básicos sobre la reacción catalítica de interés, su relevancia y la razón de su selección. Asimismo, se muestran definiciones fundamentales sobre la IA y el ML, utilizados en esta investigación como herramientas, con el objetivo de presentarlos de la manera más clara y comprensible posible.

2.2.1 Origen y definición del Machine Learning*:*

Para entender, comprender y manejar mejor el concepto de Aprendizaje Automático o *Machine Learning*, resulta esencial dominar ciertos conceptos fundamentales. A continuación, se presentan los más relevantes.

Inicialmente, es necesario visualizar de forma general pero concreta las diferencias entre el potencial de las capacidades humanas y las de los ordenadores. Esta diferencia a menudo pasa inadvertida para gran parte de la comunidad no científica. Generalmente, los ordenadores superan a los humanos en tareas que implican el manejo y procesamiento de grandes cantidades de datos, como el almacenamiento, procesamiento y realización de cálculos asociados. Por otro lado, en tareas relacionadas con el razonamiento, creatividad, planificación futura y resolución de problemas en general, los ordenadores son menos eficientes [10]. De estas diferencias se puede inferir que el potencial de los computadores como herramientas radica en apoyar la solución de problemas científicos concretos, más que en cuestiones abstractas o "emocionales". Las **Ciencias Computacionales** se definen entonces como un campo que utiliza metodologías computacionales para responder a preguntas científicas. En este campo se incluyen herramientas como la Inteligencia Artificial (IA) y el *Machine Learning* (ML) aplicadas, por ejemplo, en la química, así como el uso de modelos mecano-cuánticos computacionales, que no necesariamente requieren de la IA. Estos últimos son modelos cuya implementación puede ser tan ventajosa como costosa, y a menudo se presentan como herramientas económicamente poco accesibles y de uso limitado [8], [10].

Aunque el término "ciencias computacionales" es útil para delinear el objetivo de la investigación actual, puede resultar demasiado amplio considerando la existencia de otras definiciones más específicas. Un ejemplo es la **Ciencia de Datos (CD)**, que abarca un conjunto más restringido de herramientas o recursos computacionales. La CD se define como la combinación de matemáticas y estadística, programación especializada, Inteligencia Artificial y análisis avanzados, con el objetivo de descubrir información "oculta" en grandes conjuntos de datos [11]. Existen diversas definiciones de la CD; una de las más completas indica que su propósito principal es extraer "valor real" de los datos. Establece además que los datos pueden estar estructurados o no, ser abundantes o escasos y estar estáticos o en flujo. Asimismo, define parámetros para el manejo de datos y consideraciones clave para determinar el potencial de aplicación de su tratamiento [12]. Dada la naturaleza interdisciplinaria de la CD, no es sorprendente su aplicación en campos de las ciencias naturales como la química. Aunque su definición no lo limita, lo más común es que la CD se desarrolle casi obligatoriamente con herramientas computacionales. Por lo tanto, puede incluirse como una rama de las Ciencias Computacionales para los propósitos de este documento.

Hasta este punto, se puede concluir que la Ciencia de Datos (CD) está inmersa en las ciencias computacionales o, en otras palabras, constituye un subcampo de estas. Como implica la definición, la CD requiere el uso de **Inteligencia Artificial (IA)** para su funcionamiento, lo que conlleva a profundizar en este concepto. Desde su primera mención, el concepto de IA ha experimentado una evolución y modificación constantes, resultando en múltiples definiciones válidas en la actualidad. La siguiente es una de las más concisas y útiles para este contexto: la IA es una rama multidisciplinaria de la ciencia que estudia cómo crear máquinas (ordenadores o computadoras) capaces de realizar tareas que normalmente solo los seres humanos pueden llevar a cabo. La IA puede comprender el comportamiento humano, pero no necesariamente debe imitar los métodos biológicos observables [10], [13], [3]. La IA juega un rol central en el desarrollo de esta investigación, ya que es la herramienta principal, o al menos una rama de ella, la que se pretende implementar en la investigación actual. La IA es un campo complejo y, por ello, el nivel de profundidad con que se tratará este concepto u otros relacionados será el mínimo requerido para su aplicación.

Al definir y discutir sobre la IA, a menudo surge la idea de la "capacidad de las máquinas para reemplazar a los humanos", un pensamiento comprensible dada la creciente integración de la IA en la vida cotidiana. Un aspecto relevante relacionado con esta preocupación es la prueba de Turing, propuesta en 1950 por Alan Turing en su reconocido trabajo "*Computing Machinery and Intelligence*". En resumen, esta prueba plantea que un evaluador humano debe distinguir entre respuestas de texto, una de otro humano y otra de una máquina; si el evaluador no logra diferenciarlas, se dice que la máquina "ha pasado la prueba de Turing" y es capaz de replicar la inteligencia humana [10], [14]. Aunque la prueba ha sido objeto de debates informáticos y filosóficos [13] sigue siendo un hito en el desarrollo de la IA. Desde su primera mención, la IA ha experimentado un desarrollo significativo, especialmente en el "tercer periodo dorado", que abarca desde principios de la década de 2000 hasta la actualidad, donde su evolución ha sido más rápida y destacada [3].

Al mencionar la prueba de Turing, esta podría implicar que la IA es un ideal y que no es posible igualar, y mucho menos superar, la inteligencia humana. Esta idea es en cierta medida correcta, y con base en esta "imposibilidad" se plantean dos tipos de IA [13]:

- ✓ IA estrecha: enfocada en tareas específicas, constituye la mayoría de lo que actualmente se conoce como IA. Se puede considerarla la "IA real".
- ✓ IA robusta: teórica y existente solo en hipótesis, se subdivide en IA General, que igualaría las capacidades humanas, incluyendo aprendizaje, planificación futura y autoconsciencia; y la Superinteligencia Artificial, un caso hipotético en el que la IA superaría las capacidades del cerebro humano, perteneciente más al ámbito de la ciencia ficción. Aunque la IA robusta es aún teórica, los científicos continúan explorando formas de desarrollarla.

La IA, como forma de replicar la inteligencia humana, busca resolver problemas de manera similar a cómo lo haría un ser humano, al menos en teoría. Al igual que existen diversas maneras en que un ser humano aprende y actúa para solucionar un problema, hay también varias formas en que una máquina puede abordar una pregunta. Entre estas distintas metodologías de resolución de problemas se destaca el **Aprendizaje Automático** o, en inglés, ***Machine Learning*** **(ML)**, término que se utilizará a lo largo de este documento. Al igual que los conceptos previamente tratados, el ML tiene múltiples definiciones aceptadas. En esencia, el ML se define como una rama de la IA en la que se suministra al ordenador grandes cantidades de datos, permitiéndole extraer patrones y conclusiones para tomar decisiones sin estar programado explícitamente para ello [3], [15], [16].

El término "*Machine Learning*" fue acuñado por primera vez en 1959 por Arthur Samuel, un investigador de IBM, durante sus estudios sobre el juego de damas entre un humano y una computadora de dicha empresa, siendo el resultado favorable para la computadora [16]. Esta primera definición ha dado pie a múltiples interpretaciones posteriores y, como con los demás conceptos ya mencionados, es complicado establecer una única definición "de diccionario". Sin embargo, esto no debería impedir comprender el objetivo general del ML.

Gracias a su versatilidad, el ML ha encontrado una amplia gama de aplicaciones, algunas de las cuales interactúan más directamente con el día a día de las personas de lo que se podría pensar inicialmente. El ML está presente en áreas como el posicionamiento en motores de búsqueda, el procesamiento de lenguaje natural y la ciberseguridad, entre otras [17]. Aunque las aplicaciones más evidentes de estas herramientas computacionales puedan parecer menos relevantes en el desarrollo investigativo de las ciencias naturales, hay, sin duda, aplicaciones significativas del ML en este campo. Para poder aplicar el ML o la IA en general, es crucial tener un entendimiento claro de las herramientas y conceptos asociados a su ejecución.

2.2.2 Funcionamiento y profundización en conceptos clave de ML:

El conocimiento y uso adecuado de los datos constituyen el pilar fundamental del *Machine Learning* (ML). Este manejo de los datos abarca desde la correcta selección hasta la limpieza y procesamiento de estos, empleando los métodos más adecuados. El primer paso es comprender la diferencia entre datos brutos y **conjuntos de datos (*datasets*)**. Un *dataset* suele ser el resultado de organizar los datos brutos, preparándolos para su clasificación y obtención de un **marco de datos (*dataframe*)**, que es una versión del *dataset* más estructurada y organizada.

Un *dataset* puede dividirse en hasta tres subconjuntos: de entrenamiento, de prueba y de validación. El de entrenamiento se utiliza para alimentar el modelo de ML y establecer el patrón de predicción. El

de validación se compone de un grupo de datos ejemplares que se usan para evaluar la precisión del modelo, introduciéndolos como si fueran muestras reales. El *dataset* de prueba permite evaluar el rendimiento del modelo [18]. Con base en esta división, se definen tres etapas clave en la ejecución del ML [18]:

- ✓ **Fase de entrenamiento:** es la etapa en la que se entrena al modelo con la introducción de los datos. El resultado de esta etapa en el modelo de aprendizaje.
- ✓ **Fase de validación y prueba:** en esta etapa se dedica a usar la porción correspondiente del *dataset* para probar la precisión del modelo de ML; el resultado de esta fase suele ser un modelo mejorado.
- ✓ **Fase de aplicación:** en esta fase el modelo listo se dedica a la predicción sobre datos de "la vida real" para los cuales el modelo era pretendido en primer lugar.

A medida que se han ido definiendo conceptos, se hace más evidente la importancia de los **datos** como base de todo estudio de ML. Por lo tanto, es crucial definir ciertos términos relacionados con la constitución de los datos, que son fundamentales para la comprensión de esta investigación.

2.2.2.1 Datos:

Los datos, fundamentales en el ámbito del *Machine Learning* (ML) y la Inteligencia Artificial (IA), se presentan a través de varios conceptos clave que facilitan la comprensión del estudio. Generalmente, los datos se organizan en *datasets* o *dataframes*, que suelen representarse como tablas con filas y columnas. Estos formatos son la base para la mayoría de las definiciones en ML. A continuación, se definen algunos términos importantes [18].

- ✓ **Característica, atributo, campo o variable:** es una columna del *dataset*, y pueden ser tanto datos de entrada como de salida.
- ✓ **Instancia:** es una fila del *dataset*, también llamado **entrada**.
- ✓ **Vector de características o tupla:** es una lista de características (columnas).
- ✓ **Dimensión:** es un subgrupo de atributos que describen un dato. Por ejemplo, una **fecha** se compone de tres dimensiones: día, mes y año. La dimensión de un dato depende del dato en sí.
- ✓ **Conjunto de datos:** un conjunto de datos (*dataset*) es una colección de filas y columnas (instancias y características) y se subdivide hasta en tres subclases dependiendo de la función que cumpla en la ejecución del ML: entrenamiento, validación y prueba.

Estos términos son esenciales para simplificar las explicaciones a lo largo del documento. No obstante, existen más clasificaciones y conceptos derivados relevantes en relación con los datos.

Para ilustrar con un ejemplo práctico, se recopilaron datos hipotéticos sobre la edad, sexo y promedio de consumo semanal de frutas y/o verduras de una muestra de 10 niños de un colegio Z. Estos datos se resumen en la **Tabla 1**.

Tabla 1. Sexo, edad y consumo promedio de frutas/verduras de los estudiantes del colegio Z. "M" representa "masculino" y "F" representa "femenino".

Sexo	Edad	Frutas/verduras consumidas por semana en el último mes
M	9	2,3
M	10	2,0
F	9	1,5
M	9	0,7
F	8	0,9
F	11	3,1
F	9	0,5
M	10	1,0
F	9	1,2

Los datos representados en la **Tabla 1** se clasifican en dos categorías fundamentales: **numéricos** y **no numéricos**. Los datos numéricos, también conocidos como **cuantitativos** o **no categóricos**, son aquellos que se expresan mediante números y pueden estar comprendidos en un intervalo de valores, finito o infinito. Además, los datos numéricos pueden ser **enteros** o **flotantes (continuos)**. Por otro lado, los datos no numéricos, también llamados **cualitativos** o **categóricos**, son aquellos que se limitan a un número finito de categorías o clases distintas, y por lo general no se expresan en términos numéricos [18], [19], [20].

En el ejemplo de la **Tabla 1**, la variable "sexo" es categórica; la variable "edad" es numérica y sus datos son enteros; la variable "frutas/verduras consumidas por semana en el último mes" es numérica y sus datos son flotantes. Esta distinción es crucial, dado que la mayoría de los algoritmos de ML, especialmente en **Python** (el lenguaje utilizado en esta investigación), operan mediante cálculos matemáticos que requieren datos numéricos. Por ello, las variables categóricas en un *dataframe* deben **codificarse**, proceso mediante el cual se convierten en representaciones numéricas.

Habiendo establecido y diferenciado los términos más relevantes relacionados con los datos, es conveniente aclarar la diferencia entre un modelo y un algoritmo, conceptos que a menudo se confunden. Un **algoritmo** es un conjunto de instrucciones sistematizadas diseñadas para calcular los valores óptimos de los parámetros de un modelo. Un **modelo** se refiere a una representación matemática que combina parámetros e instrucciones sobre cómo usarlos para obtener un resultado, aplicando el algoritmo [18].

Para ilustrar estas definiciones, se considera el ejemplo de un modelo simple: la relación lineal entre dos variables, representada por la ecuación **y=mx+b**. Uno de los algoritmos más utilizados para entrenar este modelo, es decir, para encontrar los **parámetros** óptimos "m" y "b" que permitan obtener **valores de salida** o **respuestas** ("y") a partir de los **datos de entrada** ("x"), es el método de mínimos cuadrados.

2.2.2.2 Modelos de regresión de ML:

En ML, los modelos de regresión son herramientas que permiten investigar la relación entre variables y brinda una fórmula matemática ajustada (basada en una función) que permite predecir el valor de una variable dependiente (comúnmente "y") a partir de una o más variables independientes (comúnmente "x") [21]. A continuación, se describen de forma simple le funcionamiento de los modelos de regresión de ML aplicados en la investigación.

2.2.2.2.1 Regresor de Árbol de Decisión [22]*:*

Consiste en dividir consecutivamente el conjunto de datos principal en subconjuntos a partir de selección de características, es decir cada característica indica una división por eso se forman ramas o clasificaciones asignándole un valor numérico. Se diferencian de los modelos de árboles de decisión porque se enfocan en analizar valores categóricos en cambio *Decision Tree Regressor* usa valores numéricos, por eso el nombre de regresión.

2.2.2.2.2 Regresión Lineal [23]*:*

Este modelo se enfoca en buscar una línea recta que describa la relación que hay entre una variable dependiente con una o más variables independientes que son variables respuestas. Es uno de los más simples y a pesar de ser eficiente cuando la relación entre variables es próxima a la linealidad, no lo es tanto en relaciones más complejas ya que este se ciñe a la búsqueda de una línea recta que describa dicha relación.

2.2.2.2.3 Regresor LightGMB [24]*:*

Se basa en la técnica de *boosting* que significa formar varios árboles de decisión y después tiende a unirlos, esto hace que sea más rápido cuando toma la decisión de clasificarlos de acuerdo con las características que va encontrando. Cada árbol que se forma siempre mejora su eficiencia y corrige los errores; esto lo hace realizando un gradiente para minimizar la función de pérdida que es la diferencia entre las predicciones del modelo y los verdaderos valores.

2.2.2.2.4 Regresor estocástico de gradiente descendente [25]*:*

Los gradientes siempre hacen que se minimicen las funciones, y este modelo también minimiza la función de pérdida; lo que hace el algoritmo es encontrar el valor de los parámetros que minimizan la función y luego realiza una iteración con el fin de que sea más eficiente y se aproxime a los valores reales para que se más confiable al realizar la predicción.

2.2.2.2.5 Cadena Kernel [26]*:*

Los datos a los que se van a aplicar los modelos se encuentran en un espacio específico, pero puede no presentar un patrón para modelar. *Kernel Ridge* lo que hace es realizar una multiplicación vectorial de la matriz de datos para llevarlos a otro espacio dimensional y examinarlos en ese nivel. El *kernel* hace referencia ahora al modelo que se ajusta a los datos en ese nivel. Y ahí se encuentran hiperparámetros para encontrar la mejor clasificación.

2.2.2.2.6 Red Lineal Elástica [27], [28]*:*

Consisten en combinar dos funciones de regularización del modelo, se llaman *lasso* y *ridge*, ambas buscan los ceros de las funciones, pero con vectores distintos, y al combinarlas lo que se logra es

evitar que se elijan los valores donde la función a minimizar es cero, esto se ve reflejado en evitar elegir valores irrelevantes o que no aporten a la elección de características importantes.

2.2.2.2.7 Cadena Bayesiana [29]*:*
Los métodos bayesianos consideran la probabilidad. Este modelo se diferencia de los demás por no dar un valor único como respuesta a la aplicación del modelo en los datos, sino que entiende que puede haber una probabilidad en lo que se estima.

2.2.2.2.8 Regresor de gradiente de refuerzo [30]*:*
También es un modelo basado en árboles de decisión, pero no tan profundo, es decir, no hace tantas separaciones entre los datos con las características, y sigue los mismos pasos para minimizar la función, encontrar los mejores parámetros e iterar respecto a esos valores para encontrar el mejor valor de regresión.

2.2.2.2.9 Regresor Vector de Soporte [31]*:*
Cuando se tiene un problema de clasificación se utiliza un modelo que se llama "máquina vector de soporte", consistente en la multiplicación los datos por vectores para encontrar un hiperplano ya que en el espacio donde estaban los datos no se ajustaba a un modelo. Su objetivo es minimizar la complejidad de los datos, y este se caracteriza por contener un parámetro que permite contener errores dentro de una zona tolerante.

