CARACTERIZACIÓN TEÓRICA DE LAS TÉCNICAS DE LA CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL IMPLEMENTADAS EN EL CAMPO DE LA CIENCIA E INGENIERÍA DE LOS MATERIALES: OPORTUNIDADES, DESCUBRIMIENTOS E INNOVACIÓN

ANDRÉS FELIPE SIERRA ÁLVAREZ

PROGRAMA DE INGENIERÍA QUÍMICA

DEPARTAMENTO DE INGENIERÍA AMBIENTAL, CIVIL Y QUÍMICA

FACULTAD DE INGENIERÍAS Y ARQUITECTURA

UNIVERSIDAD DE PAMPLONA

PAMPLONA, Diciembre 17 de 2021

CARACTERIZACIÓN TEÓRICA DE LAS TÉCNICAS DE LA CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL IMPLEMENTADAS EN EL CAMPO DE LA CIENCIA E INGENIERÍA DE LOS MATERIALES: OPORTUNIDADES, DESCUBRIMIENTOS E INNOVACIÓN

ANDRÉS FELIPE SIERRA ÁLVAREZ

Trabajo de monografía presentado como requisito para optar al título de INGENIERO QUÍMICO

Directora: ANA MARÍA ROSSO CERÓN

Doctora en Ingeniería Química

Co-tutora: SONIA ESPERANZA REYES GÓMEZ

Doctora en Ciencia e Ingeniería de Materiales

PROGRAMA DE INGENIERÍA QUÍMICA

DEPARTAMENTO DE INGENIERÍA AMBIENTAL, CIVIL Y QUÍMICA

FACULTAD DE INGENIERÍAS Y ARQUITECTURA

UNIVERSIDAD DE PAMPLONA

Pamplona, Diciembre 17 de 2021

Esta monografía está dedicada a mis padres

Per aspera ad astra

AGRADECIMIENTOS

Agradezco a mis pilares, mis padres, por siempre apoyarme en mis retos y pasiones. A mis compañeros y amigos Brandon Martínez y Edwin Sánchez, quienes siempre fueron los primeros en escuchar todas mis ideas. Por último, pero no menos importante, a mis directoras de trabajo, Ana María Rosso y Sonia Reyes, sin su conocimiento y guía no hubiese podido materializar mis ideas.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

GLOSARIO

AI: Inteligencia artificial

ML: Aprendizaje automático

DFT: Teoría del funcional de la densidad

AGI: Inteligencia artificial general

PCA: Análisis de componente principal

ICA: Análisis de componente independiente

ANN: Red neuronal artificial

kNN: k-Vecino más próximo

LCM: Materiales compuestos laminados

MDP: Procesos de decisión de Markov

1. INTRODUCCIÓN

Actualmente, se evidencia un crecimiento significativo en la toma de decisiones a partir de los datos y todas las ciencias e ingenierías, incluyendo la ingeniería química, se están transformando gracias a nuevas fuentes de datos procedentes de experimentos de alto rendimiento, estudios de observación y simulación. Esta nueva era ha ocasionado que, en la ciencia y la ingeniería basada en los datos, el descubrimiento ya no se limita a la recopilación y procesamiento de datos, sino más bien a la gestión de los mismos, la extracción y la visualización de la información obtenida de esa cantidad de datos (Beck, 2016).

La ciencia de datos utiliza enfoques científicos como la minería, limpieza y análisis exploratorio de los datos, así como, la ingeniería de características y la aplicación de diversos algoritmos de aprendizaje automático para comprender y visualizar esa enorme cantidad de datos (Virkus, 2019). La ciencia de datos es una convergencia entre el uso de la potencia del cálculo computacional, la estadística y, sobre todo, el conocimiento del área de dominio para identificar patrones significativos y relevantes.

La integración entre la ciencia de los materiales y la ciencia de datos tiene como finalidad descubrir la naturaleza que hay detrás de los fenómenos y la producción. La combinación de los datos de los materiales, la potencia de los cálculos computacionales y el conocimiento de dominio adentran a una nueva etapa de exploración y descubrimientos de materiales nuevos o alternativos (Austin, 2016).

Por consiguiente, el presente trabajo de monografía desarrolla una recolección bibliográfica con el objetivo de evidenciar de forma teórica las técnicas de la ciencia de datos e inteligencia artificial usadas en el campo de la ciencia e ingeniería de los materiales. Para esto, el desarrollo de la monografía se divide en varias secciones, dónde inicialmente aborda un marco conceptual para definir los campos y técnicas claves que se desglosan en la monografía; en la siguiente sección se expone de forma resumida las técnicas de modelado empleadas para la predicción de propiedades, que se desglosan en algoritmos de regresión y de clasificación. A continuación, se aborda el diseño de nuevos materiales con inteligencia artificial, así como el flujo de trabajo para el descubrimiento de nuevos materiales y algunos ejemplos de desarrollo de materiales con inteligencia artificial, luego, se mencionan métodos tradicionales para el diseño de nuevos materiales, resaltando el DFT y los experimentos de alto rendimiento (HT) y por último se

evidencia una comparación del aporte entre las técnicas de enfoque tradicional y las de enfoque basadas en datos.

2. JUSTIFICACIÓN

Los métodos tradicionales de desarrollo de materiales, cómo el método empírico de ensayo y error o el método basado en la teoría del funcional de densidad (DFT) presentan largos ciclos de desarrollo, baja eficiencia y elevados costos, esto representa una desventaja para el avance científico hoy en día, que tiene que adaptarse al constante crecimiento de las necesidades de la civilización. Por el contrario, las técnicas de la ciencia de datos como el aprendizaje automático presentan un bajo coste computacional y, en consecuencia, un corto ciclo de desarrollo (Y. Liu, 2017; Wei, 2019). Es entonces donde los avances del aprendizaje automático han repercutido en el área de la ciencia de los materiales con descubrimientos de nuevos materiales y las mejoras de simulaciones moleculares, anunciando así, un nuevo paradigma en la ciencia de los materiales (Morgan & Jacobs, 2020).

La ciencia de materiales puede resumirse en cuatro paradigmas: el primer paradigma es el método empírico de ensayo y error, el segundo paradigma son las leyes físicas y químicas, el tercer paradigma es la simulación por ordenador y el cuarto paradigma es la ciencia impulsada por los grandes datos (Wei, 2019).

En este campo, los datos son el nuevo recurso y el conocimiento se extrae de conjuntos de datos de materiales que son demasiado grandes o complejos para el razonamiento humano tradicional, por lo general con la intención de descubrir materiales o fenómenos en materiales nuevos o mejorados (Himanen, 2019). De la misma forma, el continuo desarrollo de la extracción, minería de datos y la inteligencia artificial, puede generar que este paradigma fácilmente unifique los otros paradigmas en sus aspectos teóricos.

Por consiguiente, la presente monografía busca realizar una identificación y descripción de las técnicas y métodos de la ciencia de datos e inteligencia artificial que son empleados en la ciencia e ingeniería de los materiales, cómo aporte a este nuevo paradigma con la finalidad de predecir propiedades y descubrir, desarrollar, diseñar e innovar en nuevos materiales o los ya existentes.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Caracterizar de forma teórica las técnicas de la ciencia de datos e inteligencia artificial que brindan oportunidades, nuevos descubrimientos e innovación en la ciencia e ingeniería de materiales.

3.2 OBJETIVOS ESPECÍFICOS

- Identificar las técnicas de la inteligencia artificial para predecir propiedades de los materiales.
- Describir las técnicas computacionales de la ciencia de datos e inteligencia artificial para el diseño de nuevos materiales.
- Determinar las técnicas y métodos convencionales usados para el diseño de nuevos materiales.
- Contrastar las técnicas y métodos convencionales usados para el diseño de nuevos materiales con las técnicas computacionales aportadas por la ciencia de datos e inteligencia artificial.

4. METODOLOGÍA

Se efectuó una investigación exploratoria con la finalidad de realizar un estudio descriptivo y comparativo en forma teórica de las técnicas de la ciencia de datos e inteligencia artificial que son empleadas en la ciencia e ingeniería de los materiales.

Inicialmente, se realizó una revisión exploratoria de la bibliografía en bases de datos cómo Science Direct, Springerlink, Scopus, Virtual Pro, SciELO y Google Scholar, proporcionadas en su mayoría por la Universidad de Pamplona, con la finalidad de adquirir la información necesaria para el desarrollo del documento. Luego de esto y con la información encontrada, se delimitó dicha información teniendo en cuenta los objetivos planteados previamente, clasificando así la información que ofreció mayor valor a los objetivos. Reconociendo que el campo aplicado a investigar se está usando mucho en la actualidad, se mantuvieron activas las alertas de las bases de datos para estar en constante actualización sobre las nuevas

investigaciones acerca del tema. Posteriormente, se organizó la bibliografía usando la aplicación Mendeley, la cual permitió gestionar las citas bibliográficas de manera efectiva y similar a los documentos de investigación. Por último, se analizó la información recopilada y se redactó el documento final, describiendo de forma generalizada las técnicas y modelos empleados en la ciencia de materiales, evidenciando estudios que emplean dichas técnicas para predicciones y diseño de nuevos materiales y exponiendo como los paradigmas de enfoque tradicional y enfoque basado en datos se complementan.

5. MARCO CONCEPTUAL

5.1 CIENCIA DE DATOS

Debido a que la ciencia de datos es un término relativamente nuevo, la construcción de su etimología sigue describiéndonos que es lo que se conoce bajo este término, sin embargo, todavía hay una diferencia significativa en lo que es la ciencia de datos. Por ende, la ciencia de los datos se ha denominado cómo:

- Un conjunto de disciplinas necesarias para resolver importantes retos en materia de datos. La tecnología, las personas y los datos son los tres pilares de la ciencia de datos (Song, 2016).
- Nueva disciplina científica que nos brinda estrategias, técnicas y métodos para recolectar índices informativos nuevos y existentes (Sajid, 2021).
- Práctica de recopilar datos, analizarlos dentro de un espacio de problemas, inferir en descubrimientos de patrones significativos y sacar conclusiones (Riveret, 2019).
- Descubrir patrones de datos interesantes y significativos utilizando métodos de análisis computacional (Riveret, 2019).