2.2.2.3 Aprendizaje profundo:
El ML es una herramienta capaz de aprender y realizar predicciones a partir de datos. Existen numerosos modelos y algoritmos dentro del ML, y algunos son más potentes y eficientes que otros. Uno de los enfoques más avanzados en ML se inspira en la capacidad del cerebro humano para establecer conexiones y tomar decisiones, emulando el funcionamiento de las **neuronas**. Aunque es improbable que un método de ML alcance la complejidad del cerebro humano, esta aproximación ha demostrado ser muy efectiva en aplicaciones actuales. Estos modelos especiales en ML son conocidos como **Redes Neuronales Artificiales (ANN, *Artificial Neural Networks*)** y forman parte de un subcampo del ML denominado **Aprendizaje Profundo (DL, *Deep Learning*)**. El DL es un método de ML capaz de procesar grandes cantidades de datos con menos preprocesamiento, entregando resultados superiores con menor intervención humana. A través de iteraciones y aprendiendo de sus propios errores, el DL puede mejorar sus predicciones [15], [18], [32], [33].

El DL generalmente requiere menos intervención humana, ya que las ANN pueden procesar datos no estructurados y descubrir por sí mismas las características relevantes para la toma de decisiones [32], [33]. Por ejemplo, al someter una ANN a información referente a las características que diferencian a distintas especies de animales domésticos, esta puede determinar las características clave de diferenciación (como la forma de las orejas), mientras que en un algoritmo de ML tradicional sería el programador quien debería especificar estas características [32]. En resumen, el DL es un método de entrenamiento en ML más potente, que muestra mejores resultados con grandes volúmenes de datos, gracias a su funcionamiento inspirado en el cerebro humano.

Al igual que las neuronas en un cerebro, una ANN consta de **nodos (o neuronas artificiales)** interconectados que trabajan juntos para resolver un problema. Cada nodo es un módulo de software y, en conjunto, forman un modelo de DL gobernado por un algoritmo que resuelve problemas matemáticos en términos de procesamiento de datos numéricos.

Las ANN se componen de tres partes básicas, llamadas **capas** (conjuntos de neuronas/nodos) y dependiendo de la distribución pueden ser más o menos profundas. La **capa de entrada** es aquella conexión de neuronas que recibe la información exterior, la procesa y la distribuye a las siguientes capas; las **capas ocultas** toman la información bien sea de la de entrada o de otras capas ocultas, procesan aún más la información y la siguen pasando a otras capas; la **capa de salida** recibe la información procesada de todas las capas ocultas y provee el resultado dependiendo de la intención del entrenamiento [18], [33].

En la **Figura 1** se ilustra la estructura de una ANN simple, donde cada círculo representa un nodo (neurona artificial). Es notable cómo se busca replicar el comportamiento humano, aunque aún no con la misma eficiencia, pero con beneficios significativos en áreas como el análisis financiero, diagnóstico médico a través del análisis de imágenes, reconocimiento y procesamiento de voz [32], [33], aunque también se ha aplicado de forma satisfactoria al ámbito científico investigativo, con interesantes propuestas en el mundo químico.

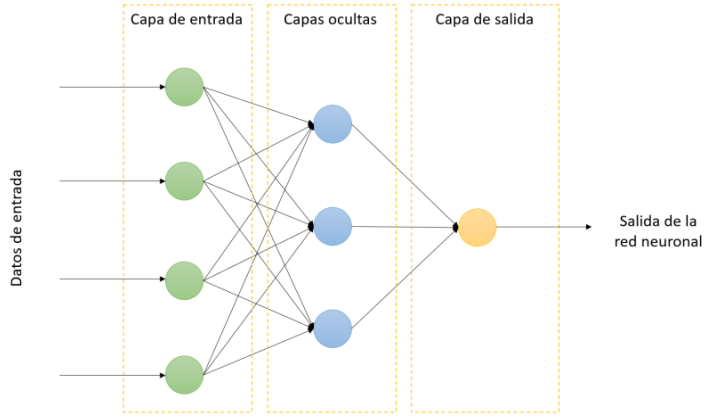


Figura 1. Estructura de una ANN simple con una capa de entrada, una capa oculta y una capa de salida. Tomado de [34].

Habiendo explicado la mayor parte de los términos más usados en el contexto del ML y algunos de los complementos y herramientas de este, es necesario entonces contextualizar de manera cronológica la evolución de las aplicaciones del ML en química y el contexto en la reacción de interés, la **epoxidación catalítica del ciclohexeno**.

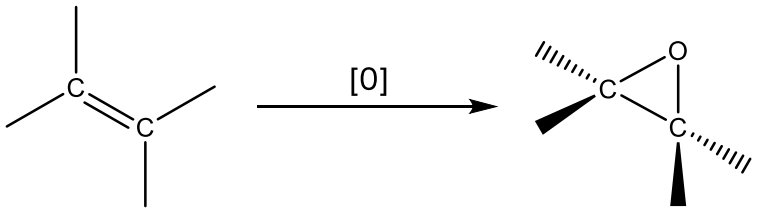
2.3 Antecedentes

La catálisis ha sido un elemento clave en el avance de innumerables reacciones químicas a lo largo de la historia, facilitando procesos desde la síntesis del ácido sulfúrico hasta la reducción del CO_2 [35], [36]. A pesar de su importancia, su desarrollo ha dependido casi exclusivamente del método de

"ensayo y error" [35], [37]. Aunque este enfoque no constituye un método estandarizado, representa una simplificación del proceso científico. Una reacción de gran relevancia industrial, en la que la catálisis ha tenido un papel crucial, es la epoxidación de alquenos. Esta reacción es especialmente significativa en la producción de intermediarios clave para la síntesis de productos químicos finos y fármacos [38]. Esta sección abordará la evolución histórica, la importancia actual y la aplicación de las Ciencias Computacionales en la reacción de epoxidación de alquenos.

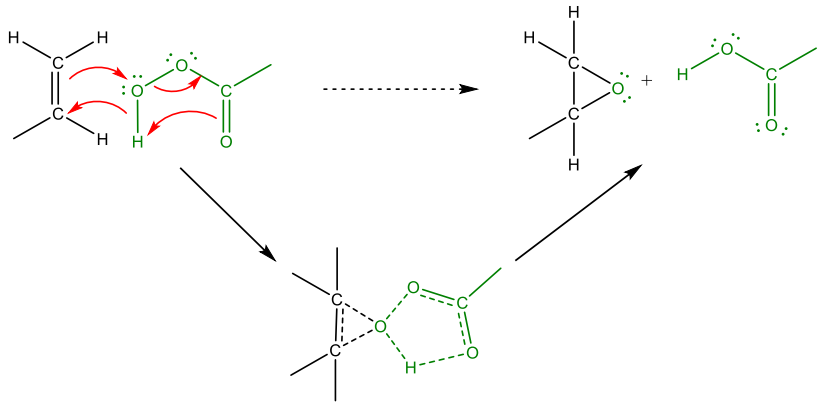
2.3.1 Descripción general de la reacción:

La reacción de epoxidación de olefinas, conocida desde hace décadas, ha sido objeto de numerosos estudios enfocados en su eficiencia, dada la útil versatilidad de los epóxidos como intermediarios para sectores productivos de mayor escala. La reacción se describe de forma general, como se muestra en el **Esquema 1**.



Esquema 1. Ecuación química general para la epoxidación de alquenos/olefinas. Tomado y adaptado de [39].

La reacción de epoxidación de alquenos ha sido, al menos desde la década de los 1900, estudiada en términos del mecanismo que involucra el uso de peroxiácidos, que hasta finales del siglo XX como mínimo, seguía siendo de gran utilidad [39]. El **Esquema 2** muestra el mecanismo resumido de la oxidación de dobles enlaces olefínicos mediante peroxiácidos carboxílicos.



Esquema 2. Mecanismo general de la reacción de epoxidación de alquenos entre el propeno y el ácido peroxiacético. Se obtienen como productos el peróxido del propeno y el ácido carboxílico correspondiente (ácido acético). Tomado y adaptado de [40].

El método de oxidación por peroxiácidos presenta desventajas notables, como el alto riesgo de descomposición no controlada, su alto costo y la producción del respectivo equivalente de ácido carboxílico [41]. Estas desventajas pueden ser mitigadas mediante la introducción de un catalizador adecuado en el sistema. Así, se han estudiado numerosos sistemas catalíticos aplicados a la epoxidación de alquenos, dando lugar a diversos resultados que han contribuido a perfeccionar la reacción a lo largo de los años [42].

La implementación de estos sistemas catalíticos ha proporcionado a la comunidad científica una considerable variedad de enfoques para llevar a cabo reacciones de epoxidación, especialmente de moléculas claves a nivel industrial como el eteno, el propeno o el **ciclohexeno** [38]. Dentro de las ventajas industriales de la epoxidación de olefinas, la producción del epóxido de ciclohexeno es de particular relevancia, ya que es una de las moléculas más utilizadas en la producción de químicos finos y en la síntesis de polímeros de poliéter [43], considerándose actualmente una molécula DE REFERENCIA para la ejecución de la reacción catalítica. Históricamente, la epoxidación del ciclohexeno se ha realizado tanto de forma homogénea como heterogénea, sin una preferencia particular por algún sistema catalítico, aunque en cada tipo de catálisis se han marcado tendencias en la evolución de la reacción.

En términos de sistemas catalíticos heterogéneos, dos de los metales de transición más estudiados y aplicados han sido el titanio (Ti) y el molibdeno (Mo). El Ti ha sido ampliamente utilizado como especie activa en catalizadores soportados, especialmente en sílica y alúmina, proporcionando resultados prometedores para aplicaciones industriales [38], [44], en particular con el uso de hidroperóxidos orgánicos como el *tert*-butil hidroperóxido (TBHP) [38]. Por su parte, el Mo ha mostrado excelentes resultados, especialmente cuando se incorpora mediante polioxometalatos (POM) y en materiales metalo-orgánicos (MOF), siendo estos algunos de los sólidos más estudiados que contienen Mo como material catalítico [45]. La incorporación de sistemas catalíticos con POM ha sido notoria, utilizando principalmente ácidos molibdovanado-fosfóricos [45]. Otra forma ampliamente estudiada de aplicar la catálisis de Mo en la epoxidación del ciclohexeno involucra materiales soportados, especialmente la sílica, con Mo como sitio activo [45].

En cuanto a la catálisis homogénea, esta también tiene una amplia representación en la reacción, siendo los complejos organometálicos los catalizadores predominantes [38]. Se emplean principalmente metales como el molibdeno (Mo), el manganeso (Mn), el renio (Re) y el vanadio (V) [38], [42], [45]. La aplicación de complejos organometálicos de Mo es evidentemente extensa en la catálisis de esta reacción y uno de los compuestos históricamente más usados para este fin ha sido el $[MoO_2(acac)_2]$, así mismo como se han usado varios otros complejos con el sitio activo MoO_2, como el $[MoO_2(oxina)_2]$ (en donde oxina equivale a 8-hidroxiquinolina) [46], [47]; sin embargo, los complejos de Mo no han sido los únicos catalizadores empleados en esta reacción y en fase homogénea, catalizadores de V, como el sistema Venturello $[PWO_4O_{24}]^{3-}$ [48], de Re como el sistema Herrmann CH_3ReO_3 [49] o el sistema de Jacobs que incluye un complejo de Mn [50] han sido aplicados.

Es evidente que el estudio de la epoxidación catalítica del ciclohexeno es lo suficientemente extenso como para suponer que la reacción ha sido optimizada significativamente. Esto ha generado una gran cantidad de datos experimentales con cada nueva investigación sobre la catálisis de esta reacción. A menudo, muchos de estos datos no se aprovechan completamente, ya que la atención se centra en variables de respuesta de eficiencia catalítica como el tiempo de recambio (TOF) o el porcentaje de rendimiento. Sin embargo, se tiende a descuidar el interés científico por el aprovechamiento de datos, como investigar qué ocurre con el TOF al modificar una condición específica. Es precisamente esta cantidad de datos generados en el laboratorio la que se pretende aprovechar al utilizar herramientas computacionales, con el fin de seguir cumpliendo con la premisa de hacer la reacción cada vez más eficiente ante el crecimiento de la demanda inherente al desarrollo poblacional. Este enfoque no solo es relevante para esta reacción específica, sino también para cualquier otro sistema catalítico.

2.3.2 Evidencias de la aplicación de las Ciencias Computacionales en catálisis:

Como ya se mencionó antes en la sección **1.1**, la incursión de la IA en el ámbito científico ha tenido gran repercusión, al menos en el surgimiento de la presente investigación a nivel institucional y regional. Pero ¿existen aplicaciones más técnicas que, a nivel de laboratorio, puedan mostrar cómo la IA y las Ciencias Computacionales han incursionado en el ámbito de la catálisis? Con el fin de brindar evidencia suficiente para dictaminar si la pregunta planteada se responde y de qué forma, se realiza un recorrido por los antecedentes más relevantes respecto de la aplicación de los recursos computacionales en la catálisis.

El ML se ha implementado en el desarrollo e investigación de la catálisis desde la década de 1990, y en sus primeros pasos usaba redes neuronales artificiales (ANN, *Artificial Neural Networks*) para relacionar propiedades fisicoquímicas de los catalizadores y las condiciones del catalizador con el rendimiento catalítico [8], aunque este trabajo se limitaba en cuanto al número de sistemas catalíticos estudiados. Los trabajos con ANN en esta misma década fueron perseverantes, observándose que este tipo de herramienta computacional es especialmente útil en el procesamiento de datos bien definidos, pero de difícil comprensión, como la espectroscopía o secuencias biológicas [51]. Fue entonces hasta el año 2000 cuando se reportó la primera revisión del estado del arte del uso de las ANN en ingeniería química [52]. Como resultado de la revisión del avance de la implementación de las ANN en la década de 1990, se observó que su aplicación ya había sido reportada en la literatura en un amplio rango de reacciones catalíticas tales como el reformado de metano, la oxidación selectiva del CO, la reducción electroquímica del CO_2, la hidrogenación del CO_2, la transesterificación de triglicéridos y la epoxidación de olefinas de cadena extensa [53].

Una de las investigaciones que mejor ayuda a dimensionar el refinado de la capacidad de predicción de las ANN ha sido aquella publicada en el 2012 por Arcotumapathy et al. [54], en la cual se estudia la reacción de reformado de metano y se entrena la red a partir de un *dataset* de 200 entradas. El resultado general de la capacidad predictiva de la ANN es aparentemente muy alto, alcanzando niveles de paridad casi ideales. Se utilizaron descriptores relativamente sencillos de obtener: % de Ni, composición y tipo de soporte, composición y tipo de promotor, temperatura de reducción del catalizador, temperatura de reformado, presión, relación vapor/carbono y tiempo en flujo; todos estos

descriptores son fácilmente obtenibles del estudio y adecuación de las condiciones de reacción, y en retrospectiva, al analizar el resultado presentado por los autores, es una ventaja enorme el hecho de no tener que depender de herramientas computacionales más costosas como los modelos mecano-cuánticos (QM). Sin embargo, es posible también hacer uso de herramientas avanzadas de QM para complementar u optimizar los estudios con ANN y proporcionar así mejores resultados [55].

La catálisis homogénea también ha tenido incursión en el campo de la IA y el ML, aunque en una medida menos notoria pero igualmente importante, en estudios que han ido desde la predicción de enantioselectividad hasta la elucidación de mecanismos catalíticos [56], [57]. La catálisis homogénea permite tener otros enfoques moleculares que hacen de los descriptores variables más relacionadas a las propiedades moleculares de la materia y permiten la inclusión de parámetros usualmente no vistos en la catálisis heterogénea, como aquellos relacionados con intermediarios (mecanismos) de reacción [57].

Las reacciones de epoxidación también han sido objeto de estudio e implementación de métodos de IA y ML. Ya en 2010, Baumes y su equipo [58] reportaban el modelado de las velocidades iniciales de la reacción catalítica de epoxidación de olefinas de alto peso molecular mediante el entrenamiento de ANN con datos generados internamente (laboratorio). Los resultados más notorios fueron la reducción de tiempo en procedimientos de laboratorio y la revelación de información más profunda acerca de la dependencia y las relaciones de la actividad catalítica con descriptores como los sustituyentes del catalizador de Ti/SiO_2.

El modelado del comportamiento catalítico de materiales basados en Ti-silicatos en epoxidación de olefinas se realizó incluso antes, en 2005, y en este trabajo se usaron descriptores extraídos del catalizador o de la composición sintética del mismo, con la importante observación de que se incluyeron datos espectrales de los mismos materiales, en particular datos de difracción de rayos X, DRX. Aunque la interpretación podría resultar ambigua en algunos aspectos, estos aportaron información relevante respecto de variables como tamaño de partícula y poro y tipo de sistema cristalográfico, los cuales se asociaron de forma sorprendente y satisfactoria con su incidencia en el rendimiento catalítico [59].

Los resultados generales de estas investigaciones parecen ser bastante precisos y prometedores para las implementaciones actuales y futuras de herramientas de ML y DL en sistemas catalíticos, e incluso en otros tipos de sistemas como el diagnóstico médico o el diseño de fármacos.

Las aplicaciones de este tipo de herramientas se extienden más allá de la mera reacción de epoxidación [8], [57] y suponen un campo de investigación de elevada importancia en el desarrollo continuado de la catálisis y la química en general, brindando aún mayor relevancia y expectativa a investigaciones como la presente.

Tras haber explorado los fundamentos más importantes y algunas de las numerosas aplicaciones de las Ciencias Computacionales en la química y la catálisis, se presentan los objetivos del trabajo de grado. Además, se expondrá cómo se lograron alcanzar dichos objetivos, así como los resultados obtenidos y los análisis correspondientes. Se explicará la relevancia de esta investigación y las razones

para apoyar futuros proyectos similares. Asimismo, se discutirán algunos aspectos susceptibles de mejora y las posibles estrategias para lograrlo.

3 OBJETIVOS

3.1 Objetivo general

Realizar la epoxidación del ciclohexeno a través de un diseño construido a partir de un modelo de *Machine Learning*.

3.2 Objetivos específicos

- Construir un conjunto de datos [experimentales] a partir del análisis del estado del arte referente a la epoxidación del ciclohexeno con los descriptores característicos de las condiciones tipo de catalizador, agente oxidante, temperatura, tiempo de reacción y solvente, necesarios para el entrenamiento de un modelo de *Machine Learning*.
- Construir un modelo de *Machine Learning* para predecir el rendimiento de la epoxidación catalítica del ciclohexeno utilizando la información de los descriptores almacenados en el conjunto de datos bajo las condiciones de reacción óptimas.
- Comparar los rendimientos de reacción obtenidos i) por el método de *Machine Learning* contra ii) el experimental, usando métricas de error.

4 METODOLOGÍA

El desarrollo completo de la presente investigación se divide en tres fases de experimentación como tal, precedidas por un prólogo —consulta y actualización del estado del arte—, y así mismo, un epílogo —consistente en la escritura del presente documento—; las tres fases (Fases 1, 2 y 3) recogen el conjunto de pasos que se siguieron con el fin de cumplir con los tres objetivos específicos del proyecto.

Cada una de las fases, se desarrolla con base en la metodología de Patrón de Investigación Iterativo de Pratt [60], la cual consiste, de forma general, en el cumplimiento de cuatro etapas: observación, identificación, desarrollo y pruebas. Cada uno de los parámetros anteriores debe ser primeramente definido dentro de cada fase, y posteriormente deben ser completados con éxito para determinar así la finalización de una fase individual.

El desarrollo y finalización de cada una de las fases 1-4 estará determinado por el resultado que se pueda obtener en cada una de ellas; esto conduce pues, a que cada una de las fases 1-3 se vea asociada a cada uno de los objetivos específicos, y la fase 4 represente el origen del producto final de la investigación.

4.1 Prólogo

En esta sección se mantuvo una visión actualizada de las investigaciones concernientes a la búsqueda bibliográfica necesaria para la construcción del proyecto en sí mismo, además de aquellas concernientes al estado del arte de la "catálisis digital" en todo momento, desde el comienzo de la investigación hasta la redacción del documento final de trabajo de grado.

4.2 Fase 1

El resultado principal que se obtuvo al completar esta fase fue un conjunto de datos experimentales obtenidos del análisis del estado del arte referente a la epoxidación del ciclohexeno con descriptores necesarios para el entrenamiento de un modelo de *Machine Learning*. Con el fin de completar esta fase se han identificado las cuatro etapas mencionadas por Pratt [60] como sigue:

- **Observación:** búsqueda en bases bibliográficas de distintas investigaciones ya publicadas en donde se lleve a cabo la epoxidación catalítica del ciclohexeno, siguiendo la metodología planteada por Petersen et al. [61] para la ejecución de un mapeo sistemático. El método general descrito por Petersen et al. [61] y adaptado a una investigación en particular (ingeniería de software), indica comenzar el mapeo sistemático por el planteamiento de algunas preguntas de investigación claras, las cuales serán objeto de respuesta a lo largo del mapeo, y que en principio ayudarán a enfocar la búsqueda bibliográfica. Las preguntas de investigación Q_n planteadas en este caso son:
    i. **Q_1:** ¿Hay un número de publicaciones suficiente, sobre la epoxidación catalítica del ciclohexeno, como para generar un conjunto de datos apto para la alimentación de un modelo de *Machine Learning*?
    ii. **Q_2:** ¿Poseen las publicaciones encontradas el enfoque investigativo adecuado para ser parte de la selección final inmediatamente anterior a la construcción del conjunto de datos?

iii. **Q_3:** A parte de poseer el enfoque investigativo adecuado, ¿poseen las publicaciones los datos mínimamente necesarios referentes a la investigación de la reacción catalítica de interés como para ser parte de la selección final?

El mapeo sistemático debe tener como objetivo primario el responder a estar tres preguntas planteadas, y las respuestas ayudan a enfocar la búsqueda bibliográfica hacia el tipo de datos que se quieren extraer de la literatura.

Una vez identificadas las preguntas de investigación, es preciso iniciar la sección de búsqueda como tal, para lo cual fue necesario definir entonces los "tipos" de palabras clave (*keywords*). Esto se puede lograr mediante la aplicación de una estrategia en particular para este estudio (tomada y adaptada de [61]) denominada PICO (de sus siglas en inglés: *Population*, *Intervention*, *Comparison and Outcomes*). Esta estrategia se aplica sobre los componentes del tema principal de la búsqueda bibliográfica, con el objetivo de facilitar la identificación de grupos de palabras clave.

Al definir entonces el objetivo de la búsqueda bibliográfica como la obtención de los datos mínimamente necesarios para la construcción de un conjunto de datos, relacionados a las características propias de un catalizador y condiciones de un sistema de reacción para la epoxidación del ciclohexeno, se establecieron entonces las definiciones para cada parámetro del PICO dentro del tema de búsqueda.

i. **Población:** referido al objeto de estudio, o concepto sobre el cual se lleva a cabo una eventual intervención, se ha definido la población en este caso como el sustrato que será objeto de su transformación catalítica a su correspondiente epóxido, siendo este sustrato el ciclohexeno.
ii. **Intervención:** definida como la forma en la que se incide el cambio deseado en la población, se toma en este caso como intervención el conjunto de condiciones, tanto propias del catalizador como de la reacción en sí, que se utilizaron en cada una de las investigaciones publicadas para la epoxidación catalítica del ciclohexeno.
iii. **Comparación:** siendo que la comparación se establece como la medición de la intervención elegida con otra forma de intervención, se compara entonces entre las distintas condiciones de reacción catalíticas la incidencia que pueden tener en el resultado el variar distintos parámetros de la reacción o el catalizador.
iv. **Resultados:** para la definición de este parámetro dentro de los límites de la búsqueda bibliográfica se tomaron entonces los resultados en rendimiento de reacción de todas las formas de intervención, es decir, de todos los conjuntos únicos de condiciones de reacción.

Una vez definidos los parámetros anteriores, la identificación de grupos de palabras clave se vuelve más precisa, facilitando el enfoque de estos grupos hacia las categorías de población, intervención y comparación. Sin embargo, debido a la especificidad y singularidad de la categoría "resultados" para los propósitos de esta investigación, no se consideró práctico incluir un grupo de palabras clave basado en este parámetro.

Por lo tanto, se establecieron cuatro grupos de palabras clave, inspirados en la estrategia PICO, pero no limitados estrictamente a cada uno de sus componentes.

i. Derivado del parámetro “intervención”, se enfoca en el uso específico de un metal M en la reacción, con la estructura general de “*use of M*”, donde M puede ser Pd, Mo o Ru. Este enfoque se debe a que, en un catalizador, el metal suele ser el componente que proporciona la actividad catalítica.
ii. También derivado de la “intervención”, este segundo grupo aclara el uso de la catálisis en la reacción de epoxidación del ciclohexeno, abarcando tanto la catálisis homogénea como la heterogénea.
iii. Basado en la “población”, el tercer grupo orienta la búsqueda hacia el tipo específico de reacción del sustrato, en este caso, la epoxidación del ciclohexeno. Este grupo posee una estructura más abierta pero siempre dirigida hacia el objetivo mencionado.
iv. No se centra directamente en ninguno de los parámetros de PICO. Este grupo incorpora elementos de ciencias de datos y computación aplicadas a la epoxidación catalítica de alquenos. El propósito es explorar antecedentes del uso de metodologías informáticas en el campo de las ciencias catalíticas.

Tras definir estos grupos de palabras clave, se empleó una matriz de Excel proporcionada por Elsevier. Esta herramienta facilita la construcción y organización de las cadenas de búsqueda, que posteriormente se introducirán en bases de datos para llevar a cabo la búsqueda bibliográfica.

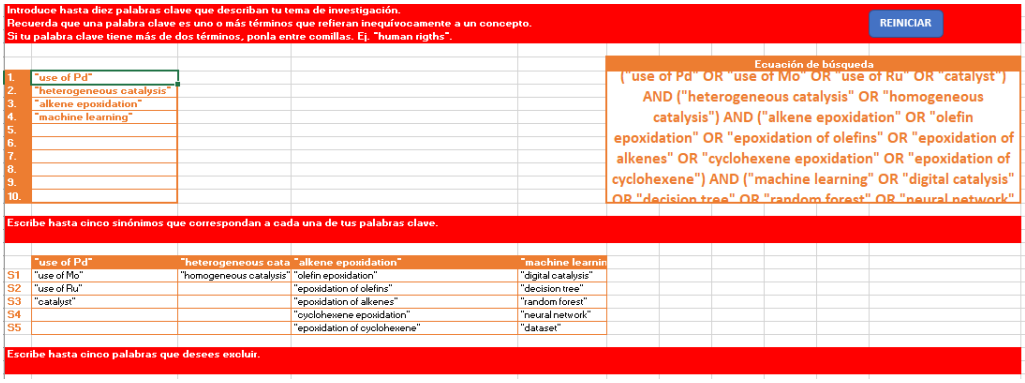


Figura 2. Resultado preliminar de la generación de grupos de palabras clave y una cadena de búsqueda base en la matriz utilizada, provista por Elsevier.

Tras definir las palabras clave "cabeza de grupo", se establecen sinónimos para cada una de ellas, formando así lo que se denomina un grupo. En el contexto de esta investigación, un grupo de palabras clave se define como una sección de la cadena de búsqueda, donde los términos están interconectados por el conector "*OR*". Esta sección se distingue de otras secciones similares (otros grupos) mediante el uso del conector "*AND*". Los sinónimos son términos que apuntan al mismo objetivo que la palabra clave principal. Por ejemplo, la cabeza de grupo 1, “*use of Pd*”, elegida como representante de los tres metales considerados, se refiere al uso del Pd como centro activo en catalizadores empleados en la epoxidación del ciclohexeno. Esta palabra clave tiene tres sinónimos: dos con la misma estructura base (“*use of Mo*” y “*use of Ru*”), y un tercero (“*catalyst*”) que se refiere

al uso de catalizadores en general para la reacción. Estas cuatro palabras clave tienen el objetivo común de restringir la búsqueda a investigaciones que utilicen un catalizador en la reacción, ya sea de manera específica (uso de un metal M) o de forma general.

Así, se han definido las palabras clave "cabeza de grupo" y sus sinónimos, siguiendo la estructura de la matriz utilizada. Como se observa en la **Figura 2**, además de la posibilidad de incluir sinónimos, existe la opción de definir términos de búsqueda que se desean excluir en cada uno de los grupos. Estos términos estarían separados del resto del grupo por el conector "*NOT*". Sin embargo, en este caso particular, no se excluyó ningún término de búsqueda.

Una vez establecidos todos los términos de búsqueda, incluyendo las cabezas de grupo y sus sinónimos, se genera la cadena de búsqueda bruta mediante la concatenación correspondiente de todos los términos, un proceso que ya está automatizado en la matriz.

En la **Tabla 2** se muestran los grupos de palabras clave ingresados a la matriz de Excel.

Tabla 2. Grupos de palabras clave obtenidos.

<table>
<tr><th>Grupo 1</th><th>Grupo 2</th><th>Grupo 3</th><th>Grupo 4</th></tr>
<tr><td>Referente al catalizador</td><td>Referente al catalizador</td><td>Referido al sustrato</td><td>Referido a la incorporación de la IA en catálisis (removido en búsqueda para extraer datos)</td></tr>
<tr><td>“use of Pd”</td><td rowspan="4">“heterogeneous catalysis”</td><td>“alkene epoxidation”</td><td>“machine learning”</td></tr>
<tr><td>“use of Mo”</td><td>“olefin epoxidation”</td><td>“digital catalysis”</td></tr>
<tr><td rowspan="2">“use of Ru”</td><td>“epoxidation of olefins”</td><td>“decision tree”</td></tr>
<tr><td>“epoxidation of alkenes”</td><td>“random forest”</td></tr>
<tr><td rowspan="2">“catalyst”</td><td rowspan="2">“homogeneous catalysis”</td><td>“cyclohexene epoxidation”</td><td>“neural network”</td></tr>
<tr><td>“epoxidation of cyclohexene”</td><td>“dataset”</td></tr>
</table>

La cadena de búsqueda base, denominada "cadena 01" (abreviada como "cad01"), fue inicialmente introducida en el editor de texto Sublime. Esto se hizo con el fin de facilitar posibles modificaciones sobre la cadena original. Esta cadena base se empleó para realizar la búsqueda bibliográfica en la base de datos proporcionada por la Universidad del Cauca, "Scopus". La consistencia de esta cadena de búsqueda "progenitora" sugiere que los resultados obtenidos están directamente relacionados con la producción de trabajos en el campo de la "catálisis digital en la epoxidación del ciclohexeno". Esto implica que la cantidad de trabajos identificados se considera antecedente relevante para la investigación en curso. Con la cadena "cad01", se identificaron 3 publicaciones que contienen

información potencialmente útil para reforzar el entendimiento de los fundamentos y el estado del arte del presente proyecto. Sin embargo, estas publicaciones no registran estudios previos con un enfoque similar al de la investigación actual.

La modificación manual de la cadena de búsqueda base resultó en la creación de cuatro cadenas de búsqueda distintas. Los resultados obtenidos en Scopus a partir de estas cadenas modificadas se presentan en la **Tabla 3**.

Tabla 3. Cadenas utilizadas en la búsqueda bibliográfica de referencia para la construcción del conjunto de datos. La cadena "cad01" es la cadena de búsqueda de base y las cadenas "cad02" a "cad04" son los resultados de la modificación manual de "cad01".

Cadena de búsqueda	Nomenclatura	Número de resultados*
("use of Pd" OR "use of Mo" OR "use of Ru" OR "catalyst") AND ("heterogeneous catalysis" OR "homogeneous catalysis") AND ("alkene epoxidation" OR "olefin epoxidation" OR "epoxidation of olefins" OR "epoxidation of alkenes" OR "cyclohexene epoxidation" OR "epoxidation of cyclohexene") AND ("machine learning" OR "digital catalysis" OR "decision tree" OR "random forest" OR "neural network")	cad01	3
("use of Pd" OR "use of Mo" OR "use of Ru" OR "catalyst") AND ("heterogeneous catalysis" OR "homogeneous catalysis") AND ("alkene epoxidation" OR "olefin epoxidation" OR "epoxidation of olefins" OR "epoxidation of alkenes" OR "cyclohexene epoxidation" OR "epoxidation of cyclohexene")	cad02	219
("use of Pd" OR "use of Mo" OR "use of Ru" OR "catalyst" OR "heterogeneous catalysis" OR "homogeneous catalysis") AND ("alkene epoxidation" OR "olefin epoxidation" OR "epoxidation of olefins" OR "epoxidation of alkenes" OR "cyclohexene epoxidation" OR "epoxidation of cyclohexene") AND ("machine learning" OR "digital catalysis" OR "decision tree" OR "random forest" OR "neural network")	cad03	6

("use of Pd" OR "use of Mo" OR "use of Ru" OR "catalyst" OR "heterogeneous catalysis" OR "homogeneous catalysis") AND ("alkene epoxidation" OR "olefin epoxidation" OR "epoxidation of olefins" OR "epoxidation of alkenes" OR "cyclohexene epoxidation" OR "epoxidation of cyclohexene")	cad04	2640

*Búsqueda realizada el 27 de julio del 2022.

Las cadenas "cad01" y "cad03" son similares, diferenciándose principalmente en que para "cad03" se eliminó el grupo de palabras clave número 2 (como se muestra en la matriz de Excel en la **Figura 2**) y se integró en un único grupo junto con las palabras clave del grupo 1. Esta modificación se debió a la similitud en las características definidas por ambos grupos, ya que todos los términos se relacionan con la catálisis. Lógicamente, se esperaría que "cad03" arroje más resultados que "cad01", dado que "cad01" está contenida dentro de "cad03". Esta inferencia se puede verificar al examinar los títulos de los resultados obtenidos con ambas cadenas de búsqueda (ver el **Anexo 1** "Matriz para obtención de cadena de búsqueda.xlsx").

De forma similar, las cadenas "cad02" y "cad04" se parecen entre sí, siguiendo la misma lógica de diferenciación que "cad01" y "cad03". Sin embargo, estas dos cadenas se diferencian del primer par en la ausencia del grupo 4 de palabras clave (ver **Figura 2**), que se enfoca en la "utilización de técnicas de las ciencias de la computación en el desarrollo de catalizadores". Al excluir este grupo, la búsqueda se centra en estudios sobre la epoxidación catalítica del ciclohexeno de forma "clásica". Esta estrategia permite encontrar un mayor número de publicaciones, que tras definir ciertos parámetros de inclusión/exclusión, servirán como base para la construcción del conjunto de datos de entrenamiento. Así como los resultados de "cad01" están contenidos en "cad03", los de "cad02" se encuentran dentro de "cad04". Considerando la limitación de tiempo en este tipo de investigaciones, se decide analizar inicialmente solo los resultados obtenidos con "cad02".

En la interfaz de Scopus, los resultados se exportan como un archivo ".csv", que posteriormente se convierte en una hoja de cálculo de Excel (".xlsx"). Una vez listo el archivo, se realiza una clasificación preliminar (arbitraria, elegida a conveniencia) de las publicaciones, asignándoles codificadamente valores de -1, 0, o 1, basados en el análisis de sus títulos y resúmenes. Publicaciones con -1 son aquellas no relevantes para la investigación; las asignadas con 0 requieren una lectura más detallada para decidir su relevancia; y las marcadas con 1 son inmediatamente pertinentes para la construcción del conjunto de datos.

La "información" que determina si una publicación recibe -1, 0 o 1, es su mención o implicación en la realización de la epoxidación del ciclohexeno. Tras dos revisiones consecutivas, solo quedaron publicaciones clasificadas como -1 o 1, puesto que aquellas clasificadas como 0 (provisionalmente) fueron revisadas y clasificadas como 1 (aprobada) o -1 (descartada).