Dicho término se ha hecho cada vez más popular en la industria y en las disciplinas académicas para referirse a la combinación de estrategias y herramientas para hacer frente a la avalancha de datos y debido a que este campo requiere de conocimientos y habilidades de diferentes disciplinas, incluyendo las matemáticas, la estadística, ciencias de la computación y tecnologías de la información, el término *científico de datos* es una descripción común de un ingeniero o científico de cualquier disciplina que está equipado y capacitado para procesar, analizar y comunicar de manera efectiva en este contexto el uso intensivo de los datos. De este modo, las áreas principales de la ciencia de datos suelen ser la gestión de datos, el aprendizaje

automático y estadístico y la visualización de datos, con el soporte de todos los algoritmos de inteligencia artificial desarrollados e implementados para aportar a la generación de conocimiento (Beck, 2016).

5.2 INTELIGENCIA ARTIFICIAL

En términos generales, la inteligencia artificial (IA) puede definirse cómo la habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana, sin embargo, abordar el término requiere de una definición más detallada, pues Rouhiainen (2018) describe la IA cómo la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. La IA puede ser aplicada en muchas situaciones dónde la inteligencia humana tiene completa participación como, por ejemplo, el reconocimiento de imágenes estáticas, clasificación y etiquetado, en las mejoras del desempeño de la estrategia algorítmica comercial, el procesamiento eficiente y escalable de datos de pacientes, mantenimiento predictivo, detección y clasificación de objetos, entre otras. Asimismo, una de las orientaciones principales de la inteligencia artificial es el aprendizaje automático (Rouhiainen, 2018).

Comúnmente, la inteligencia artificial suele clasificarse en dos grandes categorías, las cuales se les conoce cómo generales y específicas.

5.2.1 Inteligencia artificial general (*AGI*)

Es generalmente una máquina que puede aprender a resolver cualquier problema que el intelecto humano pueda resolver. De igual forma, es conocida como inteligencia artificial "completa" o "fuerte", no obstante, actualmente es hipotética y la creación de la *AGI* es una meta importante para los investigadores (Chibani, 2021).

5.2.2 Inteligencia artificial específica

En contraste, la inteligencia artificial específica o también conocida como inteligencia artificial débil está centrada en la realización de tareas específicas previamente definidas (Chibani, 2021). Por esta razón, las técnicas mencionadas en nuestra investigación están catalogadas bajo el término de inteligencia artificial específica cómo, por ejemplo, el aprendizaje automático,

que en general ha sido una de las ramas de la inteligencia artificial que más ha arrojado éxitos en los últimos años, seguido del aprendizaje profundo.

5.3 APRENDIZAJE AUTOMÁTICO (*MACHINE LEARNING*)

Algunos estudios realizados por (Langley, 2018; Rouhiainen, 2018) indican que el aprendizaje automático, Machine Learning en inglés, es una de las perspectivas principales de la inteligencia artificial, puesto que permite a un sistema aprender de los datos en lugar de hacerlo mediante la programación explícita, o dicho de otra forma, se trata de un aspecto de la informática en el que las computadoras o máquinas tienen la capacidad de aprender sin estar programadas para ello, un resultado distintivo de esto serían las sugerencias o predicciones en situaciones particulares. No obstante, esto no hace al aprendizaje automático (ML por sus siglas en inglés) un proceso simple, dado que utiliza una serie de algoritmos que aprenden iterativamente de los datos, con la finalidad de mejorarlos, describirlos y predecir resultados finales, por lo tanto, a medida que los algoritmos ingieren datos de entrenamientos, es posible producir modelos más precisos basados en esos datos.

El aprendizaje automático puede dividirse entre modelos de aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado, dependiendo esto del tipo, la veracidad, capacidad y la cantidad de datos disponibles (Butler, 2018).

5.3.1 Aprendizaje supervisado

Este tipo de aprendizaje automático es también conocido como “aprendizaje con profesor”, dado que la acción humana infiere en el etiquetado de las salidas correspondientes de los datos de entrenamiento (Wei, 2019). El objetivo del algoritmo es derivar una función que, dado un conjunto específico de valores de entrada, prediga los valores de salida con un grado aceptable de fidelidad. Por consiguiente, los datos son previamente etiquetados u organizados para indicar cómo tendría que ser categorizada la nueva información (Butler, 2018; Rouhiainen, 2018). La mayoría de los artículos encasillan la regresión y la clasificación dentro de este tipo de aprendizaje; en la regresión el resultado es de forma numérica, mientras que en la clasificación el resultado es presentado de forma categórica. Algunos de estos posibles algoritmos son la estadística bayesiana, aprendizaje de árboles de decisión (*decision tree*), bosques aleatorios (*random forest*) (Mohri, 2018; Russell, 2019).

5.3.2 Aprendizaje no supervisado

En el aprendizaje no supervisado las salidas correspondientes a los datos de entrenamiento no están etiquetadas, si el conjunto de datos disponibles consiste únicamente en valores de entrada, se puede determinar el resultado cuando hay una cantidad masiva de datos (Langley, 2018) y utilizar el aprendizaje no supervisado para intentar identificar tendencias, patrones o agrupaciones en los datos (*clustering*) (Butler, 2018) con técnicas cómo k-medias (*k-means*), agrupación espectral (*spectral clustering*) o modelos de mezcla Gaussiana o por el contrario, con técnicas de reducción de dimensionalidad cómo el análisis de componentes independientes (*ICA*) o el análisis de componentes principales (*PCA*) (Sajid, 2021).

5.3.3 Aprendizaje reforzado

En el aprendizaje reforzado, en lugar de especificarle al modelo sobre cómo producir acciones correctas, se utilizan señales de refuerzo que proporcionan al ambiente o entorno para evaluar la calidad de las acciones generadas y mejorar las estrategias de adaptación al ambiente o entorno (Wei, 2019), dicho de otra forma, es una aprendizaje basado en la experiencia; interactuando con la dinámica del entorno o ambiente para maximizar una función de recompensa. En comparación con el aprendizaje supervisado, este no necesita disponer de pares de entrada y salida etiquetados previamente. Algunas de las técnicas usadas con este tipo de aprendizaje son los procesos de decisión de Markov (*MDP*) o los métodos de aprendizaje Q (*Q-learning methods*) (François-lavet, 2018; Mohri, 2018; Riveret, 2019).

5.4 CIENCIA DE MATERIALES

La ciencia de materiales se define cómo el estudio de las características y aplicaciones de los materiales; es reconocida como una disciplina bien establecida que combina la química, la física e investigación ingenieril, que se plantea cómo objetivo diseñar nuevos materiales desde cero para su uso en la sociedad (Himanen, 2019). En gran medida, el campo de la ciencia de

los materiales se basa en experimentos y modelos basados en simulación para comprender la física de diferentes materiales con la finalidad de entender mejor sus características y descubrir nuevos materiales con propiedades mejoradas. Últimamente, la generación de esta gran cantidad de datos a causa de los experimentos y simulaciones han ofrecido oportunidades sin precedentes para la aplicación de técnicas basadas en datos en este campo, abriendo nuevas vías para acelerar el descubrimiento y el diseño de los materiales (Agrawal, 2016).

Adicionalmente, la ciencia de materiales basada en datos (*data-driven*) y la informática de los materiales son términos que encierran la práctica científica de la extracción sistemática de conocimientos a partir de conjuntos de datos de los materiales, práctica que se diferencia de los enfoques científicos tradicionales en la investigación de los materiales por el volumen de datos procesados y la forma más automatizada de extraer la información. La ciencia basada en datos se anuncia como un nuevo paradigma en la ciencia de los materiales, en este campo los datos son los nuevos recursos y el conocimiento es extraído de estos conjuntos de datos de los materiales que son tan grandes o complejos para el razonamiento humano tradicional (Himanen, 2019).

5.5 EL CUARTO PARADIGMA EN LA CIENCIA DE MATERIALES

Desde hace miles de años, la ciencia tuvo un enfoque puramente empírico (primer paradigma), ejemplo de esto corresponde a las observaciones metalúrgicas a lo largo de las épocas de piedra, bronce, hierro y el acero. Luego de todo eso, llega el paradigma de los modelos teóricos (segundo paradigma), caracterizándose por la formulación de varias leyes presentadas en ecuaciones matemáticas; como por ejemplo las leyes de la termodinámica. No obstante, a pesar de contar con la formulación teórica de las leyes que brindan un entendimiento de los fenómenos de los materiales, con el tiempo los modelos teóricos se hicieron más complejos y, por ende, más difíciles de interpretar por lo que la solución analítica ya no era factible.

Es entonces que, con la aparición de los computadores, el tercer paradigma permitió realizar simulaciones de fenómenos complejos del mundo real basados en modelos teóricos del segundo paradigma, cómo la teoría del funcional de la densidad (DFT) y las simulaciones de dinámica molecular. Este paradigma ha contribuido al mismo tiempo a hacer avanzar los otros paradigmas. Sin embargo, la enorme cantidad de datos generada debido a los experimentos y simulaciones ha dado lugar al cuarto paradigma, la ciencia basada o impulsada por datos, unificando los tres primeros paradigmas e incrementando y popularizándose en el campo de la

ciencia de materiales, generando así, un nuevo y creciente campo que se conoce como la informática de los materiales (Agrawal, 2016), término que engloba la práctica científica de extracción sistemática de conocimientos a partir de conjuntos de datos sobre materiales (Himanen, 2019), La informática de los materiales ha desplegado un sinfín de técnicas dentro de las cuales se encuentran algunas capaces de predecir las propiedades de los materiales.

6. TÉCNICAS DE MODELOS PREDICTIVOS USADOS EN LA CIENCIA DE MATERIALES

En este capítulo se abarca el primer objetivo planteado, identificar las técnicas empleadas en la ciencia de materiales con la finalidad de predecir propiedades. Se evidencian tanto los modelos que son para regresión, los cuales arrojan un valor numérico; o los modelos que son de clasificación, los cuales arrojan una categoría. De igual forma, se describen algunas aplicaciones de dichas técnicas empleadas dentro de la ciencia de materiales y se contrastan con resultados de diferentes investigaciones donde se usaron modelos similares o modelos diferentes para la misma evaluación.