Finalmente, durante la revisión de texto completo y con el objetivo de hacer una última selección de calidad, se han establecido parámetros de exclusión basados en la experiencia adquirida ante la lectura de los documentos encontrados. Los parámetros de exclusión determinados se enlistan a continuación:

- ✓ Todo artículo en idiomas distintos al español, inglés y francés se excluye, independientemente de si clasifica o no la revisión de calidad inicial. Subrayados en color negro.
- ✓ Poca o nula claridad ante la exposición de la información concerniente a la reacción catalítica de interés o ausencia de esta. Subrayados en color amarillo.
- ✓ Trabajos en los cuales las variables respuesta, es decir, aquellas relacionadas al rendimiento de la reacción (TON, TOF, % conversión, % selectividad) no estén en función de un único producto de interés: el epóxido del ciclohexeno. Subrayados en color azul.
- ✓ Enfoque distinto a la actividad catalítica y a la eficiencia de la reacción. Subrayados en color marrón.
- ✓ "Literatura gris", por ejemplo, libros. Subrayados en color gris.

Además, como una estrategia simbólica (no como un parámetro de exclusión), se han marcado en verde oscuro aquellas publicaciones que podrían tener relevancia sustancial debido a la introducción de las Ciencias Computacionales en estudios de epoxidación catalítica, incluso si no se centran en el ciclohexeno como sustrato (ver **Anexo 2**: "cad02-Eliminación")

Es importante reconocer que algunas publicaciones relevantes para el mapeo sistemático podrían no haber sido capturadas por la cadena de búsqueda. Para abordar esta posibilidad, se adopta la estrategia de "bola de nieve", que implica revisar las referencias de las publicaciones seleccionadas y someterlas a un breve análisis de calidad para, si es necesario, incluirlas en los resultados finales.

- **Identificación:** se definen las distintas características, propiedades fisicoquímicas, propiedades catalíticas y condiciones de reacción requeridas para la conformación de un conjunto de datos que permita la aplicación de un modelo de ML para obtener una predicción del rendimiento de la reacción catalítica a partir de algunos valores de partida para cada descriptor. Para ello, se han tratado de definir descriptores cuya identificación cualitativa y cuantitativa no implique el uso de tratamientos computacionales previos (como, por ejemplo, el empleo de métodos computacionales cuánticos); esto partiendo de la intención de construir un conjunto de datos simple pero preciso, que sea capaz de dar resultados hipotéticamente aceptables sin requerir de un mayor gasto de recursos. Aclarado lo anterior, los descriptores que se han identificado, así como los rangos entre los que se definen sus posibles valores a tomar, se enlistan en la **Tabla 4**.

Tabla 4. Descriptores planteados inicialmente y su definición.

Descriptor	Definición
Metal	Cuyos posibles valores se definen entre el aluminio (Al, número atómico 13) y el copernicio (Cn, número atómico 112).
Tipo de catalizador	Que se define solamente entre dos posibles valores, homogéneo y heterogéneo.
Estado de oxidación	Valor discreto que varía entre 0 y 7+ para los propósitos de esta investigación.
Configuración electrónica externa	Valor alfanumérico discreto que varía entre d^0 y d^{10}.
Geometría del centro metálico	Sus posibles valores son: lineal, trigonal planar, tetraédrica, cuadrada planar, bipiramidal trigonal, pirámide cuadrada y octaédrica.
Ligando/soporte	Valor no definido que describe brevemente el tipo de ligando o soporte de cada catalizador.
Clase de ligando/soporte	Indica la coordinación en el caso del ligando, y varía entre mono-hexadentado; en el caso de los soportes, indica si el catalizador es másico o soportado.
Oxidante	Indica el agente oxidante usado.
Clase de oxidante	Indica el tipo de oxidante dependiendo de su origen, definiéndose entre orgánico e inorgánico.
Solvente	Especifica el solvente usado.
Clase de solvente	Indica el tipo de solvente, y se define entre polar y apolar.
Temperatura	Valor continuo, definido entre 0 y 200°C debido a los resultados de las publicaciones base.
Tiempo de reacción	Valor continuo que se define entre 0 y 24 horas.
Porcentaje de catalizador en mol	Se define como un valor continuo entre 0 y 100 %, y sus unidades son % mol de catalizador (centro metálico activo) respecto al sustrato (ciclohexeno).
Porcentaje de conversión	Valor continuo entre 0 y 100 % y se basa en mol de sustrato transformado (consumido).
Porcentaje de selectividad	Valor continuo entre 0 y 100 % y se basa en la cantidad del epóxido del ciclohexeno obtenido por mol de sustrato.

Porcentaje de rendimiento	Valor continuo, definido entre 0 y 100 % y se refiere al rendimiento global de la reacción (mol producto/mol sustrato), y para efectos prácticos equivale, en casos en los que no se reporte, al producto entre % conversión y % de selectividad sobre 100 (%C*%S/100).
TON	Valor continuo que virtualmente puede variar entre 0 e infinito.
TOF	Valor continuo que virtualmente puede variar entre 0 e infinito y cuyas unidades son h^{-1}.

- **Desarrollo:** Se recopilan los descriptores definidos anteriormente en un conjunto de datos construido a partir de los resultados obtenidos en las publicaciones seleccionadas mediante la búsqueda bibliográfica con "cad02". Este conjunto se organiza de tal manera que los descriptores se ubican en las columnas, mientras que en las filas se sitúan las entradas. Cada "entrada" corresponde a un "conjunto de condiciones extraído de una publicación definida". Es importante señalar que algunas publicaciones contienen más de un conjunto de datos extraíble, por lo que no siempre es correcto asumir que una "entrada" equivale a una "publicación".

  La asignación de valores a cada descriptor se realizó mediante la lectura completa de los documentos, enfocándose especialmente en la inclusión del ciclohexeno como sustrato de estudio. Este análisis detallado también permitió asignar una clasificación definitiva (-1, 1) a todas aquellas publicaciones inicialmente marcadas con 0.

  Tras la lectura de los documentos y la asignación de la clasificación final, se extrajeron los datos de aquellos trabajos que superaron satisfactoriamente todos los filtros previos (incluyendo los parámetros de exclusión mencionados en la sección de observación). El resultado es un conjunto de datos bruto que luego se somete a un proceso de pretratamiento para "pulirlo" y prepararlo para su utilización en el modelo de ML.
- **Pruebas:** en esta parte se “limpia” el conjunto de datos obtenido en la sección de desarrollo, de forma que los datos contenidos en él cumplan con las condiciones requeridas para su uso como base en la construcción del modelo algoritmo predictivo. Cuando se termina de recopilar los datos en el conjunto, se nota que quedan algunas entradas irregulares como, por ejemplo, entradas en las cuales el valor de al menos uno de los descriptores (columnas) no se especifica ni se puede inferir a partir de la información contenida en cada publicación. Entradas con este problema y otras formas en las cuales una entrada no puede ser válida para su uso se especifican en la sección de resultados, así mismo como el resultado de la limpieza del modelo. El resultado de esta limpieza provee un conjunto de datos listo para utilizar en el armamento del algoritmo predictivo de ML.

4.3 Fase 2

El objetivo principal de esta fase es desarrollar un algoritmo predictivo que permita estimar un rendimiento teórico. Este cálculo no se basa en inferencia química, sino en la capacidad de aprendizaje

alcanzada por la IA utilizando los datos proporcionados. Para lograr esto, se emplea el conjunto de datos obtenido en la Fase 1. A pesar de que este conjunto se supone listo para usar en la construcción del algoritmo predictivo, puede requerir ajustes y procesos de "limpieza" adicionales para adaptarlo a los requisitos de la ciencia de datos. A continuación, se describen los cambios realizados al conjunto de datos de la Fase 1 para su uso en el desarrollo del modelo de ML.

Inicialmente, se crea un cuaderno en Google Colab. Lo primero que se hace es importar las librerías necesarias para ML, incluyendo aquellas específicas para *Deep Learning* (DL) y otras esenciales para la presentación de métricas de error, que son clave para evaluar la precisión del modelo.

Una vez importadas las librerías, se carga el conjunto de datos de la Fase 1 desde Google Drive. Tras la carga del conjunto de datos (a menudo referido como dataframe en el contexto del cuaderno), se realizan operaciones comunes de limpieza y organización. Esto incluye contar las entradas y separar las variables numéricas de las no numéricas. Por tanto, se escribe código para: primero, contar las clases de descriptores y clasificar cada descriptor como "objeto" o "número", es decir, distinguir entre variables **categóricas** y **no categóricas**. Posteriormente, se genera un código para visualizar el conteo de variables, mostrando parámetros como el número de entradas, el número de valores únicos de cada descriptor, la moda (para variables categóricas) o la desviación estándar (para variables no categóricas).

Una vez verificada la correcta distribución de los datos en el *dataframe*, se procede a modificar su estructura en función de las variables categóricas y no categóricas. Las variables categóricas, como aquellas que clasifican los metales con símbolos de la tabla periódica (de Al a Cn, por ejemplo), se consideran del tipo "objeto" y no son directamente reconocibles por el modelo de ML. Esto representa un desafío, ya que, para los fines del ML, el conjunto de datos debe ser puramente numérico. Para resolver esto, se implementa un código que **codifica el *dataset***, transformando cada variable categórica en numérica asignando un código único a cada valor como, por ejemplo, 11 para Mo. Asimismo, cada variable numérica del tipo entero (**int**) se convierte en flotante (**float**).

Con el *dataframe* ahora completamente numérico y compuesto por valores flotantes en cada descriptor, surge el desafío de las grandes variaciones entre columnas. Por ello, es necesario **normalizar los datos codificados**. La **normalización** es un proceso que transforma los datos para que se ajusten a una escala común, reduciendo así la probabilidad de error. Los beneficios de normalizar los datos incluyen mayor precisión, rendimiento mejorado en algoritmos que requieren escalas similares para el procesamiento de datos, y reducción del ruido causado por valores atípicos [62]. Existen varios métodos matemáticos para normalizar datos; en este caso, se elige la **normalización basada en mínimos y máximos**, que sitúa todos los datos de cada descriptor en una escala conveniente dentro del intervalo [0,1]. La fórmula para normalizar los datos se presenta en la Ecuación 1 [63]:

$$z = \frac{x' - min\ (x)}{máx(x) - min\ (x)}\ Ecuación\ 1$$

En donde x' representa el dato a transformar, mientras que min(x) y máx(x) representan los datos mínimo y máximo en cada columna, respectivamente.

Una vez normalizado el *dataset*, se procede a revisar su estructura. Tras confirmar que todo está correctamente normalizado, se divide el conjunto de datos en dos partes: **entrenamiento (80%)** y **prueba (20%)**. Esta división implica que, según la proporción establecida, los datos de prueba se usarán una vez que el modelo esté entrenado, comparando la variable de respuesta predicha por el modelo en los datos de prueba, como si estos fueran datos de una entrada desconocida, con el valor conocido de la variable de respuesta en el conjunto de datos. De esta manera, se obtiene una medida de la precisión del modelo, comparando entre lo que debería predecir y lo que realmente predice, utilizando datos extraídos del mismo conjunto y no de otra fuente.
Habiendo realizado este proceso previo de organización de datos, se procede a seguir cada una de las etapas definidas por Pratt en su metodología de conducción de investigaciones [60].

- **Observación:** en esta etapa se realiza la recopilación de algoritmos de regresión que han de ser la "biblioteca" de la que se deben elegir aquellos a implementar. Se deben elegir algoritmos especializados en la tarea de regresión ya que lo que se pretende es predecir el valor de una variable denominada "de respuesta" a partir de los valores de entrada de la reacción.
- **Identificación:** selección de algoritmos de regresión para su implementación en el modelo predictivo. Los algoritmos de regresión elegidos, de forma arbitraria y basada en aquellos más comunes usados para propósitos similares, fueron:
    i. *Decision Tree Regressor*
    ii. *Linear Regression*
    iii. *LightGBM Regressor*
    iv. *Stochastic Gradient Descent Regressor*
    v. *Kernel Ridge*
    vi. *Linear Elastic Net*
    vii. *Bayesian Ridge*
    viii. *Gradient Boosting Regressor*
    ix. *Support Vector Regressor*

- **Desarrollo:** implementación de al menos tres algoritmos de aprendizaje máquina para tareas de regresión utilizando el lenguaje de programación Python y los *frameworks* Tensorflow y ScikitLearn. Los algoritmos que rigen los modelos de regresión explicados en el ítem anterior fueron aplicados, uno por uno, mediante su importación al cuadernillo de Google Colaboratory en donde se escribió el código base para el entrenamiento. La importación y ejecución de cada uno de estos modelos es un proceso automatizado por la escritura de código y la observación se remite a los resultados en términos de las métricas de error utilizadas.

- **Pruebas:** ejecución de los algoritmos de aprendizaje máquina con el objetivo de obtener las métricas de validación de los procesos de predicción del rendimiento teniendo como referencia el error cuadrático medio y el coeficiente de correlación.
  A parte de la ejecución de algoritmos para modelos de regresión ya nombrados, y como aplicación adicional a nivel más profundo, se utilizó Aprendizaje Profundo (DL, por sus siglas

en inglés) para el entrenamiento con los datos del *dataset* y cuyo fundamento ya se especificó de forma sencilla en la sección de Fundamento teórico.

El procedimiento de aplicación de DL fue similar al de la aplicación de modelos de ML y partió de la importación de librerías, codificación y normalización del conjunto de datos, todo similar a la primera vez que se realizó. Ya en el entrenamiento, los fundamentos son algo distintos, por cómo se puede suponer de la sección de Fundamento teórico, en donde se describe el funcionamiento del DL y las ANN. La aplicación de DL debe comenzar con la construcción de una ANN que se ha de entrenar alrededor de los datos normalizados, para lo cual había que dispones de **capas** con un número definido de **neuronas** y establecer un **optimizador** y una **función de pérdida** que permitieran obtener el mejor resultado de capacidad de predicción. Las capas y neuronas ya están definidas, así que se hará la aclaración de lo que es un optimizador y una función de pérdida, que ya tuvo una mención anteriormente.

Un **optimizador** es un algoritmo que usa la red neuronal para modificar atributos como el sesgo y la interconexión de neuronas, cambiando así los parámetros del modelo entrenado y mejorando con cada repetición. Básicamente el optimizador es la forma en la que la ANN se corrige y mejora a sí misma, minimizando la función de pérdida [18].

Una **función de pérdida** es una función que mide, en promedio, la diferencia entre los datos de salida del modelo entrenado y los datos reales, esto respecto a la porción de validación o prueba del *dataset*. Entre menor sea este valor se dice que el modelo es mejor [18].

Estos dos conceptos son importantes porque hacen parte de la arquitectura de la ANN; sin embargo, su fundamento y funcionamiento requieren una base matemática y teórica superior y no se considera necesaria su explicación a profundidad.

Una vez establecida la arquitectura de la ANN se ejecutó en la plataforma Google Colab y se determinó la estructura de mejor rendimiento predictivo.

4.4 Fase 3

En esta fase se compararon los rendimientos i) predicho por el modelo de ML y ii) calculado mediante el seguimiento de la reacción realizada en el laboratorio.

- **Observación:** se eligieron los valores de entrada de los 14 descriptores establecidos mostrados en la **Tabla 5** de la sección de Resultados y Análisis. Las especificaciones sobre los mismos se encuentran en dicha sección.
- **Identificación:** los valores de la **Tabla 5** se sometieron al mismo proceso de codificación y normalización que al *dataset* con la misma escritura de código y algoritmos usados en el cuadernillo de Google Colab, esto con el fin de que el modelo procese los datos de entrada bajo la misma base que los del conjunto. Una vez normalizados los datos de entrada se ejecutó en el modelo la orden de procesamiento y predicción del rendimiento.
- **Desarrollo:** luego de la introducción de los valores de entrada de cada descriptor al modelo y la obtención posterior de un rendimiento predicho, se utilizan dichos valores de entrada como condiciones para la realización de la reacción catalítica de epoxidación del ciclohexeno.
- **Pruebas:** ejecución de la realización en el laboratorio y comparación con el porcentaje brindado por el modelo entrenado.

Materiales y reactivos:

Se utilizaron los reactivos necesarios para la reacción catalítica: ciclohexeno estabilizado para síntesis Merck 99%, n-dodecano para síntesis Merck 99% y tolueno Fisher Chemical 99%, todos del grupo de investigación Catálisis.

Se utilizó bisulfito de sodio sólido del Departamento de Química, Universidad del Cauca, área de docencia.

Se utilizaron el tert-butil hidroperóxido (TBHP) Alfa Aesar solución acuosa 70% y el principal catalizador usado en la reacción codificado como MoO_2Cl_2@COMOC-4, cuya estructura unitaria o monomérica se muestra en la **Figura 3**, y provenientes del Departamento de Química de la Universidad Nacional de Colombia, sede Bogotá.

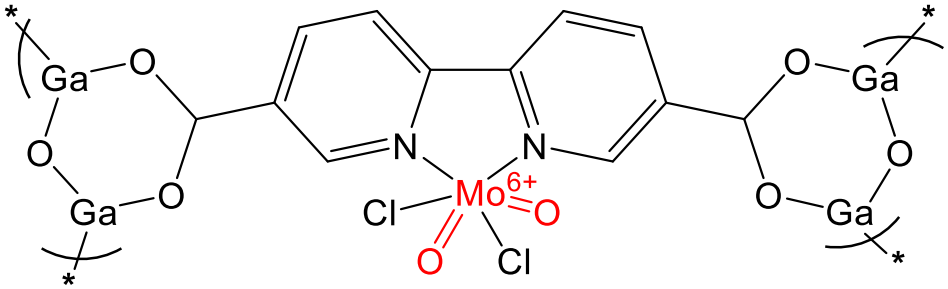


Figura 3. Estructura unitaria del catalizador tipo MOF, MoO_2Cl_2@COMOC-4 [64].

Se usaron los equipos: balanza analítica Precisa ZB220A; centrífuga HERMLE Z 206 A con una máxima velocidad de rotación de 6000 rpm; cromatógrafo de gases Shimadzu GC-14A, detector FID, columna ZB-WAX 100% polietilenglicol, del grupo de investigación Catálisis.

Métodos:

El montaje de reacción se ilustra en la **Figura 4**.

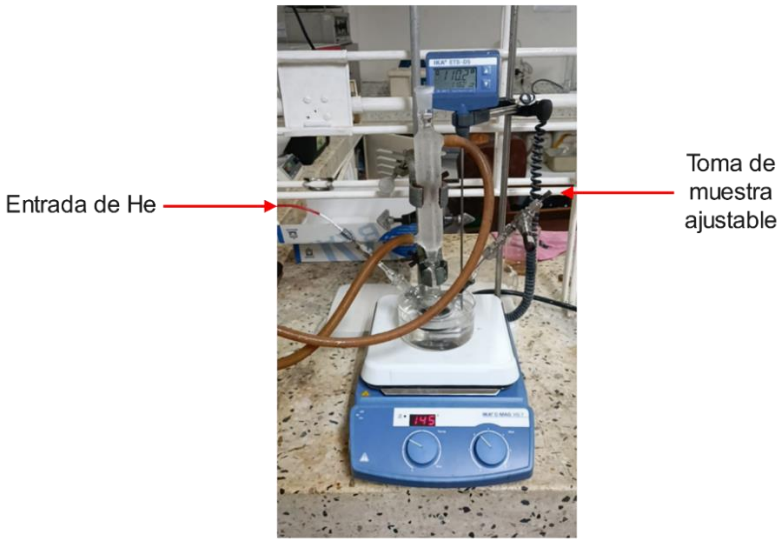


Figura 4. Montaje de reacción utilizado para la epoxidación catalítica de ciclohexeno.

La configuración del montaje se detalla en la **Figura 4**, este consiste en un reactor tipo semi-batch de tres bocas, sumergido en un baño de aceite mineral. En la boca central se conecta un condensador de reflujo. Según se muestra, las bocas laterales se equipan con una entrada de gas inerte y una salida ajustable para muestras. Tanto el aceite del baño como la mezcla de reacción se agitaron de manera continua durante la reacción.

En el reactor se adiciona la mezcla compuesta de la siguiente forma: 4 mmol de ciclohexeno (406 µL), 4 mmol de dodecano (911 µL), 8 mmol de TBHP (1036 µL) y 0,04 mmol del catalizador MoO_2Cl_2@COMOC-4 (1% mol respecto al Mo VI, 66 mg del sólido), todo disuelto en un volumen del 10 mL de tolueno (procedimiento adaptado de [64]).

Inicialmente, se preparó la mezcla con ciclohexeno y dodecano en tolueno, la cual se introdujo al balón de reacción. Una vez estabilizado el sistema, se agregó el TBHP, momento en que se comenzó a registrar el tiempo de reacción (t=0). Se extrajeron alícuotas en intervalos de 0.5, 1.0, 2.0, 3.0 y 4.0 horas. Cada muestra se centrifugó a 4000 rpm durante 5 minutos. Posteriormente, se tomaron aproximadamente 200 µL del sobrenadante y se transfirieron a un vial de separación líquido-líquido (**Figura 5**), al cual se añadieron aproximadamente 200 µL de una solución de $NaHSO_3$ al 15% m/v. Tras una agitación vigorosa y un periodo de reposo, se extrajeron 2 µL de la fase superior para inyectarlos en el cromatógrafo de gases para su respectivo análisis.

Figura 5. Viales de separación utilizados en el procedimiento de análisis por CG-FID.

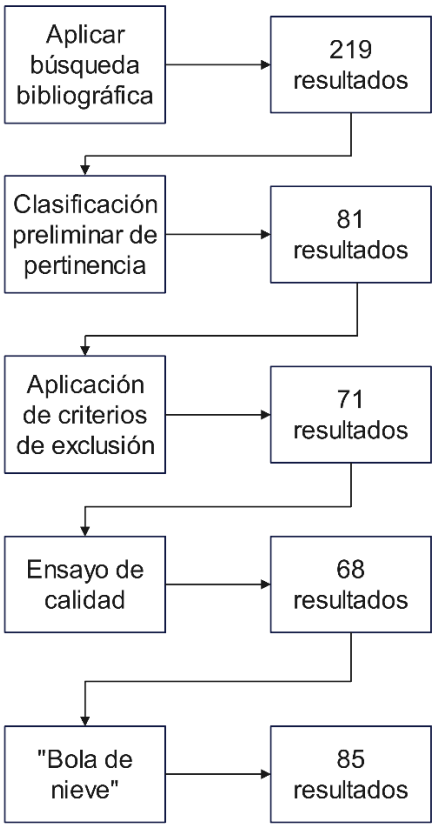
Para el análisis cromatográfico el volumen de inyección fue de 2 µL; las temperaturas del inyector y del detector fueron 80°C y 200°C respectivamente. La temperatura inicial de la columna fue de 60°C la cual se mantuvo por un minuto; posteriormente se aumenta la temperatura de la columna a razón de 10°C/min hasta 140°C y se mantiene en esta temperatura por dos minutos; finalmente, se aumenta de nuevo la temperatura de la columna

a razón de 10°C/min hasta 180°C y se mantiene así por dos minutos, dando un tiempo total de corrida de 17 minutos.

La recolección de datos se hace con ayuda del software *Chromatostation* y el análisis de estos se basó en la respuesta en términos del área de cada señal.

5 RESULTADOS Y ANÁLISIS

En la sección de metodología se presentó el procedimiento seguido para la obtención de publicaciones y el filtrado posterior, todo con el fin general de obtener un conjunto de datos bruto que permitiese entrenar un modelo de ML. La lectura y filtrado de las publicaciones obtenidas se ejecuta como se menciona **Metodología** (sección 3.2), y se resume en el **Esquema 3**.



Esquema 3. Resumen de la revisión y filtrado de las 219 publicaciones obtenidas como resultado de la búsqueda bibliográfica con la cadena “cad02”.

El ensayo de calidad previo a la aplicación de la “bola de nieve” corresponde a lectura de texto completo y evaluación de la calidad general del trabajo, en orden de clasificarlo como apto o no para su inclusión/exclusión en el conjunto de datos bruto. En el **Anexo 3** “Conjunto de datos bruto v1.docx” se muestra la organización de todas las publicaciones que atravesaron los filtros aplicados y que dan origen al conjunto de datos; en él se puede corroborar como los descriptores se encuentran ubicados en columnas, mientras que cada fila corresponde a cada uno de los estudios o entradas. Es necesario

aclarar que cada entrada no corresponde obligatoriamente a una publicación, ya que una sola publicación puede contener más de un estudio catalítico sobre el sustrato deseado, ya que se considera a una entrada un conjunto de descriptores que corresponde a un solo ensayo catalítico.
A pesar de que en el **Anexo 3** aparecen columnas con las variables "TON", "TOF" "% conversión" y "% selectividad", estas se han suprimido en una siguiente versión del conjunto de datos, mostrada en el **Anexo 4** "Conjunto de datos versión final.xlsx", ya que se determinó que la variable respuesta, es decir, la salida del modelo, debe ser solamente una y la característica que mejor representa la eficiencia general de una reacción es el "% rendimiento", por lo cual esta última es la única columna referente a la eficiencia de reacción en esta versión del conjunto de datos en formato Excel. Es necesario convertir al conjunto de datos en un archivo del tipo Excel ya que este es el formato en el cual se ha de introducir (luego, como *dataframe*) a la interfaz de Google Colab para el entrenamiento del modelo.
Una vez el conjunto de datos mostrado en el **Anexo 4** ha sido revisado, se encuentra que algunas entradas aún tenían "datos perdidos", datos que no pueden ser calculados por ningún método al alcance del grupo de trabajo y no pueden ser obtenidos por ningún medio de contacto con los autores, lo que convirtió a estas entradas en descartables. Ante este hecho, la entrada debe ser eliminada por completo, ya que el conjunto de datos debe proporcionar información completa acerca de sus descriptores en cada entrada a fin de que la precisión del modelo, evaluada por las métricas usadas sea confiable y no dependa de "datos perdidos". El resultado de haber limpiado por completo el conjunto de datos del **Anexo 4** es un *dataset* listo para operar, mostrado en el **Anexo 5** "20-07-2023-Dataset.xlsx" y que representa el conjunto de datos final en su versión alfanumérica que será usado para las operaciones computacionales posteriores. En la **Figura 6** se muestra una fracción de este *dataset*, resultado principal de la Fase 1, para una mejor visualización de lo explicado.

Metal	Tipo de catalizado	Estado de oxidació	Configuración electrónic	Geometrí	.igandos/Soport	Clase ligando/soport
Mo	Heterogéneo	6+	d0	N.A.	Mo-UN-1000 (MOF de Mo con Zr/O especies)	Másico

Clase ligando/soport	Oxidante	Clase oxidant	Solvente	Clase solvent	Temperatura (°C	Tiempo (h	% catalizador (%mol	% Rendimient
Másico	H2O2	Inorgánico	Acetonitrilo	Polar	65	5	0.1	1.67

Figura 6. Fracción del dataset "20-07-2023-Dataset.xlsx".

Es evidente que los resultados de una búsqueda bibliográfica extensiva sobre la epoxidación catalítica del ciclohexeno serán mucho más abundantes, en términos numéricos, y las investigaciones relacionadas no son ni siquiera de cerca parecidas en cantidad a las reportadas como resultado de las búsquedas hechas con las cadenas presentadas (particularmente "cad02" y "cad04"); sin embargo, este fue el resultado para la búsqueda hecha en la presente investigación y siguiendo la metodología para un **mapeo sistemático**, el cual tiene como finalidad establecer una sistematicidad en la búsqueda bibliográfica, establecer grupos específicos dentro de la catálisis y reducir la búsqueda a un número razonable de publicaciones, razonable en términos de revisión, filtrado, lectura y extracción de datos, ya que todo el proceso se hizo sin automatización alguna. La ventaja que brinda el mapeo sistemático es que cada vez que se quiera actualizar el conjunto de datos se puede realizar sobre la misma base,

además de que, si se quisiera ampliar más la cantidad de entradas o explorar otras áreas o características de la reacción en sí, se puede iniciar un nuevo mapeo sistemático en lugar de hacer una búsqueda bibliográfica suelta e irregular cada vez. De la misma manera, y con el objetivo principal de buscar razonabilidad en la revisión de las publicaciones, se eligió de base para la construcción del *dataset* ya que si se lee de forma detenida es posible inferir que los resultados de la "cad02" están inmersos en los resultados de la "cad04", además de que los resultados obtenidos con esta última cadena podrían ser demasiados como para la revisión y depurado de parte del personal.

El conjunto de datos obtenido tiene la siguiente estructura general: 201 entradas (filas), 15 columnas de las cuales 14 corresponden a los descriptores y la columna restante corresponde a la variable respuesta (% Rendimiento).

Una vez obtenido el *dataset* limpio, se procedió con la implementación de los resultados en la plataforma Google Colaboratory, herramienta sobre la cual se realizó la sección relativa al entrenamiento del modelo y predicción. La interfaz de esta plataforma se muestra de manera general en la **Figura 7**, en la cual se observan las librerías y herramientas principales que se utilizan para la ejecución de ML en lenguaje Python.

Las **librerías** para ML en lenguaje de Python son conjuntos de funciones y métodos que permiten "sintetizar" operaciones en el único hecho de importarlas [65], [66].

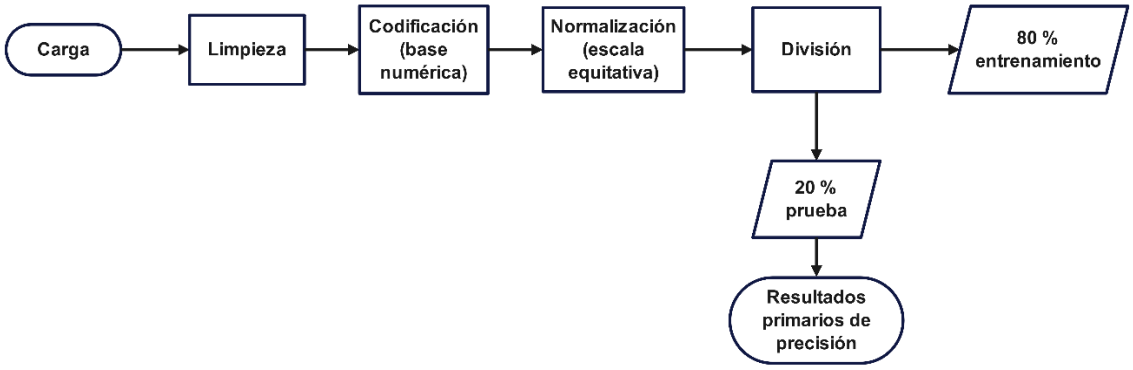
Figura 7. Interfaz de la plataforma Google Colaboratory (Google Colab) que muestra parte inicial del código usado para la construcción del modelo entrenado.

El *dataset* del **Anexo 5** es el producto principal de la Fase 1, resumida en la realización del mapeo sistemático y la obtención del conjunto de datos principal. Como se mencionó en la sección de metodología, es un requerimiento del procesamiento de ML que el conjunto de datos se encuentre en base numérica, por lo cual la forma en la que se muestra en el **Anexo 5** requiere un tratamiento posterior; para poder proseguir, es necesario convertir todas las variables categóricas del *dataset* final en variables numéricas (no categóricas), por lo cual se utilizó el método de **codificación**. La

codificación utilizada para este *dataset* es mediante el método de **codificación ordinal**, la cual asigna valores numéricos enteros a las variables categóricas con base en un orden jerárquico implícito en las mismas [19]. Para ejemplificarlo mejor, se toma el descriptor **configuración electrónica**; este descriptor es una variable categórica que se define entre 11 posibles categorías, desde d^0 hasta d^{10}. El algoritmo de codificación asignará, o al menos intentará asignar, los valores enteros en orden creciente al orden jerárquico de las categorías, esto es, 0 para d^0, 1 para d^1, y así sucesivamente. A pesar de que no a todas las variables categóricas se les puede determinar esta jerarquía el método funciona aparentemente bien para este *dataset* y puede ser debido a la predominancia de variables categóricas jerárquicas presentes en el mismo. El resultado de la codificación es un nuevo *dataset* completamente en base numérica (**Anexo 6**), lo cual en principio cumple el requerimiento inicial; sin embargo, las escalas resultantes no son conmensurables entre sí, y tal como una escala puede variar entre distintos órdenes de magnitud, otra puede variar unas pocas unidades. Para evitar una posible fuente de grandes varianzas en los resultados de precisión del modelo se hace necesario ajustar todos los intervalos de variación de cada descriptor a una escala prudente y única. Esto es lo que se logra con la **normalización**, y como se mencionó en la metodología se utilizó el método basado en mínimos y máximos regido por la Ecuación 1. La normalización por este método implica generalizar las escalas de todos los descriptores (ahora variables numéricas) a un intervalo de 0 a 1, en donde 0 siempre corresponderá al valor menor del intervalo y el 1 corresponde al mayor. Ahora con el *dataset* normalizado se puede iniciar el proceso de alimentación y entrenamiento de los modelos de regresión, por lo cual este se puede denominar entonces *dataframe*.

Con el *dataframe* se debe poder tener una forma de validar la precisión del modelo entrenado (una vez esté listo) que no dependa de incluir más datos externos; para lograr lo anterior se dividió el conjunto en dos secciones: de entrenamiento y de prueba (o validación). La proporción de la división tomada fue de 80% para entrenamiento, es decir, la porción del *dataframe* que ha de servir como alimento y aprendizaje del modelo, y 20% de prueba, con la cual se pretende validar la precisión del modelo ya entrenado. Lo anterior se ejemplifica como sigue: cualquiera que sea el modelo de regresión aplicado, este necesita encontrar patrones y aprender a generar predicciones con base en una variable de respuesta, y para ellos necesita ver el comportamiento de distintos datos ya asociados; para ello es el subconjunto de entrenamiento (80%), que le permite al modelo entrenarse y por ello debe ser la porción mayoritaria, ya que debe tener una cantidad de datos mínimamente confiable como para poder encontrar patrones reales y que permitan tener un modelo preciso. Una vez el modelo ha aprendido cómo funcionan las asociaciones implícitas en los datos suministrados, en este caso el comportamiento del % Rendimiento respecto del ajuste de las condiciones de experimentación (descriptores), es necesario probar qué tan amplia es la capacidad de predecir un resultado; es para ello que se utiliza la porción del subconjunto de validación (20%), con la cual lo que el modelo realiza es, de una forma muy resumida, pretender que ese 20% de entradas no tienen ningún valor de la variable respuesta, es decir, que no existe el valor del % Rendimiento para ellas, y toma cada una de esas entradas como “incógnitas”, luego entonces se da a la tarea de predecir cuales serían los valores del % Rendimiento para ese 20% de entradas y, como en realidad ya se cuenta con ese valor, lo que hace el modelo es comparar intrínsecamente qué tan distantes se encuentran los dos valores de % Rendimiento, el real que está incluido en el *dataset* y el que ha predicho después de ser entrenado

con el subconjunto de entrenamiento. En el **Esquema 4** se muestra un resumen del procedimiento seguido hasta esta sección.



Esquema 4. Resumen de la modificación del *dataset* hasta la obtención del *dataframe* y entrenamiento del modelo. La **limpieza** corresponde a la revisión del *dataset* en busca de datos perdidos o escalas no correspondientes a los datos experimentales.