No obstante, el primer paso antes de cualquier modelado es entender el formato y la representación de los datos y realizar el pre procesamiento necesario para garantizar la calidad de los datos, lo que conlleva a eliminar o tratar adecuadamente el ruido, los valores atípicos, los valores perdidos, o los casos de datos duplicados que en ocasiones causan una mala ejecución del modelo.

Algunos ejemplos de pre procesamiento de datos incluyen la discretización, el muestreo, la normalización, la conversión de tipos de atributos, la extracción y selección de características, entre otros. Estos procesos de limpieza y homogeneización de los datos son un paso clave para construir modelos predictivos más precisos (Chibani, 2021).
Además, el pre procesamiento puede ser supervisado o no supervisado en función de si el proceso depende de los atributos deseados (Agrawal, 2016). Inmediatamente después que se ha realizado un pre procesamiento adecuado y los datos están listos para el modelado, se pueden emplear técnicas de minería de datos supervisadas para el modelado predictivo. Es necesario aclarar que la división adecuada entre los datos en conjuntos de entrenamiento y prueba debe

ser muy cuidadosa debido a que de lo contrario el modelo estará propenso a sobre ajustarse (*overfitting*) y mostrar una precisión demasiado óptima.

Si el atributo objetivo es numérico como, por ejemplo, la resistencia a la fatiga o la energía de formación, las técnicas de regresión pueden utilizarse para la elaboración de modelos predictivos (Mohri, 2018). Y si por el contrario es categórica, por ejemplo si un compuesto es metálico o no, se pueden usar técnicas de clasificación (Russell, 2019).

La **Tabla 1** evidencia técnicas de modelado con capacidad de regresión, clasificación o ensamblaje. Algunas de estas técnicas pueden ser usadas tanto para clasificación y regresión, ya que permiten trabajar con diferentes variables, ya sea numéricas o categóricas. Por otro lado, existen varias técnicas de aprendizaje por conjuntos que combinan los resultados de los aprendices base de diferentes maneras, y que demuestran mejorar la precisión y la solidez de los modelos en algunos casos (Agrawal, 2016) o dicho de otra forma, estos modelos de ensamblaje (*ensembles*) lo que hacen es utilizar otros sub modelos más pequeños para entrenar muchos de esos y luego hacer un voto entre estos modelos con la predicción de cada uno y promediar o tomar el más frecuente y con el resultado final hacer la predicción.

A continuación, en la **Tabla 1**, se observa una lista de algunas técnicas de modelado predictivo más conocidas.

Tabla 1 Algoritmos populares de modelos predictivos

Técnica de modelado	**Capacidad**	**Breve Descripción**
Algoritmos Naive Bayes (John & Langley, 2013)	Clasificación	Un clasificador probabilístico basado en el teorema de Bayes
Redes Bayesianas (Bouckaert, 2004)	Clasificación	Un modelo gráfico que codifica las relaciones condicionales probabilísticas entre variables
Regresión logística (Hosmer, 1989)	Clasificación	Ajusta los datos a una curva logística sigmoidal

Regresión lineal (Weher, 1977)	Regresión	Ajuste lineal por mínimos cuadrados de los datos con respecto a las características de entrada
Vecino más cercano (Aha, 1991)	Clasificación y regresión	Usa la instancia más similar en los datos de entrenamiento para hacer predicciones
Redes neuronales artificiales (Bishop, 1995; Fausett, 1994)	Clasificación y regresión	Usa capas ocultas de neuronas para conectar las entradas y las salidas, los contrapesos de los bordes se aprenden mediante retro-propagación
Máquinas de vectores de apoyo (Vapnik, 2000)	Clasificación y regresión	Sobre la base de la minimización del riesgo estructural, construye hiperplanos espaciales de características multidimensionales
Tabla de decisión (Kohavi, 1997)	Clasificación y regresión	Construye reglas que implican diferentes combinaciones de atributos
Decisión stump (Witten, 2016)	Clasificación y regresión	Modelo de aprendizaje automático basado en un árbol débil que consiste en un árbol de decisión de un solo nivel
J48 (C4.5) árbol de decisión (Salzberg, 1994)	Clasificación	Modelo de árbol de decisión que identifica el atributo de división en función de ganancia de información e impureza de Gini
Árbol de decisión alternativo (Mason, 2002)	Clasificación	Se compone de nodos de predicción y nodos de decisión alternados, una instancia recorre todos los caminos aplicables
Árbol de modelo logístico (Landwehr, 2005; Sumner, 2005)	Clasificación	Un árbol de clasificación con funciones de regresión logística en las hojas o bordes
Árbol de modelo M5 (Quinlan, 1992; Wang, 1997)	Regresión	Un árbol con función de regresión lineal en los nodos.

Árbol aleatorio (Quinlan, 1992)	Clasificación y regresión	Considera un subconjunto de atributos elegidos al azar
Árbol de reducción de cantidad de errores (Wang, 1997)	Clasificación y regresión	Construye un árbol utilizando la ganancia de información/varianza y lo reduce utilizando la depuración de errores para evitar el sobreajuste
AdaBoost (Freund, 1996)	Ensamblaje	El refuerzo puede reducir significativamente la tasa de error de un algoritmo de aprendizaje débil
Embolsado (Breiman, 1996)	Ensamblaje	Construye múltiples modelos a partir de subconjuntos de datos de entrenamiento con bootstrap para mejorar la estabilidad del modelo reduciendo la varianza
Subespacio aleatorio (Ho, 1998)	Ensamblaje	Construye múltiples árboles de forma sistemática mediante la selección pseudo-aleatoria de subconjuntos de características
Bosques aleatorios (Breiman, 1996)	Ensamblaje	Un conjunto de múltiples árboles aleatorios
Bosque de rotación (Rodriguez, 2006)	Ensamblaje	Genera conjuntos de modelos basados en la extracción de características seguida de rotaciones de ejes

Fuente: (Agrawal, 2016).

6.1 ALGORITMOS DE REGRESIÓN

Cómo se ha evidenciado, el objetivo de un algoritmo de aprendizaje automático es aprender a partir de los datos de entrenamiento y la función de mapeo de la entrada y la salida, por ejemplo, una estructura química y sus propiedades, donde los datos de entrada sería la estructura y los de salida las propiedades. Por lo tanto, el algoritmo tendrá la tarea de ser capaz de hacer una

predicción ante nuevos datos, con un nivel aceptable de precisión una vez esté entrenado (Chibani, 2021).

Dentro de los algoritmos mencionados en la **Tabla 1** se ilustran uno de los más sencillos que es la regresión lineal.

6.1.1 Regresión lineal

La regresión lineal es uno de los métodos más sencillos y potentes de regresión en el que la relación entre las variables explicativas y las objetivas se modela mediante una ecuación lineal. Un modelo de regresión lineal multivariante puede expresarse como:

$$y = \sum_{i}^{p} a_i x_i$$

Dónde y es la variable dependiente (u objetivo), x_i es la variable independiente y a_i es un coeficiente constante para las variables independientes x_i.

Shiraiwa (2018) empleó un modelo de regresión lineal para la predicción de la resistencia a la fatiga en aceros, indicó que si bien, un modelo más complejo proporciona una predicción más precisa en el conjunto de datos de entrenamiento, si el modelo es demasiado complejo la precisión del conjunto de datos de validación disminuye debido al sobreajuste (*overfitting*). Para evitar este problema y reducir el error en la validación, recurrió a la técnica de validación cruzada. Para esto se dividieron 360 muestras de entrenamiento, 10 muestras en 36 grupos. Se calcularon los coeficientes de regresión lineal para 35 grupos, excepto el primero y se predijo la resistencia a la fatiga del primer grupo utilizando dichos coeficientes calculados. Seguido, calcularon el error cuadrático medio de dicha predicción. Luego de esto, utilizaron 35 grupos, excepto el segundo, para calcular los coeficientes y el error cuadrático medio para los segundos grupos. Con la repetición del procedimiento anteriormente mencionado, se calcularon 36 valores de error cuadrático medio y dicho valor total del error cuadrático medio se registró cómo el error de la validación cruzada. Se encontró que, con todos los coeficientes normalizados, dos tratamientos térmicos mostraron que tienen una gran importancia en la resistencia a la fatiga y que además la resistencia a la fatiga aumenta con la adición de elementos de C, Mn y Cu, lo que se confirmaba con la teoría referente al fortalecimiento de la solución sólida y el endurecimiento por precipitación (Shiraiwa, 2018).

En la **Figura 1** se muestra el resultado de la predicción con la resistencia a la fatiga experimental.

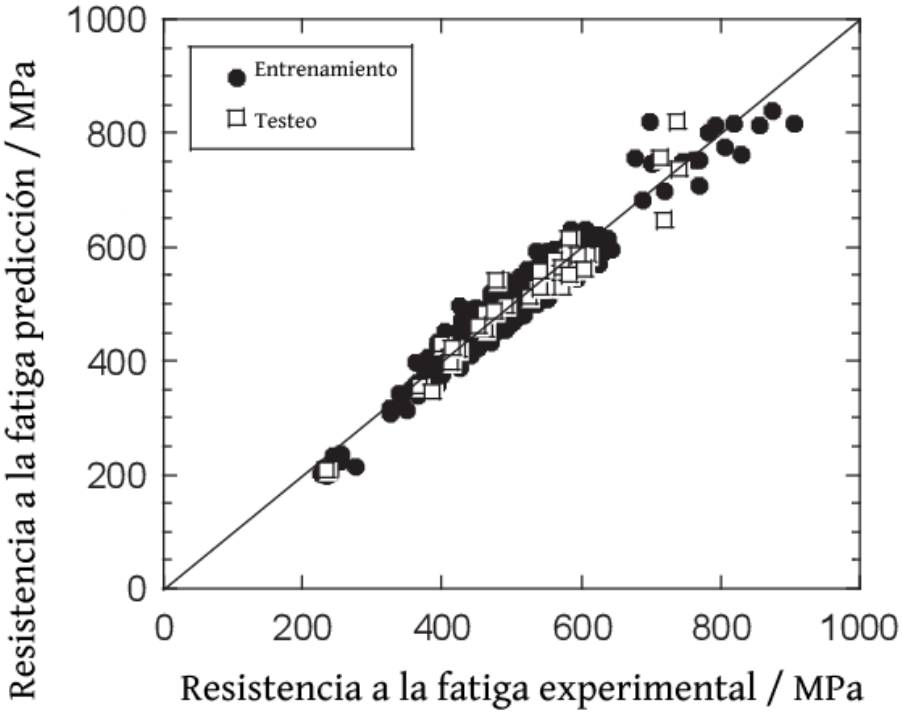

Figura 1. Gráfico de dispersión de la resistencia a la fatiga predicha por el modelo de regresión lineal.

Fuente: (Shiraiwa, 2018).

Por otra parte, Agrawal (2014) empleó modelos predictivos para predecir de igual forma la resistencia la fatiga en aceros. Los resultados mostraron que los modelos empleados (redes neuronales, árboles de decisión y regresión polinómica de multivariable) fueron más precisos para una gama amplia de materiales que el modelo de regresión lineal propuesto anteriormente.

6.1.2 Redes neuronales artificiales (*ANN*)

Las redes neuronales artificiales (ANN, por sus siglas en inglés) son una forma de aprendizaje automático que se inspira en las redes neuronales biológicas, se utilizan habitualmente para el modelado estadístico no lineal de datos para modelar relaciones complejas entre datos de entrada y salida (Agrawal, 2014). Sin embargo, este tipo de técnica de modelado puede usarse tanto de forma de regresión como de clasificación.

La red incluye una o más capas ocultas de múltiples neuronas artificiales conectadas a las entradas y salidas con diferentes pesos (*weight*). Los pesos internos de los bordes aprenden durante el proceso de entrenamiento mediante técnicas como la retropropagación (*Backpropagation*).

La arquitectura de la red neuronal artificial es la siguiente:

$$h_j = \varphi\left(\sum_k \omega_{jk} x_k + b_j\right)$$

Dónde x_k es la variable de entrada de la unidades de k-veces la capa de entrada, ω_{jk} y b_j son los pesos y los sesgos (*bias*), respectivamente; h_j es la variable de salida y φ es la función de activación.

En el estudio realizado por Shiraiwa (2018) el número de unidades en cada capa se fijó en diez unidades en la primera capa oculta y cinco unidades en la segunda capa oculta. En la capa de salida, usaron la función identidad $\varphi(x) = x$ como función de activación. Por otro lado, en las capas ocultas utilizaron tres tipos de funciones no lineales: la función sigmoide, la tangente hiperbólica, la cual se describe como una función más adecuada para la red que la función sigmoide, y por último la función ReLU o Unidad Lineal Rectificada.

$$\varphi(x) = \frac{1}{1 + e^{-x}}$$

$$\varphi(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$\varphi(x) = \max(0, x)$$

LeCun (2015) reportó que ReLU es la mejor función de activación.

El diseño de una red neuronal artificial se puede observar en la **Figura 2 (a)**. En la **Figura 2 (b)** se observa la representación gráfica de las funciones de activación usadas.

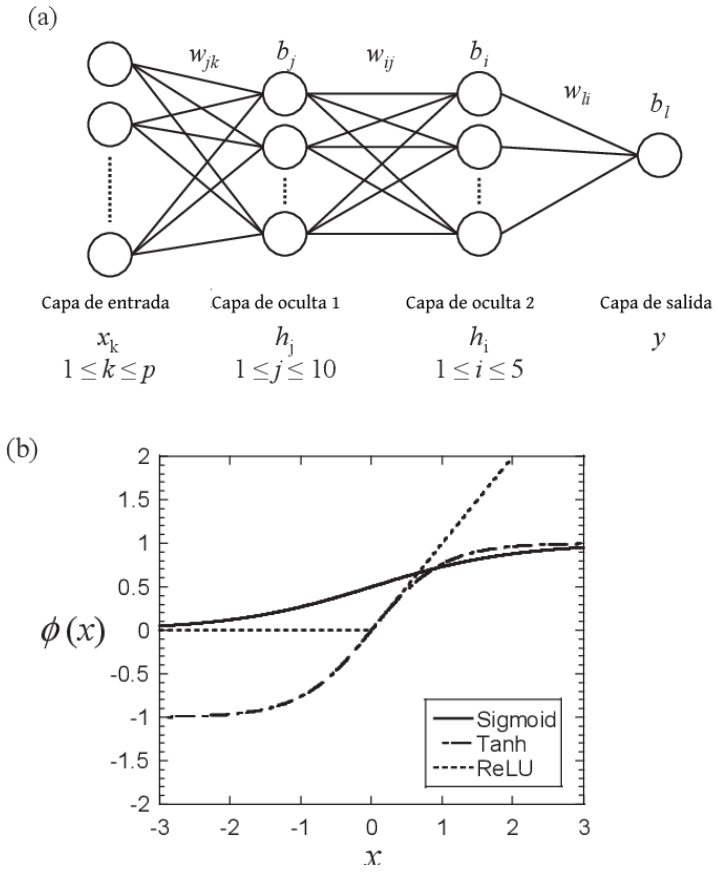

Figura 2 (a) Arquitectura de una red neuronal y (b) función de activación.
Fuente: (Shiraiwa, 2018)

El entrenamiento de la red neuronal consiste en establecer correctamente los pesos y los sesgos, generalmente usando el algoritmo de entrenamiento conocido como retropropagación. Inicialmente, se asignan valores aleatorios a los pesos y a los sesgos y el valor de salida se calcula utilizando estos pesos y sesgos iniciales. Una vez obtenido el valor de salida con los pesos y sesgos iniciales, se compara con el valor experimental y se actualiza el sesgo de la capa de salida y el peso de la capa oculta. Luego de esto, comparando el valor de la unidad oculta con el valor anterior, se actualiza el sesgo de la capa oculta y el peso de la capa anterior y de forma similar, se actualizan todos los pesos y sesgos. Un ciclo que actualiza los pesos y los valores en todas las capas se llama una época (*epoch*) (Shiraiwa, 2018).
De esta forma, el error en la capa de salida se propaga a la capa de entrada a través de la red cumpliendo el objetivo de minimizar el error.

En la **Tabla 2**, se muestran los resultados de los coeficientes de correlación y los del error cuadrático medio (RMSE), evidenciando que, las redes neuronales son más precisas en

comparación a la regresión lineal como menciona Agrawal (2014), sin embargo, este último no reporta la construcción de su red neuronal, por lo que no se tiene la constitución de sus capas ocultas y de salida. No obstante, presenta su valor de coeficiente de correlación y error cuadrático medio para los resultados de su red neuronal, siendo 0.9861 y 31.0545, respectivamente. Estos valores de coeficiente de correlación y error cuadrático son especificados únicamente para el testeo. Además, comparando los resultados obtenidos por Shiraiwa (2018) en la regresión lineal, nuevamente se evidencia en las redes neuronales artificiales, arroja que la temperatura de templado representa un atributo importante para predecir la resistencia a la fatiga, manifestado anteriormente en los resultados de Agrawal (2014).

Tabla 2. Coeficiente de correlación y error cuadrático medio (RMSE) del modelo de regresión lineal multivariante y del modelo de red neuronal artificial.

Modelo	Coeficiente de correlación	Error cuadrático medio (RMSE)
Entrenamiento Regresión Lineal	0.973	22.8
Testeo Regresión Lineal	0.968	23.0
Entrenamiento Red Neuronal Artificial	0.991	13.4
Testeo Red Neuronal Artificial	0.992	16.4

Fuente: (Shiraiwa, 2018).

6.1.3 Modelo de árboles M5

Los modelos de árboles M5 son una reconstrucción del algoritmo M5 de Quinlan (Quinlan, 1992; Wang, 1997) para inducir árboles de modelos de regresión, que combina un árbol de decisión convencional (*decision tree*) con la opción de funciones de regresión lineal en los nodos.

La construcción de un árbol modelo es similar a la del árbol de decisión. En primer lugar, se construye el árbol inicial y, a continuación, se poda (se reduce) para superar el problema del sobreajuste. Por último, se emplea el proceso de suavización (*smoothing*) para compensar las discontinuidades bruscas entre los modelos lineales adyacentes en las hojas del árbol podado (Solomatine, 2004).

En la **Tabla 3** se observan los valores de coeficiente de correlación y el error cuadrático medio para los modelos de árboles de M5 y redes neuronales empleados por Agrawal (2014). De la misma forma, Agrawal reporta que las técnicas empleadas son capaces de alcanzar una alta precisión predictiva, sin embargo, Shiraiwa (2018) reporta valores mayores en función de su predicción, con coeficientes de correlación en su modelo de red neuronal de 0.992. Cabe destacar que dentro del análisis predictivo de datos es cada vez más difícil aumentar la precisión de la predicción más allá de un determinado punto y que algunos modelos y evaluaciones estadísticas son más útiles que otros dependiendo de los tipos y tratamiento de datos que se empleen.

Tabla 3. Coeficiente de correlación y error medio cuadrático (RMSE) del modelo de árboles M5 y del modelo de red neuronal artificial.

Modelo	Coeficiente de correlación	Error cuadrático medio (RMSE)
Modelo de árboles M5	0.9890	27.6065
Red neuronal artificial	0.9861	31.0545

Fuente: Elaboración propia.

Los datos presentados en la **Tabla 3** se obtuvieron de Agrawal (2014).

6.2 ALGORITMOS DE CLASIFICACIÓN

Los algoritmos de clasificación sirven comúnmente para predecir una variable categórica, es decir, atribuir una etiqueta a los datos de entrada. En el caso más sencillo, estas categorías se reducen a dos en una variable binaria cómo, por ejemplo, si el material es conductor o aislante u otro ejemplo podría ser, si el material es poroso o no lo es (Chibani, 2021).

Se pueden utilizar un gran número de algoritmos de clasificación diferentes cómo lo son, la regresión logística, k-vecino más cercano (*kNN*), Naive Bayes y máquinas de vectores de apoyo (*SVM*), por mencionar algunos.

6.2.1 Algoritmos Naive Bayes

Los clasificadores Naive Bayes son un conjunto de algoritmos de clasificación basados en el teorema de Bayes que identifican la hipótesis más probable, dados los datos como conocimiento previo sobre el problema. El teorema de Bayes proporciona una manera formal de calcular la probabilidad de que una hipótesis sea correcta, dado un conjunto de datos existentes. A continuación, se pueden probar nuevas hipótesis y actualizar los conocimientos previos. De este modo, se puede seleccionar la hipótesis (o el modelo) con la mayor probabilidad de representar correctamente los datos (Bolstad, 2004; Butler, 2018).