Antes de proseguir con la discusión de las pruebas y hallazgos de regresión, se han de aclarar las dos métricas utilizadas para medir la precisión del modelo una vez estuvo entrenado. La primera de ellas es el error absoluto medio (MAE, por sus siglas en inglés), y que se define como el promedio de las diferencias absolutas entre los valores predichos y los valores reales de cada registro o entrada. La fórmula con la cual se calcula el MAE se detalla en la Ecuación 2 [18].

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \quad Ecuación\ 2$$

En donde "N" corresponde al total de datos, "y_i" corresponde al valor real, mientras que "$\hat{y}_i$" corresponde al valor predicho por el modelo. Mientras menor sea el valor del MAE este será mejor.
La otra métrica usada es el coeficiente de correlación (R^2) que en palabras menos técnicas, resume el grado de correlación entre las variables dependientes e independientes comparado con el modelo. El valor de R^2 tiene un mínimo de 0,0 y un máximo de 1,0 y cuando mayor sea su valor (siempre menor que el máximo) se dice que el modelo es mejor. El valor de este coeficiente determina qué tan bien puede predecir el valor "y" de una muestra desconocida "x", en este caso, qué tan bien podrá predecir el porcentaje de rendimiento (y) respecto de los valores de los descriptores (x), dependiendo qué tan cercano a 1,0 sea. Este valor puede ser negativo debido a la arbitrariedad de la correlación, y no significa que sea incorrecto y hace que se pueda tomar el valor absoluto de R^2. La Ecuación 3 muestra cómo se calcula este valor [67].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, donde\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad Ecuación\ 3$$

Una vez establecidas las métricas de error que se han de utilizar para medir el alcance de predicción del modelo, se procedió con la explicación de la aplicación de los modelos de regresión implementados y los resultados que arrojaron.

En la sección **3.3.** de metodología se mostraron los prospectos de modelos de regresión a aplicar, y se importaron en el cuadernillo de Google Colab para su ejecución (**Anexo 7**). Dado que la aplicación de estos modelos es dependiente de la IA y el algoritmo del programa en Python y no se hace por medios manuales, no se entrará en detalle más profundo del explicado en la metodología sobre sus principios.

Después del autoajuste y el ajuste manual de los parámetros de los modelos que proveían un mejor resultado de predictibilidad, en busca de mejores resultados de precisión, se encontró que con el **Árbol de Decisión** el valor de R^2 fue de **0,4207** y el error absoluto de **15,99**. Ningún modelo fue capaz de alcanzar estas cifras, ni siquiera cercanas, en términos de predictibilidad. Aunque el objetivo primordial era lograr implementar el uso de herramientas informáticas como el ML en el ámbito catalítico a nivel institucional, y no se fijaron objetivos en términos de eficiencia predictiva —ya que se conocían las posibles limitaciones en torno al sistema catalítico elegido y la profundidad de los descriptores seleccionados— los resultados aparentan ser poco llamativos en términos de aplicabilidad a situaciones reales, tanto en el ámbito académico, como investigativo y, para futuras proyecciones, industrial. En este contexto, los resultados podrían ser explicados por las limitaciones generales a las que están sujetos los modelos utilizados; sin embargo, al no tener conocimiento a profundidad de lo que sucede en el procesamiento de los datos, no es garantía utilizar estas limitaciones como argumento para tratar de explicar las desviaciones.

Figura 8. Aspecto de la interfaz Google Colab cuando se importan los modelos de regresión utilizados.

Cuando se obtiene un estimado de la capacidad de predicción de un modelo de regresión, ya sea alta o baja, es difícil saber con certeza el porqué de este valor, en gran medida porque no se conoce de primera mano una por una las asociaciones que está desglosando el modelo y los patrones que está encontrando para predecir. Los modelos de regresión son susceptibles de varias desviaciones, lo que los expone a limitaciones y, en consecuencia, a imprecisiones o errores. Las limitaciones a las que

puede someterse cada modelo dependen precisamente del fundamento de este, y algunos pueden ser más susceptibles a ciertas limitaciones que otros.

Aunque la normalización de los datos codificados puede ayudar bastante a evitar la disminución considerable en la precisión por efectos de las escalas no unificadas y la mayor varianza en los datos en modelos como *Linear Regression*, *Stochastic Gradient Descent Regressor* o *Support Vector Regression* [68], [69], [70], la alta sensibilidad a las escalas no unificadas o a los valores atípicos no es el único factor que puede afectar a los modelos implementados, e incluso el mejor modelo (*Decision Tree Regressor*) se ve afectado por estas variaciones [71], [72], por lo cual no debería ser esta la causa mayoritaria de los resultados obtenidos. En el ámbito del ML se conocen dos principales inconsistencias referentes a la cantidad de datos y/o a la capacidad de ajuste del modelo: el **sobreajuste (*over-fitting*)** y el **subajuste (*under-fitting*)**. Un modelo cae en el sobreajuste cuando está expuesto a modelar una enorme cantidad de datos, casi que obligando al modelo no a aprender de los datos, sino a memorizarlos, por lo cual la precisión de modelamiento suele caer drásticamente y disminuir por ende la eficiencia representada en las métricas de error [18]. Esto le suele ocurrir a modelos complejos que se ajustan milimétricamente a los patrones provistos por los datos de los cuales se alimentan, como varios de los modelos aquí utilizados [71], [72], [73], [74], [75]; sin embargo, no sólo depende de la complejidad del modelo, sino de la cantidad de datos, y es probable que el aporte combinado de estos dos factores sea lo que podría estar provocando sobreajuste. Sin embargo, puede haber también subajuste, que es el caso contrario, y ocurre mayormente cuando la cantidad de datos no es suficiente para un buen modelado como, por ejemplo, si se tuvieran dos puntos de un modelado de datos exponencial estos podrían verse como una recta y obtener una pésima correlación, que proviene solamente de la poca cantidad de datos [18]. Aunque aparentemente ninguno de los modelos suele ser especialmente sensible al subajuste [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], es una posible respuesta a la baja capacidad de predicción para estos modelos, ya que al ser un procedimiento tan complejo desde el punto de vista molecular, la epoxidación catalítica del ciclohexeno puede tener un número de correlaciones con tendencia al infinito cuando se varíen los descriptores tomados para esta investigación, e incluso es sabido que no son todos los factores de los cuales depende el rendimiento de una reacción catalítica de esta complejidad.

Dada la [relativamente] baja capacidad de predicción del modelo entrenado hasta este punto se optó por aplicar un método de entrenamiento y predicción un poco más profundo y exigente, por llamarlo de cierta forma, siendo este el **Aprendizaje Profundo (*Deep Learning*, DL)**. El DL, como se mencionó, es una rama del ML que requiere una menor intervención humana y es capaz de profundizar más en los patrones que pueda haber en los datos suministrados.

En la experiencia particular de la investigación actual, la mayor intervención humana, y tal vez la más exhaustiva, fue la estructuración de la red neuronal, la cual puede tener tantas variaciones como ideas tenga el programador, así que en este punto la necesidad de mejorar este resultado obligaba a una exploración adicional a la estimada inicialmente; sin embargo, se hizo luego evidente el beneficio. La base de los ensayos con ANN consistía en la adición de capas con un determinado número de neuronas y la variación entre distintas funciones de pérdida y optimizadores a disposición. La **Figura 9** muestra la arquitectura final de la ANN que proporcionó un mejor resultado (**Anexo 8**).

```
[ ]  model = tf.keras.Sequential([
         tf.keras.layers.Dense(1024, activation='relu', input_shape=(x_train.shape[1],)),
         tf.keras.layers.Dense(512, activation='relu'),
         tf.keras.layers.Dense(256, activation='relu'),
         tf.keras.layers.Dense(128, activation='relu'),
         tf.keras.layers.Dense(64, activation='relu'),
         tf.keras.layers.Dense(32, activation='relu'),
         tf.keras.layers.Dense(16, activation='relu'),
         tf.keras.layers.Dense(32, activation='relu'),
         tf.keras.layers.Dense(64, activation='relu'),
         tf.keras.layers.Dense(128, activation='relu'),
         tf.keras.layers.Dense(256, activation='relu'),
         tf.keras.layers.Dense(512, activation='relu'),
         tf.keras.layers.Dense(1024, activation='relu'),
         tf.keras.layers.Dense(1)  # Output layer with one neuron for regression
     ])

     # Compile the model
     model.compile(optimizer='adam', loss='huber')
```

Figura 9. Arquitectura de la red neuronal (ANN) final utilizada para realizar la regresión sobre los datos de entrada. ANN simétrica con 13 capas, número de neuronas en verde, última línea de código muestra el optimizador y la función de pérdida usados.

Como bien se ha expuesto en la sección de fundamentos teóricos, las ANN trabajan de una forma especial respecto de los demás modelos de ML, y debido a su profundidad de aprendizaje se denomina Aprendizaje Profundo (DL) a este tipo de ML. En DL suele usarse al tener una mayor cantidad de datos, aunque no significa que no pueda trabajar con datos no estructurados ni gran cantidad de estos, y en este caso se ha decidido optar por este método ante la respuesta deficiente de los modelos convencionales de ML.

Al probar varias combinaciones posibles para la arquitectura de la ANN, se identificó que el mejor resultado de precisión se obtenía cuando se reasignó la proporción del modelo, destinada a la validación, siendo esta de un 10% (para más condiciones remitirse a sección de metodología, Fase 2). En parte esto aparenta coherencia debido a que la ANN pudo entrenarse mayor con un mayor número de datos, lo cual es pertinente al usar DL. Sin embargo, la cantidad de datos de entrenamiento que se suministraron de más no fue sustancialmente mayor respecto de los modelos de regresión de ML, por lo cual la evidente mejoría que se presenta a continuación no parece tener mucha influencia de parte de este factor. Luego de realizar el entrenamiento y procesamiento, la red neuronal provee un valor de **MAE de 30,54** y un valor de **$|R^2|$ de 0,7928**; aunque el valor de MAE aumentó, es de recordar que este valor no refleja un porcentaje si no un promedio entre las diferencias de datos predichos y reales (respecto del subconjunto de validación) y la menor cantidad de datos pudo causar que en promedio las diferencias observadas fueran mayores. El valor del coeficiente de correlación cercano a 0,8 es prometedor ya que los descriptores y los valores de porcentaje de rendimiento se correlacionan de forma más eficiente y, aunque es una aproximación matemáticamente incorrecta, expresado en valor porcentual el modelo es aproximadamente un 80% preciso (al menos potencialmente), lo cual es un resultado verdaderamente positivo para las pretensiones de la investigación. La profundidad de análisis y extracción de información de la ANN se ha probado con la

diferencia entre los resultados, siendo este el modelo final entrenado para la predicción del % Rendimiento de la reacción a partir de las condiciones iniciales (descriptores). Este ha sido el resultado de la Fase 2 del proyecto y se demostró, como se mostrará, que el modelo es capaz de brindar valores no atípicos, lo que en primera instancia fue un logro del entrenamiento.

Terminado el entrenamiento del modelo y su refinamiento, se ejecuta la reacción en el laboratorio con el fin de obtener el rendimiento de reacción "real". La reacción se condujo como se menciona en la sección de metodología (Fase 3), y en esta sección se hace necesario aclarar algunos aspectos importantes de esta sección experimental.

En este punto se contaba con dos opciones: introducir en el sistema de reacción un catalizador ya incluido en el conjunto de datos generado en la Fase 1, sabiendo que la respuesta del modelo iba a ser conocida de antemano o incluir en el sistema de epoxidación un catalizador ajeno al conjunto de datos que permitiera evaluar la robustez del modelo entrenado y su comportamiento ante datos "desconocidos".

Por colaboración entre el grupo de investigación en Catálisis de la Universidad del Cauca y el grupo de investigación en Estado Sólido y Catálisis Ambiental (ESCA) de la Universidad Nacional de Colombia-Sede Bogotá fue posible hacer uso de un catalizador recientemente sintetizado para epoxidación de aceite de soya. El catalizador se describe a profundidad en la publicación de Castellanos y colaboradores [64], y en este documento se mencionarán de forma breve algunas de las apreciaciones más relevantes, con el fin de retomar el enfoque dedicado a la aplicación del ML y DL a este experimento.

El catalizador tiene una estructura definida como un *Metal-Organic Framework (MOF)* o Red Metalo-Orgánica en español, sólidos que tienen como característica principal un metal enlazado sistemáticamente a ligandos orgánicos con sustituyentes adecuados, generando una estructura sólida altamente cristalina, lo cual provee a estos materiales de áreas superficiales usualmente altas (por encima de 7000 m^2/g), bajas densidades y alto contenido del metal, que sigue siendo el sitio activo (o al menos parte del mismo) [78], [79], [80], [81], [82]. Los MOF han tenido un amplio rango de aplicaciones en catálisis [64] y en particular en la catálisis de alquenos, aunque con frecuencia en moléculas modelo de baja masa molar [83] como, por ejemplo, el **ciclohexeno**. Esta última es la razón por la cual el grupo ESCA se interesó en aplicar este catalizador en la reacción de epoxidación de moléculas naturales de mayor masa molar [64].

El catalizador tiene al Mo como centro metálico y parte del sitio activo, una estructura soportada en principio por el ligando **ácido 2,2'-bipiridin-5,5'-dicarboxílico** posteriormente modificado con la introducción del Ga en la red cristalina, y cuya estructura base se muestra en la **Figura 3**, corroborada según los autores con la ayuda de estudios de difracción de rayos X [64]. El área superficial y volumen de poro del catalizador fueron 214 m^2/g y 0,78 cm^3/g respectivamente, mientras que la carga total de Mo fue de 5,9 % [64]. El catalizador se ha simbolizado como MoO_2Cl_2@COMOC-4 y se determinó que la especie activa corresponde a MoO_2; además, como ya se mencionó durante la sección **Antecedentes**, el Mo es uno de los metales más estudiados en el campo de la catálisis de olefinas, habiendo provisto incontables resultados catalíticos para este sistema.

La reacción conducida en el grupo ESCA con el aceite de soya como sustrato implicó un estudio completo de optimización de las condiciones, condiciones que se han implementado como referencia para llevar a cabo la reacción en el grupo de Catálisis. Es evidente que el estudio realizado para el

sustrato natural por el grupo ESCA no implica que bajo las mismas condiciones el rendimiento de epoxidación del sustrato ciclohexeno vaya a ser optimo, pero es el punto de partida para probar el catalizador con esta molécula y para evaluar la capacidad del modelo. Es importante recordar que las condiciones de reacción son equivalentes a los valores de entrada de cada descriptor, esto es, las variables independientes del modelo entrenado que han de conducir a la regresión (predicción) del % Rendimiento, es decir, la variable dependiente (de respuesta). En la **Tabla 5** se muestran estas condiciones.

Como ya se ha descrito en varias ocasiones, la elección de los descriptores está ligada principalmente a la simplicidad de obtener los mismos y es lógico pensar que los 14 descriptores numerados en la **Tabla 5** son fácilmente extraíbles de la sección experimental de cada una de las publicaciones, cálculos simples y/o, en casos de máxima complejidad, contacto con los autores. Para ejemplificar lo anterior, se toma el rendimiento de la reacción, el cual corresponde a la variable respuesta; esta columna en el *dataset* final (aún sin codificar) aparecía vacía en algunas entradas, por lo cual se temía no poder contar con un número mayor de estudios en el conjunto. Por esta razón se tomó la determinación de buscar una forma simple de calcular el % de rendimiento a partir de otras variables que pudieran o no estar incluidas en el *dataset*, y estas variables fueron % de conversión y de selectividad, y se muestra el cálculo realizado mediante la Ecuación 3.

$$\%\, rendimiento\ (\%R) = \frac{\%conversión * \%selectividad}{100} \quad Ecuación\ 3$$

Tabla 5. Valores de entrada de los 14 descriptores utilizados para la predicción del % de rendimiento.

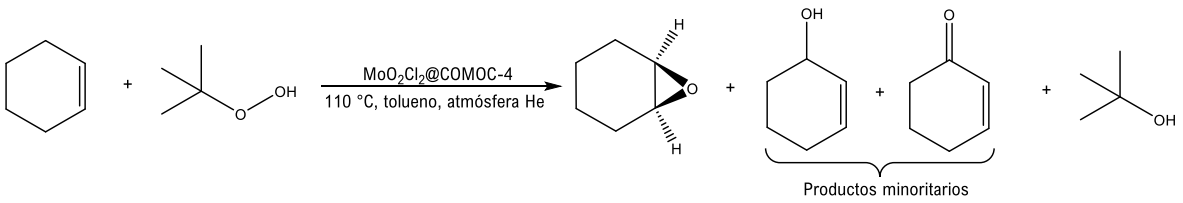
Descriptor	Valor	Código normalizado
Metal	Mo	0.687500
Tipo de catalizador	Heterogéneo	0.250000
Estado de oxidación	6+	0.875000
Configuración electrónica	d^0	0.083333
Geometría	N.A.	0.285714
Ligando/soporte	Mo-UN-1000 (MOF de Mo con Zr/O especies)	0.664179
Clase ligando/soporte	Soportado	0.416667
Oxidante	TBHP	0.666667
Clase oxidante	Orgánico	0.666667
Solvente	Tolueno	0.848485
Clase solvente	Apolar	0.200000
Temperatura (°C)	110	1.000000
Tiempo (h)	4	0.163775
% mol catalizador (Mo)	1	0.021584

De la misma manera se pudieron obtener los valores de todos los descriptores para todos los estudios incluidos, y en este aspecto radica gran parte del valor de este estudio: la construcción y el refinamiento de un conjunto de datos, sea el presentado en este documento o en cualquier otro

proyecto, pueden ser logrados mediante extracción simple de datos poco complejos que no requieran de la implementación de herramientas computacionales más avanzadas o costosas, y puede ser de gran utilidad para estudios sencillos de aplicación de la IA en el desarrollo de las ciencias catalíticas. Por supuesto, la profundidad y capacidad de análisis y predicción que los conjuntos de datos sencillos les brindan a los modelos de regresión no son los mejores posibles, pero es un recurso valido que se puede utilizar en estudios primarios de implementación.