De hecho, Addin (2007) indica que los clasificadores Naive Bayes en particular son uno de los sistemas de clasificación más exitosos para simular la detección de daños en materiales de ingeniería.
Usando clasificadores Naive Bayes y un método para la selección de subconjuntos basados en los valores medios y máximos de las amplitudes de ondas, el clasificador y el método de selección de subconjuntos de características se analizaron y probaron en dos conjuntos de datos. Los conjuntos de datos se llevaron a cabo sobre la base de daños artificiales creados en materiales compuestos laminados (*LCM*) cuasi isotópicos del sistema grafito/epoxi AS4/3501-6 y rodamientos de bolas de acero del tipo 6204 (Addin, 2007). Los materiales compuestos laminados (*LCM*) se fabrican apilando placas o capas de materiales compuestos para adquirir propiedades únicas (por ejemplo, alta resistencia y rigidez, y peso ligero) que no pueden ser garantizadas por los componentes individuales del laminado (Kessler, 2001).

Por otra parte, Liu (2014) empleó clasificadores Naive Bayes, *k*-vecinos próximos y red de funciones de base radial en un experimento en el que se utilizó un dedo robótico especialmente diseñado para reconocer materiales de la superficie de los objetos.
Los clasificadores de Naive Bayes mostraron una alta precisión de clasificación en los experimentos de Addin (2007), la mejor asignación de clasificación obtenida fue del 94.65%,

mientras que para Liu (2014) los resultados indicaron que la clasificación de Naive Bayes superó a los otros dos métodos de clasificación, con una tasa media de éxito del 84.2% (Asaduzzaman, 2021; Cai, 2020).

6.2.2 *K*-vecino más cercano (kNN)

En los métodos de *k*-vecino próximo o *k*-vecino más cercano, se calculan las distancias entre las muestras y los datos de entrenamiento en un hiperespacio descriptor. Llamados así porque el valor de salida de una predicción se basa en los valores de los *k* "vecinos más cercanos" de los datos, donde *k* es un número entero. Los modelos de vecinos más cercanos pueden utilizarse tanto en modelos de clasificación como de regresión. En la clasificación, la predicción viene determinada por la clase de la mayoría de los *k* puntos más cercanos; en la regresión, viene determinada por la media de los *k* puntos más cercanos (Shakhnarovich, 2005).

Nigsch (2006) aplicó la técnica de *kNN* a la predicción de puntos de fusión. Utilizando un conjunto de datos de 4119 moléculas orgánicas diversas (conjunto de datos 1) y un conjunto adicional de 277 drogas o fármacos (conjunto de datos 2) para comparar el rendimiento en diferentes regiones del espacio químico, además, investigó la influencia del número de vecinos más cercanos utilizando diferentes tipos de descriptores moleculares (2D o 3D).

Los resultados obtenidos evidenciaron que las predicciones para los fármacos resultaron ser considerablemente mejores, presentando un RMSE (expresado en °C) de 46.3 con un r^2 de 0.30, a diferencia de las predicciones basadas en no fármacos, presentando un RMSE (expresado en °C) de 50.3 con un r^2 de 0.30. No obstante, optimizaron mediante algoritmo genético los conjuntos de datos, esta vez obteniendo para el conjunto de datos 1 valores de RMSE (expresado en °C) de 46.2 con r^2 de 0.49 y para el caso del conjunto de datos 2, valores de RMSE (expresado en °C) de 42.2 con r^2 de 0.42.
La comparación de estos resultados demuestra que le método *kNN* introduce intrínsecamente un error sistemático en la predicción del punto de fusión (Nigsch, 2006). Por otra parte, Hansen (2013) utilizó 7165 muestras para la predicción de la atomización molecular.

6.2.3 Árboles de decisión

Los árboles de decisión son diagramas de flujo que se utilizan para determinar un curso de acción o un resultado. Cada rama del árbol representa una posible decisión, suceso o reacción. El árbol está estructurado para mostrar cómo y por qué una elección puede llevar a la siguiente, con ramas que indican que cada opción es mutuamente excluyente. Los árboles de decisión se componen de un nodo raíz, nodos hoja y ramas (Maimon, 2005).
El nodo raíz es el punto de partida del árbol. Tanto el nodo raíz como los nodos hoja contienen las preguntas o los criterios que hay que abordar. Las ramas son flechas que conectan los nodos, mostrando el flujo de la pregunta a la respuesta (Maimon, 2005).
Los árboles de decisión se utilizan a menudo en métodos de conjunto, que combinan varios árboles en un modelo de predicción para mejorar el rendimiento, cómo el modelo de árbol M5 usado en Agrawal (2014).

Fernandez (2014) empleó árboles de decisión para clasificar agregados de nanopartículas en una de las clases morfológicas (esferoidal, elipsoidal, lineal, ramificado). Usó carbón negro como caso de estudio, dado que este funciona como un nanorefuerzo utilizado en muchas industrias. La metodología consistió en tres pasos. Primero se realizó un procesamiento de imágenes de microscopía electrónica de transmisión para calcular un conjunto de 21 características morfológicas para cada agregado. Seguido de un análisis multivariante del conjunto de datos con la finalidad de reducir la dimensionalidad del problema y, por último, una creación y evaluación de árboles de decisión basados en los algoritmos evolutivos para clasificar agregados. Se observó que el modelo clasifica un agregado en una de las cuatro clases morfológicas en un sencillo proceso de comparación y que la precisión del modelo aplicado para clasificar nuevos agregados fue del 75% (Fernandez, 2014), evidenciando un modelo exitoso para determinar la forma de las partículas, lo que podría utilizarse consecuentemente para obtener una mejor comprensión de cómo el relleno afecta a la propiedades finales del material compuesto.

Dentro del marco aplicativo de la ciencia de datos e inteligencia artificial en la ciencia e ingeniería de los materiales, existen muchos modelos que, aunque no son completamente diseñado para, ya sea la predicción de propiedades o descubrimiento de materiales, pueden

ajustarse para ello, ejemplo de esto son los algoritmos usados para la predicción y descubrimientos de fármacos y nuevas drogas (Paul, 2021).
Por consiguiente, cada vez existen y se crean más técnicas que llegan a ser más robustas y, por ende, precisas. Algunos modelos de conjuntos (*ensembles*) utilizan varios modelos básicos cómo los mencionados anteriormente, logrando reducir la inestabilidad del modelo o el error, sin aumentar el sesgo del modelo (Jablonka, 2020).

7. TÉCNICAS DE DISEÑO DE NUEVOS MATERIALES CON INTELIGENCIA ARTIFICIAL

El desarrollo de este capítulo abarca el segundo objetivo planteado y examina el flujo de trabajo de diseño de nuevos materiales, evidenciando algunas de las áreas de desarrollo de materiales con la ayuda de la inteligencia artificial. A su vez, se discute sobre las técnicas tradicionales y cómo éstas están siendo complementadas y aceleradas con los modelos de aprendizaje automático, apremiando el proceso de descubrimiento y, por consiguiente, la generación de conocimiento.

No obstante, el número de estudios que exploran sistemáticamente diversas familias de materiales con el objetivo de descubrir materiales existentes con propiedades desconocidas o diseñar nuevos materiales con propiedades específicas se ha acelerado en la última década (Chibani, 2021).

Esto ha llevado a que las simulaciones moleculares hayan ampliado su escala, lo que permite a los científicos predecir la estructura y las propiedades de materiales complejos incluso antes de que sean sintetizados. No obstante, esto es una tarea difícil y larga, puesto que el papel tradicional de la computación en el diseño de materiales ha sido comprender mejor los materiales existentes (Hautier, 2012) y es entonces dónde este nuevo paradigma emergente, el enfoque impulsado por datos, acelera el descubrimiento de materiales diseñando nuevos compuestos *in silico* o en computadoras.

En la **Figura 3** se observa cómo la ciencia basada en datos surge como el cuarto paradigma científico tras la recopilación de las tres primeras eras de descubrimientos científicos impulsados por la experimentación, la teoría y la computación (Himanen, 2019).

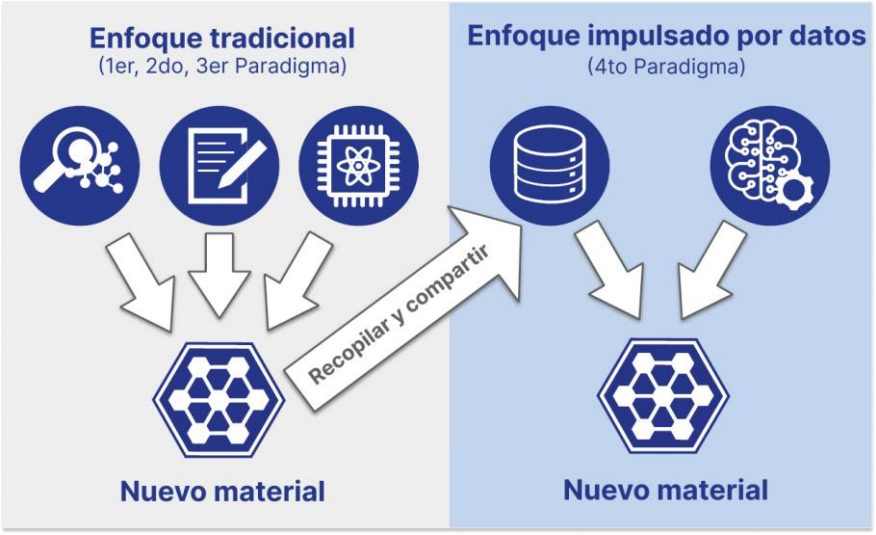

Figura 3. Esquema de descubrimiento de materiales.
Fuente: (Himanen, 2019)

Inicialmente, las iniciativas de datos de materiales comenzaron como bases que albergaban datos y ofrecían funciones de búsquedas, nos obstante, con la puesta en marcha de la Iniciativa del Genoma de los Materiales, MGI por sus siglas en inglés, marcó un momento decisivo para la ciencia de los materiales basadas en datos, debido a que las bases de datos evolucionaron hasta convertirse en centros de datos que ofrecían servicios de análisis de datos y materiales (Ward, 2012). Sin embargo, el interés emergente alrededor de la minería de datos y la IA hizo que los científicos cada vez apostaran más por estos algoritmos en sus investigaciones, en consecuencia, los centros de datos se enfocaron en el desarrollo de flujos de trabajo que permitieran a los científicos buscar, minar y consultar las bases de datos, migrando a la creación de infraestructuras que se han convertido en plataformas de descubrimiento de materiales (Himanen et al., 2019).