A pesar de la obtención de todos los descriptores de forma satisfactoria, aún existe una discrepancia que seguramente influiría en la predicción final del modelo: la descripción del catalizador, es decir, el valor de entrada del descriptor "Ligando/soporte". Dado que el catalizador utilizado en la reacción fue recientemente sintetizado y aplicado a una reacción de epoxidación sobre un sustrato diferente al deseado por el grupo de trabajo, no fue posible contar con una publicación que provisionara al menos una entrada al *dataset* con el valor del descriptor equivalente al catalizador usado, por lo cual, al entrenarse el modelo con los datos provistos el mismo no fue capaz de aprender sobre este valor del descriptor "Ligando/soporte" dado que no existe en el *dataset*. Esta es la principal inferencia que se realizó de la evaluación de la robustez del modelo: no responde bien ante datos ajenos al *dataset*. Esto causaría que, si se ejecuta el modelo y se requiere el valor de rendimiento de la reacción con las características tal cual se ajustaron en laboratorio, este no sería capaz de brindar un resultado pues el valor de entrada de dicho descriptor no coincide con ninguno en la lista de alimentación. Por esta razón la solución menos desviada posible de la realidad fue buscar un valor fisicoquímicamente aproximado de entre todos los catalizadores descritos en el *dataset* y hacer uso de este como el presunto catalizador utilizado en la reacción; aunque evidentemente esta acción no es para nada correcta desde el punto de vista científico, es el recurso que se ha utilizado para poder ejecutar la predicción con el modelo entrenado y no representa un error en todo el procedimiento previo de entrenamiento y pruebas. Ya que el estado del arte se puede actualizar y así mismo el *dataset*, se podrán seguir incluyendo más datos que podrán ayudar a refinar el modelo y se podrá incluir en algún punto un resultado con el catalizador usado realmente para hacer más precisa la predicción. Entonces, aunque aparenta ser erróneo, es valioso recordar que el % Rendimiento predicho por el modelo será equivalente a si se hubiese usado el catalizador descrito como "Mo-UN-1000 (MOF de Mo con Zr/O especies)".

La reacción se llevó a cabo por un periodo de 4 horas tomando alícuotas como se especifica en la sección de metodología de la Fase 3, y los resultados de área de las señales se muestran en la **Tabla 6**, identificando el tolueno, ciclohexeno y dodecano con la ayuda de la medición de un blanco de reacción en ausencia de catalizador y agente oxidante. El **Esquema 5** muestra la reacción general.



Esquema 5. Esquema general de la reacción de epoxidación del ciclohexeno.

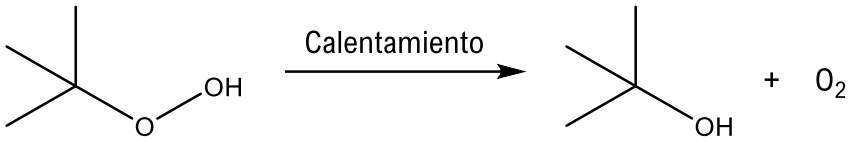
Tabla 6. Áreas de las señales desconocidas observadas mediante el análisis por CG-EM.

Tiempo (h)	Área del compuesto *unk1*	Área del compuesto *unk2*	Área del dodecano (PI)
0,5	2569,200	7556,160	73992,953
1,0	3560,500	7507,341	76659,727
2,0	3280,900	7770,044	85185,867
3,0	3318,800	6666,931	82101,211
4,0	3412,600	4046,367	85394,703

Los resultados del análisis mostraron la aparición de, en promedio, dos nuevas señales cromatográficas denominadas "desconocido 1" y "desconocido 2" (unk1 y unk2) respectivamente, las cuales parecen aumentar en intensidad a medida que se miden las alícuotas de tiempos más avanzados; sin embargo, al medir la alícuota de tiempo t=4 horas se observa una aparente disminución en las intensidades de estas señales, lo cual podría indicar una evidente desaparición de estos productos, ya sea por perdida o degradación de los mismos, o por una deficiencia del método cromatográfico o procedimiento de inyección. Como el valor para el descriptor "tiempo" en horas corresponde a 4, sobre este valor es el que se toman las áreas. Se indica el valor del patrón interno (PI) ya que respecto de éste se cuantifica.

Con la ayuda del blanco en presencia de TBHP y sin catalizador fue posible observar la aparición de una señal coincidente con el tiempo de retención del compuesto "unk1", por lo cual se asume que esta señal no corresponde necesariamente a la formación del epóxido del ciclohexeno, ya que a la vez se asume que este compuesto aparece en tiempos dentro de las cuatro horas gracias a la acción del catalizador; la aparición de la señal del compuesto "unk1" permite sugerir que si hay aparición del epóxido es más probable que este corresponda al compuesto "unk2". La aparición de los compuestos desconocidos 1 y 2, así mismo como la aparente disminución y aumento de la cantidad de estos, puede ser debido a varias causas, y a continuación se discuten las que parecen ser más probables de ocurrir. Sin necesidad de la presencia del catalizador, el hidroperóxido es capaz de oxidar al ciclohexeno hacia varios subproductos, siendo los más destacados el ciclohexenol y la ciclohexenona [84], [85]. La aparición inicial del epóxido puede ir seguida de su oxidación a ciclohexenol y posteriormente a ciclohexenona, especialmente a tiempos de calentamiento muy prolongados [84]; además, en un mayor grado y también en condiciones de elevada temperatura, el TBHP puede descomponerse en el alcohol y el oxígeno, como lo muestra el **Esquema 6** (tomado y adaptado de [84]). En la **Tabla 6** se puede ver como después de la primera media hora el área del compuesto unk1 aumenta y se mantiene, en promedio, constante durante la reacción mientras que el área del compuesto unk2 luego de su aparición disminuye paulatinamente durante el avance de reacción. Al tomar en cuenta que el compuesto unk2 es el único que aparece con la presencia del catalizador en el sistema y tomando su continuo decrecimiento en términos de área se puede suponer que se está hablando como mínimo de uno de los productos principales de oxidación, probablemente el epóxido o el ciclohexenol. A esta conclusión se llega debido a que, en caso de ser el epóxido el compuesto que genera dicha señal, este podría ver disminuido su rendimiento debido a la ruptura consecuente del anillo epoxirano hacia los subproductos ya mencionados; de estos subproductos, el que se podría

descomponer con mayor probabilidad al avanzar la reacción es el ciclohexenol, que en presencia de TBHP y calentamiento continuo puede aumentar la producción de ciclohexenona.



Esquema 6. Descomposición térmica del TBHP.

La simbología asignada a estas señales de la forma en que se ha hecho y el análisis derivado de la misma tiene su fundamento en la dificultad que representó contar, en primer lugar, con un estándar del epóxido del ciclohexeno para generar una curva de calibración por patrón interno y cuantificar de forma adecuada, y luego, la dificultad que representó poder hacer uso del equipo de cromatografía de gases necesario para las mediciones. Es por ello que se ha tomado la determinación de utilizar un método que implica relaciones matemáticas simples entre las áreas de los compuestos analizados y utiliza el fundamento de la cuantificación por patrón interno.

La variable elegida como marcador de la eficiencia del sistema catalítico ha sido el porcentaje de rendimiento (%R) de la reacción, y ya sea que se interprete como se muestra en la Ecuación 3 o como su clásica connotación de lo producido por cada 100 unidades de lo esperado, la respuesta es la misma, y por ende se puede asumir en este caso que el rendimiento se calcula con base en lo máximo esperado a producir, esto es, 4,0 mmol del epóxido (que sería el caso idealista, considerando la inclusión de 4,0 mmol de alqueno). Si se usó el dodecano como patrón interno, se espera que relacionando las áreas de las nuevas señales con las de este se pueda calcular un estimado de la cantidad, sabiendo claro la cantidad de patrón interno presente en un inicio y que presuntamente se mantiene constante durante el análisis. La Ecuación 4 muestra la relación que se plantea entre el PI y las especies producto de la oxidación, asumiendo que las hay y que el comportamiento es tal que se puede aproximar a dicha expresión.

$$\frac{A_i}{A_{PI}} = \frac{mmol_i}{mmol_{PI}} \quad Ecuación\ 4$$

Donde el subíndice *i* indica el área o las mmol de la especie a cuantificar, según corresponda. Luego de calcular la cantidad de los compuestos causantes de cada una de las nuevas señales, se utiliza la expresión de la Ecuación 5 para calcular el rendimiento, similar a lo hecho convencionalmente para analizar la eficiencia de las reacciones químicas, siempre teniendo en cuenta que la máxima cantidad posible producida de un producto de oxidación [del alqueno] es de 4,0 mmol.

$$\%\ R = \frac{mmol_i}{4{,}0\ mmol} * 100 \quad Ecuación\ 5$$

En la **Tabla 7** se presentan los resultados de rendimiento de las señales de los compuestos unk1 y unk2 a un tiempo de 4 horas que, como ya fue dicho, es el tiempo final de reacción y es el valor del descriptor "Tiempo".

Tabla 7. Rendimientos para los compuestos desconocidos calculados con el método señalado.

Compuesto	Unk1	Unk2
% Rendimiento	4,00	4,74

Los rendimientos mostrados no superan en ningún caso el 5 %, siendo ligeramente mayor el rendimiento hacia el compuesto del cual se tiene mayor sospecha de ser el epóxido. La dificultad de no haber podido contar con el estándar y la técnica adecuados para la correcta cuantificación implica que no se puede corroborar con certeza si existe o no producción del epóxido y de ocurrir, no se puede saber a qué señal se le puede atribuir; sin embargo la aparición de diferentes señales durante la realización de la reacción arrojan resultados prometedores ante posibles secuelas o refinamiento de la investigación actual, y permiten tener una idea del comportamiento del catalizador de Mo provisto por el grupo ESCA, además de dejar indicios sobre las posibles correcciones a hacer sobre el sistema catalítico, siendo la más notoria de ellas el tiempo de reacción, infiriéndose esto del comportamiento del sistema catalítico hasta antes de la última alícuota tomada en la cual empezó a aparecer un líquido lipofílico de color amarillento-marrón y que no se mostraba en otras alícuotas, así como la estrecha relación con la disminución evidente de las áreas de los demás componentes de la mezcla —desde el alqueno hasta el solvente y el patrón interno— que hacen sospechar que después de tres horas de reacción había una degradación o pérdida de gran parte de la mezcla, ya sea por reacciones indeseadas o fugas en algún punto del sistema de análisis.

La última parte del estudio consistió en la introducción de los valores de entrada al modelo de DL entrenado para su procesamiento y predicción rendimiento de la reacción. Lo anterior se hizo de manera similar al entrenamiento inicial del modelo, programando sobre un cuadernillo nuevo en la plataforma Google Colab, e importando a él el *dataset* codificado y normalizando de nueva cuenta; luego se importó el modelo y se le pidió al modelo que, después de eliminar la columna de la variable de respuesta de la lista de descriptores, brindase la predicción del % Rendimiento con base en los datos de la **Tabla 5** (que tuvo que codificar y normalizar para coincidencia de las escalas, claro está). En la **Figura 10** y en el **Anexo 9**, se muestra la apariencia del resultado obtenido en la plataforma Google Colab y que es, en principio y siendo este un logro en sí mismo, un valor no atípico, enmarcado dentro de la definición del rango de la variable de respuesta (entre 0 y 100 %). El porcentaje de reacción predicho por el modelo es de **84,54 %** aproximadamente, y aunque es aparentemente lejano si se compara con el rendimiento mostrado en la **Tabla 7**, se nota efectivamente distante, con un porcentaje de error absoluto de 94,39 %. La **Tabla 8** muestra la comparación entre ambos rendimientos (resultados sometidos a verificación).

Tabla 8. Comparación entre los rendimientos obtenidos.

Rendimiento real	Rendimiento predicho
4,74 %	84,54 %

Esta lejanía aparenta ser muy alta, pero se pueden señalar dos principales fuentes de error que pueden ser refinadas en un futuro cercano: el método de cuantificación y la descripción del catalizador en el valor de entrada. Está claro que, para ser el primer resultado de implementación de estas herramientas computacionales en el grupo de Catálisis, este es prometedor y refleja el esfuerzo realizado por conseguir ser pioneros de esta rama de investigación a nivel institucional; la línea de investigación se ha establecido en el grupo y se esperas que se asiente a corto y mediano plazo, lo que permitirá un constante mejoramiento del modelo obtenido en esta investigación, aumentando progresivamente su precisión de predicción. La corrección de los factores de error ya tratados, sumado al hecho del constante refinamiento del *dataset* puede conducir a un potencial resultado mucho más prometedor, que, con un probable aumento de la correlación del modelo y disminución del error absoluto del mismo, puede impulsar esta rama de investigación nueva en el grupo a obtener más y mejores resultados.

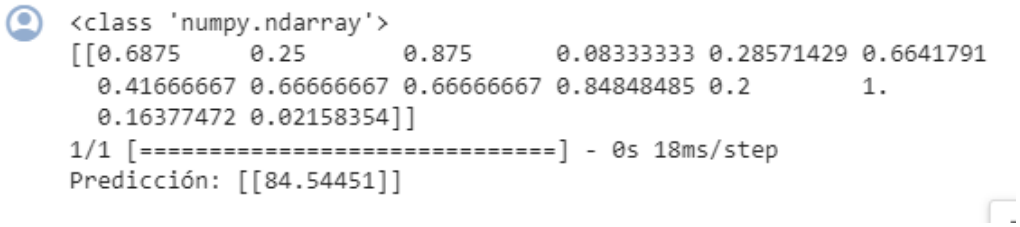


Figura 10. Interfaz de Google Colaboratory que muestra el valor del porcentaje de rendimiento predicho por el modelo entrenado bajo la leyenda "Predicción".

6 CONCLUSIONES

Más allá de la utilidad inmediata que pueda tener el producto principal de esta investigación en aplicaciones industriales o académicas, se ha demostrado que la utilización de herramientas computacionales es tangible para impulsar el desarrollo de la catálisis. Se ha logrado emplear la Inteligencia Artificial de tal manera que complementa el análisis y esfuerzo del analista humano promedio, obteniendo resultados intelectuales más sólidos.

En la Fase 1 fue posible construir un *dataset* a partir de datos extraídos de la literatura que a su vez se seleccionó mediante la aplicación de un mapeo sistemático, una metodología de búsqueda y filtrado de bibliografía que no suele ser usada el ámbitos como el químico, y desde la perspectiva personal fue un logro de suma importancia para la automatización y uniformidad de los procesos investigativos; además de la implementación relativamente novedosa en el grupo del mapeo sistemático, el conjunto de datos construido recopila gran cantidad de estudios en los que se puede visualizar a modo de *review* la gran variedad de condiciones y catalizadores aplicados a la epoxidación catalítica del ciclohexeno, esto incluso entre unos pocos metales como Mo, Pd, Co y Ni, entre otros, convirtiendo a este *dataset* en algo más que alimento de entrenamiento computacional.

La Fase 2 requirió un trabajo más extenso del planteado inicialmente debido a las complicaciones inherentes al manejo de grandes cantidades de datos y la característica categórica de la mayoría de ellos, lo que supone una gran dedicación a la limpieza y preprocesamiento de los mismo, pero que condujo sin duda a uno de los avances más importantes en el ámbito local de aplicación de las Ciencias Computacionales y la IA en el desarrollo de las ciencias catalíticas, y que fue el modelo entrenado de predicción de porcentaje de rendimiento. A parte del modelo, la experiencia adquirida por el equipo de trabajo en el concepto de la programación es de alta valía y seguramente serán de ayuda en futuras investigaciones en la misma rama, además de que la exploración de distintos modelos de regresión de ML y la profundización “inesperada” en el terreno del DL proveen sensación de satisfacción e impulso de mejoría ante la inminente presencia de la Ciencia de Datos en el andar catalítico.

La Fase 3 permitió observar el funcionamiento del modelo en sí y aportó valiosa información e inferencias útiles, siendo tal vez la más importante el hecho de que el modelo no haya tenido irregularidades a la hora de brindar un porcentaje de rendimiento y se haya mantenido en el rango típico; este dato cobra especial importancia ya que es algo que los modelos de regresión de ML suelen presentar cuando trabajan con bastantes datos categóricos y no numéricos, y permite cimentar una base sólida para el posterior refinamiento de todos los resultados obtenidos en torno a este importante avance computacional en catálisis.

Además de lo relativo al modelo, una apreciación por demás interesante fue la aplicación del nuevo catalizador sintetizado por el grupo ESCA en la ciudad de Bogotá y que a grandes rasgos muestra potencial para reacciones de epoxidación y una estructura y comportamiento interesantes de analizar en futuras colaboraciones, no solamente para epoxidaciones sino también para reacciones catalíticas de alto interés industrial.

La investigación en general provee una valiosísima base para el aprovechamiento de las herramientas computacionales en la química y ha permitido conectar instituciones, más allá de la interdisciplinariedad, permitiendo un trabajo conjunto a gran escala entre tres universidades y expertos en ambas áreas de estudio, catálisis y ciencia de datos. Ha sido muy importante poder darse cuenta que en la química actual la fundamentación teórica sobre el funcionamiento de los modelos

atómicos y las leyes moleculares de la materia no son lo único que construye al químico sino que también cobra especial relevancia el papel que juegan ciencias “marginadas” por las naturales como las ingenierías de software, la programación y la ciencia de datos, y a raíz de la visualización de lo anterior surgió durante el desarrollo del presente trabajo el objetivo personal de dar a conocer la importancia que tiene actualmente la inclusión de una formación en ciencias computacionales, más específicamente en química computacional, en los currículos de pregrado en química en todo programa en el que no se cuente con ello, ya que el trabajo del químico se ve año tras año más “opacado” por otras profesiones e incluso por otras inteligencias cuando no debería ser de esa manera.

Por supuesto que los resultados no solamente reflejan aportes del todo reconfortantes y deja aspectos sobre los cuales mejorar, siendo tal vez los dos más importantes la refinación del conjunto de datos en cuanto a cantidad de los mismos y estructuración de los descriptores, que aunque siendo la pretensión la de buscar facilidad en la obtención, es visible que otro tipo de técnicas computacionales —con las que no toda institución cuenta— serian de gran ayuda en la extracción de parámetros que permitan identificar con autenticidad real cada catalizador y por ende, cada sistema catalítico; el segundo aspecto de importancia seria incluir en el refinamiento todo dato que pueda ser posible de replicar en el laboratorio y así evitar la utilización de aproximaciones un tanto burdas como la descripción del catalizador intercambiada, ya que resulta en una evidente desviación que no es observable en el laboratorio. Sin embargo, es lo anterior y cualquier otro aspecto a mejorar lo que promete aun un futuro que necesite de químicos, mejor preparados y con nuevos conocimientos, que sean capaces de implementar recursos computacionales tan útiles de una manera certera y correcta en las ciencias químicas y que sean capaces de corregir cualquier error a conciencia, siendo aprendices en todo momento y destacando cada vez más la profesión del químico.

7 REFERENCIAS

[1] J. García Martínez, “The new chemist”, *Chemical & Engineering News (C&EN)*, vol. 96, núm. 6, feb. 2018, Consultado: el 11 de julio de 2023. [En línea]. Disponible en: https://cen.acs.org/articles/96/i6/new-chemist.html?ref=search_results

[2] V. Duros *et al.*, “Human versus robots in the discovery and crystallization of gigantic polyoxometalates”, *Angewandte Chemie*, vol. 129, núm. 36, pp. 10955–10960, 2017.

[3] C. Zhang y Y. Lu, “Study on artificial intelligence: The state of the art and future prospects”, *J Ind Inf Integr*, vol. 23, p. 100224, 2021.

[4] A. T. Bell y M. Head-Gordon, “Quantum mechanical modeling of catalytic processes”, *Annu Rev Chem Biomol Eng*, vol. 2, pp. 453–477, 2011.

[5] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, y K. Shimizu, “Machine learning for catalysis informatics: recent applications and prospects”, *ACS Catal*, vol. 10, núm. 3, pp. 2260–2297, 2019.

[6] W. Yang, T. T. Fidelis, y W.-H. Sun, “Machine learning in catalysis, from proposal to practicing”, *ACS Omega*, vol. 5, núm. 1, pp. 83–88, 2019.

[7] Z. Li, S. Wang, y H. Xin, “Toward artificial intelligence in catalysis”, *Nat Catal*, vol. 1, núm. 9, pp. 641–642, 2018.

[8] B. R. Goldsmith, J. Esterhuizen, J. Liu, C. J. Bartel, y C. Sutton, “Machine learning for heterogeneous catalyst design and discovery”, 2018.

[9] L. O. Jones, M. A. Mosquera, G. C. Schatz, y M. A. Ratner, “Embedding methods for quantum chemistry: Applications from materials to life sciences”, *J Am Chem Soc*, vol. 142, núm. 7, pp. 3281–3295, 2020.

[10] T. Allen, “Computers as Scientists”, en *Machine Learning in Chemistry: The Impact of Artificial Intelligence*, 1a ed., H. Cartwright, Ed., Oxford, United Kingdom: Royal Society of Chemistry, 2020.

[11] IBM, “¿Qué es la ciencia de datos?” Consultado: el 21 de diciembre de 2023. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/data-science

[12] W. Van Der Aalst y W. van der Aalst, *Data science in action*. Springer, 2016.

[13] IBM, “¿Qué es la inteligencia artificial (IA)?” Consultado: el 21 de diciembre de 2023. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/artificial-intelligence

[14] C. Machinery, “Computing machinery and intelligence-AM Turing”, *Mind*, vol. 59, núm. 236, p. 433, 1950.

[15] Coursera Staff Editorial Team, “Deep learning vs. Machine learning: Guía para principiantes”, Coursera. Consultado: el 21 de diciembre de 2023. [En línea]. Disponible en: https://www.coursera.org/mx/articles/ai-vs-deep-learning-vs-machine-learning-beginners-guide

[16] IBM, “¿Qué es machine learning?” Consultado: el 21 de diciembre de 2023. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/machine-learning

[17] Platzi, “¿Qué es Machine Learning? Descubre el poder del aprendizaje automático”. Consultado: el 22 de diciembre de 2023. [En línea]. Disponible en: https://platzi.com/blog/machine-learning-que-es/

[18] S. Gollapudi, “Introduction to Machine Learning”, en *Practical Machine Learning*, 1a ed., Birmingham, Reino Unido: PACKT Publishing, 2016, pp. 1–40.

[19] El mundo de los datos, “Técnicas para codificar las variables categóricas (I): codificación ordinal y one-hot”. Consultado: el 23 de diciembre de 2023. [En línea]. Disponible en: https://elmundodelosdatos.com/tecnicas-para-codificar-variables-categoricas-ordinal-one-hot/

[20] Minitab.com, “¿Qué son variables categóricas, discretas y continuas?” Consultado: el 23 de diciembre de 2023. [En línea]. Disponible en: https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistical-modeling/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/

[21] Microsoft.com, “Introducción a los modelos de regresión para Machine Learning [Parte 5] | Aprendizaje automático para principiantes”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://learn.microsoft.com/es-es/shows/machine-learning-for-beginners/introduction-to-regression-models-for-machine-learning-machine-learning-for-beginners#time=00m09s

[22] E. Frank, Y. Wang, S. Inglis, G. Holmes, y I. H. Witten, “Using model trees for classification”, *Mach Learn*, vol. 32, pp. 63–76, 1998.

[23] J. M. Vasallo, “Análisis de regresión lineal simple y múltiple”, en *Estadística aplicada a las ciencias de la salud*, Elsevier España, 2015, pp. 157–223.

[24] A. Natekin y A. Knoll, “Gradient boosting machines, a tutorial”, *Front Neurorobot*, vol. 7, p. 21, 2013.

[25] GeeksforGeeks, “Stochastic Gradient Descent Regressor”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://www.geeksforgeeks.org/stochastic-gradient-descent-regressor/

[26] IBM, “Documentación IBM: Regresión Ridge del kernel”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-statistics/28.0.0?topic=statistics-kernel-ridge-regression

[27] Interactive Chaos, “Elastic Net | Interactive Chaos”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/elastic-net

[28] Cienciadedatos.net, “Regularización Ridge, Lasso y Elastic Net con Python y Scikitlearn”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python

[29] Tutorialspoint.com, “Scikit Learn - Bayesian Ridge Regression”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://www.tutorialspoint.com/scikit_learn/scikit_learn_bayesian_ridge_regression.htm

[30] Interactive Chaos, “Gradient Boosting | Interactive Chaos”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/gradient-boosting

[31] Medium, “An Introduction to Support Vector Regression (SVR) in Machine Learning”. Consultado: el 9 de enero de 2024. [En línea]. Disponible en: https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a

[32] IBM, “¿Qué es Deep Learning?” Consultado: el 8 de enero de 2024. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/deep-learning

[33] Inc. Amazon Web Services, “¿Qué es una red neuronal?” Consultado: el 8 de enero de 2024. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/neural-network/#:~:text=Las%20redes%20neuronales%20artificiales%20aprenden,diferentes%20en%20la%20red%20neuronal.

[34] Interactive Chaos, “Estructura de una red neuronal”. Consultado: el 8 de enero de 2024. [En línea]. Disponible en: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/estructura-de-una-red-neuronal

[35] J. A. Moulijn y R. A. van Santen, “History of catalysis”, 2017.

[36] A. Modak, P. Bhanja, S. Dutta, B. Chowdhury, y A. Bhaumik, “Catalytic reduction of CO 2 into fuels and fine chemicals”, *Green Chemistry*, vol. 22, núm. 13, pp. 4002–4033, 2020.

[37] L. Wang, J. Wu, S. Wang, H. Liu, Y. Wang, y D. Wang, “The reformation of catalyst: From a trial-and-error synthesis to rational design”, *Nano Res*, pp. 1–41, 2023.

[38] T. Oyama, “Rates, Kinetics, and Mechanisms of Epoxidation: Homogeneous, Heterogeneous, and Biological Routes”, en *Mechanisms in HOMOGENEOUS AND HETEROGENEOUS EPOXIDATION CATALYSIS*, 1a ed., T. Oyama, Ed., Oxford, Reino Unido: ELSEVIER, 2008, pp. 4–72.

[39] K. A. Jørgensen, “Transition-metal-catalyzed epoxidations”, *Chem Rev*, vol. 89, núm. 3, pp. 431–458, 1989.

[40] L. Wade, “Reacciones de los alquenos”, en *Química Orgánica*, 7a ed., vol. 1, G. López, Ed., Ciudad de México, México: Pearson Educación, 2011, pp. 355–356.

[41] J. Davarpanah y A. R. Kiasat, “Synthesis and characterization of SBA-polyperoxyacid: An efficient heterogeneous solid peroxyacid catalyst for epoxidation of alkenes”, *Catal Commun*, vol. 46, pp. 75–80, 2014.

[42] S. Huber, M. Cokoja, y F. E. Kuehn, “Historical landmarks of the application of molecular transition metal catalysts for olefin epoxidation”, *J Organomet Chem*, vol. 751, pp. 25–32, 2014.

[43] S. Boudjema, M. Zerrouki, y A. Choukchou-Braham, “Experimental Design for Modeling and Multi-response Optimization of Catalytic Cyclohexene Epoxidation over Polyoxometalates”, *Journal of the Chinese Chemical Society*, vol. 65, núm. 4, pp. 435–444, 2018.

[44] J. M. Thomas *et al.*, “The Identity in Atomic Structure and Performance of Active Sites in Heterogeneous and Homogeneous, Titanium− Silica Epoxidation Catalysts”, *J Phys Chem B*, vol. 103, núm. 42, pp. 8809–8813, 1999.

[45] Y. Shen, P. Jiang, P. T. Wai, Q. Gu, y W. Zhang, “Recent progress in application of molybdenum-based catalysts for epoxidation of alkenes”, *Catalysts*, vol. 9, núm. 1, p. 31, 2019.

[46] C.-C. Su, J. W. Reed, y E. S. Gould, “Metal ion catalysis of oxygen-transfer reactions. II. Vanadium and molybdenum chelates as catalysts in the epoxidation of cycloalkenes”, *Inorg Chem*, vol. 12, núm. 2, pp. 337–342, 1973.

[47] F. Trifiro, P. Forzatti, S. Preite, y I. Pasquon, “Liquid phase epoxidation of cyclohexene by tert-butyl hydroperoxide on a Mo-based catalyst”, *Journal of the Less Common Metals*, vol. 36, núm. 1–2, pp. 319–328, 1974.

[48] C. Venturello y R. D’Aloisio, “Quaternary ammonium tetrakis (diperoxotungsto) phosphates (3-) as a new class of catalysts for efficient alkene epoxidation with hydrogen peroxide”, *J Org Chem*, vol. 53, núm. 7, pp. 1553–1557, 1988.

[49] W. A. Herrmann, R. M. Kratzer, H. Ding, W. R. Thiel, y H. Glas, “Methyltrioxorhenium/pyrazole—A highly efficient catalyst for the epoxidation of olefins”, *J Organomet Chem*, vol. 555, núm. 2, pp. 293–295, 1998.

[50] D. De Vos y T. Bein, “Highly selective epoxidation of alkenes and styrenes with H 2 O 2 and manganese complexes of the cyclic triamine 1, 4, 7-trimethyl-1, 4, 7-triazacyclononane”, *Chemical communications*, núm. 8, pp. 917–918, 1996.

[51] J. A. Burns y G. M. Whitesides, “Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition”, *Chem Rev*, vol. 93, núm. 8, pp. 2583–2601, 1993.

[52] D. M. Himmelblau, “Applications of artificial neural networks in chemical engineering”, *Korean journal of chemical engineering*, vol. 17, pp. 373–392, 2000.

[53] M. Erdem Günay y R. Yıldırım, “Recent advances in knowledge discovery for heterogeneous catalysis using machine learning”, *Catalysis Reviews*, vol. 63, núm. 1, pp. 120–164, 2021.

[54] V. Arcotumapathy, A. Siahvashi, y A. A. Adesina, “A new weighted optimal combination of ANNs for catalyst design and reactor operation: methane steam reforming studies”, *AIChE journal*, vol. 58, núm. 8, pp. 2412–2427, 2012.

[55] Z. Li, X. Ma, y H. Xin, “Feature engineering of machine-learning chemisorption models for catalyst design”, *Catal Today*, vol. 280, pp. 232–238, 2017.

[56] D. M. Lustosa y A. Milo, “Machine learning classifies catalytic-reaction mechanisms”. Nature Publishing Group UK London, 2023.

[57] J. D. Hirst *et al.*, "ML meets MLn: machine learning in ligand promoted homogeneous catalysis", *Artificial Intelligence Chemistry*, p. 100006, 2023.

[58] L. A. Baumes, P. Serna, y A. Corma, "Merging traditional and high-throughput approaches results in efficient design, synthesis and screening of catalysts for an industrial process", *Appl Catal A Gen*, vol. 381, núm. 1–2, pp. 197–208, 2010.

[59] A. Corma, J. M. Serra, P. Serna, y M. Moliner, "Integrating high-throughput characterization into combinatorial heterogeneous catalysis: unsupervised construction of quantitative structure/property relationship models", *J Catal*, vol. 232, núm. 2, pp. 335–341, 2005.

[60] K. S. Pratt, "Design Patterns for Research Methods: Iterative Field Research", en *AAAI Spring Symposium: Experimental Design for Real*, 2009, pp. 1–7.

[61] K. Petersen, S. Vakkalanka, y L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update", *Inf Softw Technol*, vol. 64, pp. 1–18, 2015.

[62] C. Páez, "¿Por qué es importante normalizar los conjuntos de datos?", DEV Community .

[63] Azure Machine Learning, "Componente Normalizar datos", Microsoft.

[64] D. C. Martínez R, C. A. Trujillo, J. G. Carriazo, y N. J. Castellanos, "Soybean Oil Epoxidation Catalyzed by a Functionalized Metal–Organic Framework with Active Dioxo-Molybdenum (VI) Centers", *Catal Letters*, vol. 153, núm. 6, pp. 1756–1772, 2023.

[65] Immune Technology Institute, "Librerías de Python, ¿qué son y cuáles son las mejores?" Consultado: el 6 de enero de 2024. [En línea]. Disponible en: https://immune.institute/blog/librerias-python-que-son/

[66] ID Digital School-Bootcamps, "¿Qué son las librerías de Python?" Consultado: el 6 de enero de 2024. [En línea]. Disponible en: https://iddigitalschool.com/bootcamps/que-son-las-librerias-de-python/

[67] scikit-learn developers, "3.3. Metrics and scoring: quantifying the quality of predictions". Consultado: el 7 de enero de 2024. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

[68] G. James, D. Witten, T. Hastie, y R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[69] L. Bottou, "Large-scale machine learning with stochastic gradient descent", en *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, Springer, 2010, pp. 177–186.

[70] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, y V. Vapnik, "Support vector regression machines", *Adv Neural Inf Process Syst*, vol. 9, 1996.

[71] T. Hastie, R. Tibshirani, J. H. Friedman, y J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[72] W. Loh, "Classification and regression trees", *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 1, núm. 1, pp. 14–23, 2011.

[73] G. Ke *et al.*, “Lightgbm: A highly efficient gradient boosting decision tree”, *Adv Neural Inf Process Syst*, vol. 30, 2017.

[74] J. Shawe-Taylor y N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[75] J. H. Friedman, “Greedy function approximation: a gradient boosting machine”, *Ann Stat*, pp. 1189–1232, 2001.

[76] D. J. C. MacKay, “Bayesian interpolation”, *Neural Comput*, vol. 4, núm. 3, pp. 415–447, 1992.

[77] H. Zou y T. Hastie, “Regularization and variable selection via the elastic net”, *J R Stat Soc Series B Stat Methodol*, vol. 67, núm. 2, pp. 301–320, 2005.

[78] J. Liu, S. Wu, y Z. Li, “Recent advances in enzymatic oxidation of alcohols”, *Curr Opin Chem Biol*, vol. 43, pp. 77–86, 2018.

[79] A. J. Howarth *et al.*, “Chemical, thermal and mechanical stabilities of metal–organic frameworks”, *Nat Rev Mater*, vol. 1, núm. 3, pp. 1–15, 2016.

[80] X.-L. Ni *et al.*, “Synthesis, characterization and catalytic performance of Mo based metal-organic frameworks in the epoxidation of propylene by cumene hydroperoxide”, *Chinese Chemical Letters*, vol. 28, núm. 5, pp. 1057–1061, 2017.

[81] Y. Liu *et al.*, “Bimetallic–Organic Framework as a Zero-Leaching Catalyst in the Aerobic Oxidation of Cyclohexene”, *ChemCatChem*, vol. 5, núm. 12, pp. 3657–3664, 2013.

[82] J. L. C. Rowsell y O. M. Yaghi, “Metal–organic frameworks: a new class of porous materials”, *Microporous and mesoporous materials*, vol. 73, núm. 1–2, pp. 3–14, 2004.

[83] K.-G. Liu, Z. Sharifzadeh, F. Rouhani, M. Ghorbanloo, y A. Morsali, “Metal-organic framework composites as green/sustainable catalysts”, *Coord Chem Rev*, vol. 436, p. 213827, 2021.

[84] G. Lewandowski y E. Milchert, “Parameters of epoxidation of cyclohexene by tert-butyl hydroperoxide”, *Ind Eng Chem Res*, vol. 40, núm. 11, pp. 2402–2408, 2001.

[85] S. Khare y S. Shrivastava, “Epoxidation of cyclohexene catalyzed by transition-metal substituted α-titanium arsenate using tert-butyl hydroperoxide as an oxidant”, *J Mol Catal A Chem*, vol. 217, núm. 1–2, pp. 51–58, 2004.