A pesar de tener una amplia cantidad de datos, producto del conocimiento generado por los paradigmas del enfoque tradicional y un crecimiento significativo en el número de proyectos e infraestructuras de la informática de los materiales (**Figura 4**), no todos los datos son óptimos para usar y los sistemas de descubrimiento de materiales requieren de un trabajo previo de los datos como se mencionó durante el desarrollo del documento.

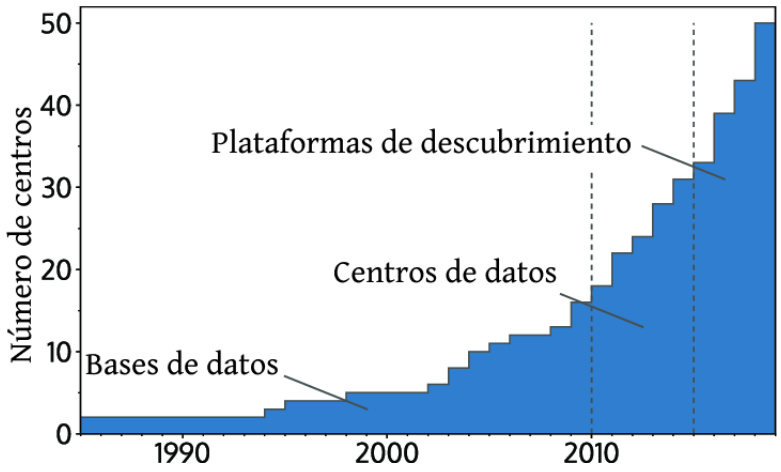

Figura 4. Número de proyectos e infraestructuras de informática de materiales en función del tiempo.

Fuente: (Himanen, 2019)

7.1 FLUJO DE TRABAJO DE APRENDIZAJE AUTOMÁTICO PARA EL DESCUBRIMIENTO DE NUEVOS MATERIALES

Los sistemas de aprendizaje automático para el descubrimiento de nuevos materiales incluyen dos partes: un sistema de aprendizaje y un sistema de predicción, que en conjunto crean todo un esquema escalable para varias familias de materiales.

El sistema de aprendizaje realiza las operaciones de limpieza de datos y selección de características, entrenamiento y prueba del modelo. El sistema de predicción aplica el modelo obtenido del sistema de aprendizaje para la predicción de componentes y estructuras (Y. Liu, 2017).

La implementación del sistema de aprendizaje, la extracción y construcción de los datos, puede ser obtenida de varias formas, una de estas incluye las bases de datos disponibles. La mayoría de estas bases de datos son accesibles tanto a través de una web, para fines de exploración y visualización sencilla y también por medio de una interfaz de programación de aplicaciones (*API*), la cual es una interfaz web con un comportamiento documentado, cuyas consultas y resultados son legibles por la máquina de acuerdo a un formato establecido (Chibani, 2021).

Estas *API* suelen ir acompañadas con una capa de software que facilita la integración en los proyectos. Un ejemplo de esto es el paquete *Python Materials Genomics* (pymatgen) (Ong, 2013) la cual es de código abierto y que se integra con la *API* RESTful o la Infraestructura y Base de Datos Interactiva Automatizada, también conocida como AiiDA (Pizzi, 2016).

Con los datos recolectados, se procede a la limpieza y a la selección de las características.
La limpieza es un paso crucial dado que permite posteriormente obtener un modelo de predicción eficiente y también ayuda para reducir la cantidad de cálculos. Por otro lado, la ingeniería o selección de las características es el proceso de extracción de las características más apropiadas de los datos y las tareas, el objetivo es obtener las características de los datos de entrenamiento para que los algoritmos puedan acercarse a su mejor rendimiento (Cai, 2020).

El sistema de aprendizaje se completa con el entrenamiento y prueba del modelo. Los pasos del modelado incluyen la selección de algoritmos adecuados, el entrenamiento a partir de los datos de entrenamiento y la realización de predicciones precisas (Wei et al., 2019).
En la **Figura 5** se evidencia un esquema generalizado del flujo de trabajo del aprendizaje automático, en la cual se muestra de dónde son obtenidos los datos para la realización y puesta en marcha del sistema de aprendizaje.

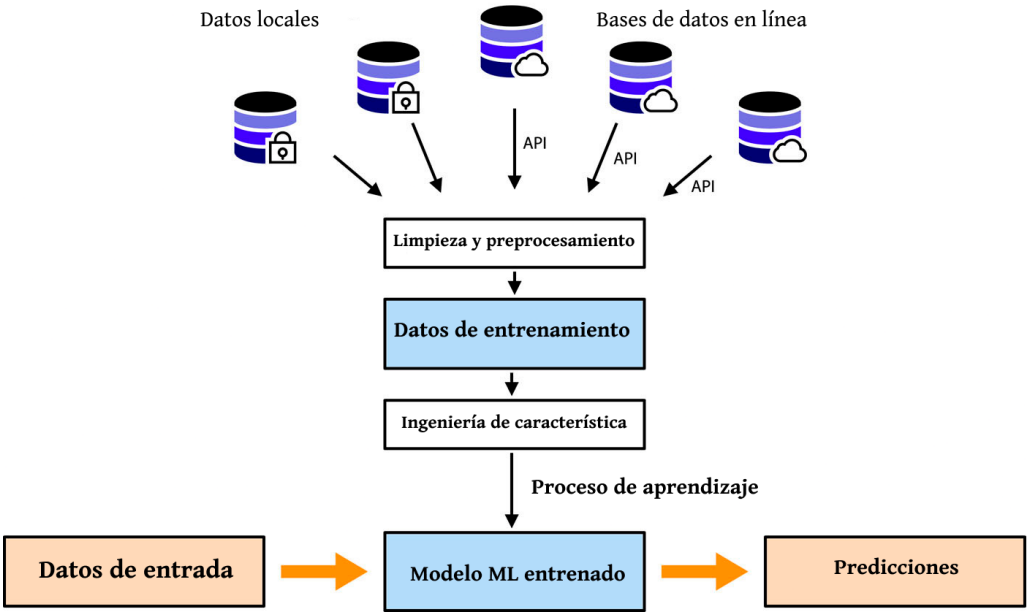

Figura 5. Flujo de trabajo de aprendizaje automático simplificado.
Fuente: (Chibani, 2021)

Los nuevos materiales suelen "predecirse" mediante un enfoque de sugerencia y prueba: el sistema de predicción selecciona las estructuras candidatas mediante la recomendación de la composición y la recomendación de la estructura, y los cálculos de DFT se utilizan para comparar su estabilidad relativa (Y. Liu, 2017).

En la **Figura 6** se observa a detalle el proceso general del aprendizaje automático en el descubrimiento de nuevos materiales, exponiendo cada etapa en el sistema de aprendizaje y en el sistema de predicción.

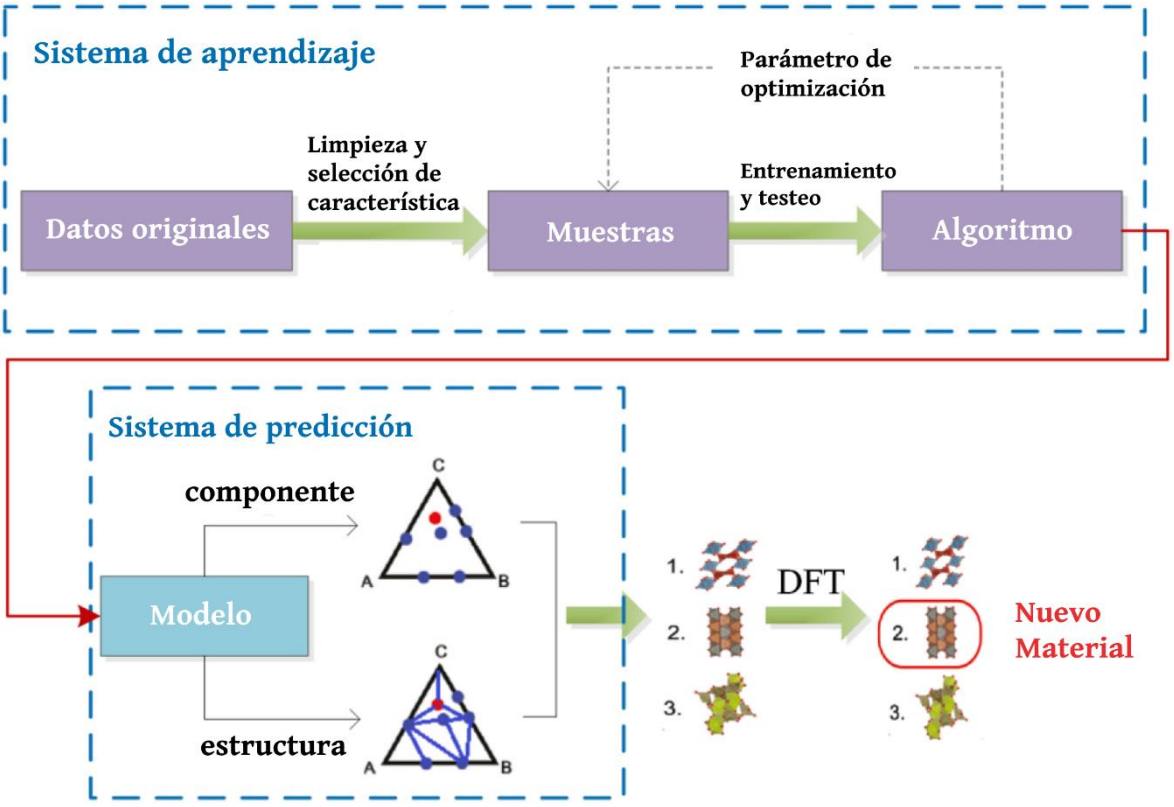

Figura 6. Proceso general de aprendizaje automático en el descubrimiento de nuevos materiales. Fuente: (Y. Liu, 2017).

7.2 DESARROLLO DE MATERIALES CON AYUDA DE LA INTELIGENCIA ARTIFICIAL

7.2.1 Diseño inverso para compuestos deseados

El objetivo del diseño inverso es encontrar materiales con funcionalidades particulares o materiales deseados. El método toma una funcionalidad como entrada y da como resultado una

estructura molecular (Wei, 2019), partiendo de las propiedades deseadas y termina en el espacio químico, es posible considerar que este proceso parte de condiciones específicas para encontrar posibles soluciones entre una amplia gama de materiales candidatos y combinaciones de materiales. Por el contrario, el diseño tradicional hacia adelante consiste en obtener los materiales objetivo a través de experimentos y luego juzgar las funcionalidades de los materiales.

Una idea básica del diseño inverso tiene su origen en el cribado virtual de alto rendimiento. Por ejemplo, Sánchez-Lengeling (2018) exploró a fondo el método de diseño inverso. Exhibía como los modelos generativos profundos se han aplicado a numerosas clases de materiales como por ejemplo, el diseño racional de posibles fármacos, rutas sintéticas de compuestos orgánicos y optimización de la energía fotovoltáica y las baterías de flujo redox, así como también a una variedad de otros materiales de estado sólido.

No obstante, aclara que desde el punto de vista actual, la idea de investigación y el método de aplicación del diseño inverso son bastante diversos; sin embargo, todavía no están maduros. Algunos procedimientos específicos, como la representación digital de las moléculas, la selección de los métodos de *ML* y el diseño de las herramientas de diseño inverso, todavía necesitan más estudio (Sánchez-Lengeling, 2018).

7.2.2 Visión computacional para el análisis de imágenes de materiales

La visión por computadora es básicamente un sistema de *IA* que puede extraer información de imágenes o datos multidimensionales. En la ciencia de los materiales, la visión por ordenador puede analizar propiedades de materiales poco claras o desconocidas a partir de enormes cantidades de imágenes, lo que puede ayudar en gran medida a los científicos a comprender las propiedades físicoquímicas y las relaciones internas de materiales similares (Cai, 2020).

Algunos ejemplos son la detección de la corrosión de las vigas de hormigón de las vías férreas mediante el análisis de las imágenes de los rieles (Gibert, 2017); la exploración de las morfologías de las partículas y las texturas de la superficie de los polvos mediante el análisis de las imágenes de la microestructura de los polvos para los procesos de fabricación aditiva

(DeCost, 2017). En un caso similar, Bastidas-Rodriguez (2016) utilizó *ANN* y máquinas de vectores de soporte (*SVM*) para clasificar la fractura de materiales metálicos para el análisis de fallas. Los resultados arrojaron qué el rendimiento de la aplicación de *ANN* era ligeramente superior a la de *SVM* y el porcentaje de precisión más alto fue del 84,95%.

7.2.3 Cribado de alto rendimiento y big data en el descubrimiento de materiales

El cribado de alto rendimiento en el campo del descubrimiento de nuevos materiales utiliza enormes volúmenes de datos para realizar tareas computacionales con el fin de detectar las propiedades de los materiales y diseñar los materiales objetivo. El Big Data puede definirse como un método de investigación que extrae información y detecta relaciones a partir de conjuntos de datos extraordinariamente grandes (Philip Chen, 2014).
Los investigadores recopilan grandes volúmenes de datos sobre los materiales objetivo y utilizan el cribado de alto rendimiento para analizar las propiedades de los materiales o la posibilidad de sintetizarlos. Teniendo en cuenta la necesidad de datos a la hora de aplicar el aprendizaje automático (ML), estos dos métodos se han convertido literalmente en los cimientos del campo del descubrimiento de nuevos materiales (Cai, 2020).

Existen algunos casos que pueden demostrar la importancia de estos dos métodos, por ejemplo, Thomas (2011) realizó un cribado de alto rendimiento para estimar la seguridad, la nanotoxicología, la ecotoxicología, la evaluación medioambiental y otras propiedades de los nanomateriales diseñados. En otro estudio también se utilizó el cribado de alto rendimiento basado en DFT para estimar la actividad de más de 700 aleaciones superficiales binarias con el fin de encontrar un electrocatalizador adecuado para la reacción de evolución del hidrógeno, e identificaron con éxito el BiPt como el material objetivo deseado (Greeley, 2006).

Por su parte, Agrawal (2016) describió sistemáticamente el importante papel que evidentemente desempeñan al día de hoy el Big Data en la informática de los materiales.

En general, además del ML, otros algoritmos de IA se utilizan ampliamente para el descubrimiento de nuevos materiales y mejoras de los existentes. De igual forma, se evidencian numerosos casos donde se desenvuelve la importancia, los efectos y las perspectivas de

desarrollo de la IA. A pesar de que la IA tiene fuertes efectos y un futuro promisorio en la ciencia de los materiales, todavía necesita un mayor desarrollo (Cai, 2020).

8. MÉTODOS DE ENFOQUE TRADICIONAL PARA EL DISEÑO DE NUEVOS MATERIALES

Este capítulo se enfoca en el tercer objetivo, métodos convencionales involucrados en el proceso de diseño de un nuevo material, por consiguiente, esboza los métodos fundamentales y nos exhibe aplicaciones logradas con los mismos. Por otra parte, estos métodos no son ajenos a la inteligencia artificial ni a su puesta en marcha con aprendizaje automático, debido a que la son implementados de manera conjunta y el cuarto paradigma, el enfoque basado en datos, ha ayudado al crecimiento de los métodos de enfoque tradicional cómo se evidencia en la **Figura 7**.

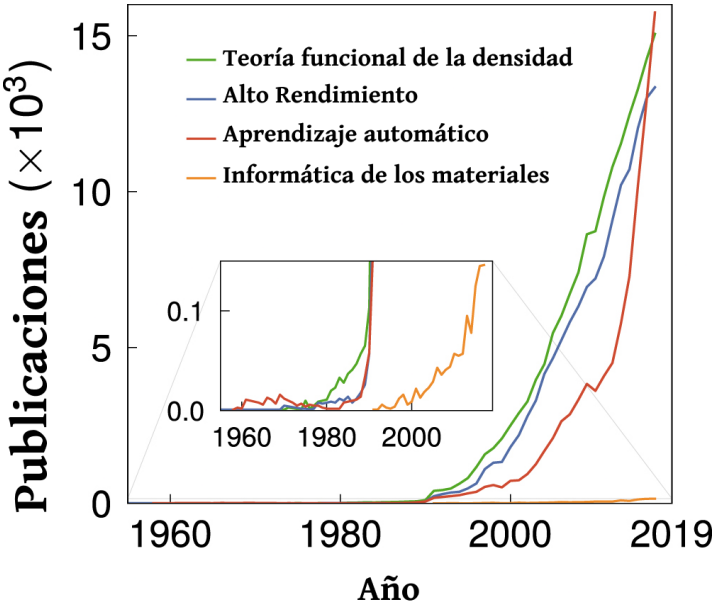

Figura 7. Evolución cronológica del número de publicaciones sobre Teoría Funcional de la Densidad (DFT), Alto Rendimiento (HT), Aprendizaje automático (ML) e informática de materiales.

Fuente: (Schleder, 2019)

Teniendo en cuenta el desarrollo histórico de la investigación en la ciencia computacional de los materiales, podemos clasificar los diferentes problemas y métodos utilizados para abordarlos en dos generaciones.

La primera generación está relacionada con la obtención de propiedades de los materiales dada su estructura, utilizando algoritmos de optimización local, normalmente basados en cálculos de DFT realizados de uno en uno. Sigue siendo el enfoque más extendido, debido a las grandes mejoras que permiten los cálculos de alto rendimiento a gran escala. La segunda generación está relacionada con la predicción de la estructura cristalina dada una composición fija, utilizando tareas de optimización global como los algoritmos genéticos y evolutivos. Este enfoque requiere la realización de un número considerable de cálculos de forma sistemática, por lo que depende en gran medida de los métodos de alto rendimiento (Butler, 2018; Schleder, 2019).

8.1 TEORÍA FUNCIONAL DE LA DENSIDAD (*DFT*)

La DFT se fundamenta en el teorema de Hohenberg-Kohn (H-K), que establece que, primero, todas las propiedades del estado base de un sistema, incluyendo la energía total, son una función de la densidad de carga del estado base; y segundo, la densidad de carga correcta del estado base minimiza la función de energía. El teorema de H-K implica que el estado base de cualquier sistema puede determinarse variando la densidad de carga hasta encontrar el mínimo global en el funcional de energía (Hautier, 2012).
Los cálculos DFT proporcionan un método fiable para estudiar los materiales una vez que se conoce la estructura cristalina o molecular. Basándose en el teorema de Hellman-Feynman (Feynman, 1939), se pueden utilizar los cálculos de DFT para encontrar un mínimo estructural local de materiales y moléculas. Sin embargo, la optimización global de estos sistemas es un proceso mucho más complicado.

Desde su desarrollo inicial, la DFT ha evolucionado desde unos cálculos limitados capaces de proporcionar resultados aproximados hasta una metodología cada vez más precisa y predictiva, lo que ha dado lugar a importantes contribuciones en diversas áreas como el descubrimiento y

diseño de materiales, el diseño de fármacos, las células solares, los materiales para la separación del agua, etc. (Schleder, 2019).

8.2 EXPERIMENTOS DE ALTO RENDIMIENTO (*HT*)

Debido a que, a medida que avanza el tiempo y la capacidad de cálculo aumenta drásticamente, los experimentos de alto rendimiento traducen en una importante reducción del tiempo empleado para realizar los cálculos, por lo que aplica un tiempo relativamente mayor a la configuración y el análisis de las simulaciones. Esto dio lugar a un cambio en el flujo de trabajo teórico, brindando nuevas estrategias de investigación; puesto que, en lugar de realizar muchas simulaciones preparadas manualmente, ahora se puede automatizar la creación de entradas y realizar varias simulaciones en paralelo o simultáneamente (Nosengo, 2016).
La idea es generar y almacenar grandes cantidades de propiedades termodinámicas y electrónicas ya sea mediante simulaciones o experimentos para materiales existentes o hipotéticos, y luego realizar el descubrimiento o la selección de materiales con las propiedades deseadas a partir de estas bases de datos (Curtarolo, 2013). Este enfoque no implica necesariamente el aprendizaje automático, sin embargo, hay una tendencia creciente a combinar estas dos metodologías en la ciencia de materiales.

Además, es importante destacar que el enfoque de alto rendimiento es compatible con las metodologías teóricas, computacionales y experimentales. Por otra parte, el principal obstáculo de un método determinado es el tiempo necesario para realizar un solo cálculo o medición, por lo que el motor de alto rendimiento tiene que ser rápido y preciso para producir cantidades masivas de datos en un tiempo razonable.

El Instituto de Física de la Academia China de las Ciencias desarrolló un método experimental único de alto rendimiento basado en el concepto de ingeniería del genoma de los materiales, que supuso un gran avance en el diseño de la composición y la exploración de las aleaciones amorfas de alto rendimiento, y realizó un rápido cribado de las aleaciones amorfas, y desarrolló un nuevo sistema de materiales de aleación amorfa de alta temperatura y alta resistencia (Li, 2019).

9. COMPARACIÓN ENTRE EL ENFOQUE TRADICIONAL Y EL ENFOQUE BASADO EN DATOS

Durante el desarrollo de este capítulo se evidencia el cuarto objetivo, contrastar los métodos empleados en el enfoque tradicional y los usados en el enfoque basado en datos. Esto, conlleva a notar que los modelos basados en datos no son comparables con las técnicas desarrolladas en el enfoque tradicional sino, complementarios, debido a que; en primera instancia, la diferencia radica en la cantidad de datos y la capacidad de procesamiento para poder descubrir características no conocidas de los materiales y, en segunda instancia, la validación de los resultados obtenidos se fundamenta en los paradigmas de enfoque tradicional.

Debido a que desafortunadamente, los estudios repetitivos de caracterización experimental y teórica suelen ser lentos e ineficientes, el rango de tiempo para descubrir nuevos materiales es notablemente largo, en algunas ocasiones entre 10 o 20 años desde la investigación inicial hasta el primer día de uso (Y. Liu, 2017).

En la **Figura 8** se observa que la investigación de nuevos materiales comprende siete etapas discretas, específicamente el descubrimiento, el desarrollo, la optimización de las propiedades, el diseño y la integración del sistema, la certificación, la fabricación y el despliegue, y las diferentes etapas pueden ser incluso llevadas a cabo por diferentes equipos de ingeniería o científicos (Ward, 2012).

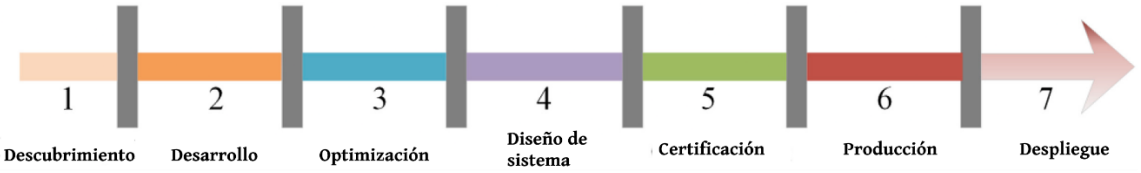

Figura 8. Proceso de búsqueda de nuevos materiales con métodos tradicionales.
Fuente: (Ward, 2012).

Es bien sabido que la simulación computacional y la medición experimental son dos métodos convencionales ampliamente adoptados en el campo de la ciencia de los materiales. Sin embargo, es difícil utilizar estos dos métodos para acelerar el descubrimiento y el diseño de materiales debido a las limitaciones implícitas tanto a las condiciones experimentales como a los fundamentos teóricos (Y. Liu, 2017).

En comparación con la medición experimental, la simulación computacional requiere menos tiempo y es ventajosa para suplir los experimentos reales, ya que se tiene un control total sobre las variables relevantes. Sin embargo, la simulación computacional también presenta muchos retos; por ejemplo, depende en gran medida de las microestructuras de los materiales implicados; requiere equipos de computación de alto rendimiento, normalmente en grandes clusters de computación en los que se pueden ejecutar los programas de simulación computacional y, no se pueden utilizar explícitamente los resultados de cálculos anteriores cuando se estudia un nuevo sistema por lo que no es escalable y en algunas ocasiones, automatizable (Y. Liu, 2017).

Por ende, en los últimas dos décadas, las actividades computacionales relacionadas con ciencia de los materiales se han ido desplazando constantemente desde el desarrollo de técnicas y los estudios puramente computacionales de los materiales hacia el descubrimiento y el diseño de nuevos materiales guiados por los resultados computacionales, el aprendizaje automático y la minería de datos o por la estrecha colaboración entre las predicciones computacionales y la validación experimental (Chen, 2015; Hautier, 2012).

Las ventajas de las estrategias modernas de investigación de materiales residen en su capacidad para encontrar un buen equilibrio entre unos requisitos experimentales razonables y una baja tasa de error, para aprovechar al máximo los amplios datos disponibles y para acelerar el proceso de investigación de materiales. Es aquí donde la colaboración entre lo teórico, lo computacional y lo basado en datos pone en proceso las capacidades de generar nuevo conocimiento de manera exponencial como se evidencia en la **Figura 7.**

10. CONCLUSIONES

La identificación de las técnicas de inteligencia artificial para predecir propiedades de los materiales nos brinda un panorama inicial del proceso de aplicación de estos modelos y nos exhibe ejemplos de posibles usos y resultados como los presentados. Los modelos descritos y sus aplicaciones a la ciencia de materiales mostraron ser efectivos en su función predictiva, resaltando que dichos modelos descritos son los básicos o los de aprendizaje base y que los modelos de conjunto (*ensembles*) presentan mejor *performance* predictivo, pero, aumentan su complejidad.

La descripción de técnicas de ciencia de datos e inteligencia artificial para el diseño de nuevos materiales muestran un amplio espectro de aplicación dentro de este campo, brindando oportunidades de uso en entornos locales de investigación y trayendo la oportunidad de seguir generando nuevos conocimientos, descubrimientos e innovación en el desarrollo de nuestras investigaciones.

El uso de técnicas convencionales no son desplazadas dentro del flujo de trabajo de diseño de nuevos materiales, por el contrario, se soportan en las nuevas técnicas basadas en datos para acelerar el proceso desde el descubrimiento hasta el despliegue y puesta en marcha de un nuevo material.

Las técnicas empleadas tanto en el enfoque tradicional como en el enfoque basados en datos, no son comparables entre sí, son complementarias. Esto se debe a que el cuarto paradigma acelera la producción y el descubrimiento de nuevos materiales, encontrando puntos de inflexión de características no conocidas en materiales existentes y validando los resultados de los modelos de predicción o descubrimiento para conocer su capacidad aplicativa, basado en las técnicas de enfoque tradicional.

11. RECOMENDACIONES

La ciencia basada en datos (*data-driven*) sigue siendo un paradigma creciente y cada vez influye más en las decisiones del entorno científico e ingenieril, sin embargo, esto trae consigo retos muy grandes; como la capacidad de manipular conjuntos de datos cada vez más robustos y complejos, y a su vez, la capacidad de extraer información más específica y puntual de esas enormes cantidades de datos. Por consiguiente, estas habilidades serán de nuestra competencia cada vez más y un buen punto de partida es aprovechar las herramientas de código abierto (*open source*) que apoyan la gestión, el aprendizaje estadístico y automático y la visualización.

12. REFERENCIAS BIBLIOGRÁFICAS

Addin, O., Sapuan, S. M., Mahdi, E., & Othman, M. (2007). A Naïve-Bayes classifier for damage detection in engineering materials. *Materials and Design*, *28*(8), 2379–2386. https://doi.org/10.1016/j.matdes.2006.07.018

Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials*, *4*(5). https://doi.org/10.1063/1.4946894

Agrawal, A., Deshpande, P. D., Cecen, A., Basavarsu, G. P., Choudhary, A. N., & Kalidindi, S. R. (2014). *Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters*. 1–19.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66. https://doi.org/10.1007/BF00153759

Asaduzzaman, M., Hossain, N., Ahmed, B., Fotouhi, M., Islam, S., Ali, R., & Abul, M. (2021). Recent machine learning guided material research - A review. *Computational Condensed Matter*, *29*(June), e00597. https://doi.org/10.1016/j.cocom.2021.e00597

Austin, T. (2016). Towards a digital infrastructure for engineering materials data. *Materials Discovery*, *3*, 1–12. https://doi.org/10.1016/J.MD.2015.12.003

Bastidas-Rodriguez, M. X., Prieto-Ortiz, F. A., & Espejo, E. (2016). Fractographic classification in metallic materials by using computer vision. *Engineering Failure Analysis*, *59*, 237–252. https://doi.org/https://doi.org/10.1016/j.engfailanal.2015.10.008

Beck, D. A. C., Carothers, J. M., Subramanian, V. R., & Pfaendtner, J. (2016). Data Science: Accelerating Innovation and Discovery in Chemical Engineering. *AIChE Journal*, *62*. https://doi.org/10.1002/aic

Bishop, C. M., Bishop, P. N. C. C. M., Hinton, G., & Press, O. U. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press. https://books.google.com.co/books?id=-aAwQO%5C_-rXwC

Bolstad, W. (2004). Introduction to Bayesian Statistics. In *University of Waikato*. John Wiley.

Bouckaert, R. (2004). *Naive Bayes Classifiers That Perform Well with Continuous Variables* (Vol. 3339). https://doi.org/10.1007/978-3-540-30549-1_106

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Butler, K. T., & Daniel, W. (2018). Machine learning for molecular and materials science. *Nature*. https://doi.org/10.1038/s41586-018-0337-2

Cai, J., Chu, X., Xu, K., Li, H., & Wei, J. (2020). *Machine learning-driven new material discovery*. 3115–3130. https://doi.org/10.1039/d0na00388c

Chen, L.-Q., Chen, L.-D., Kalinin, S. V, Klimeck, G., Kumar, S. K., Neugebauer, J., & Terasaki, I. (2015). Design and discovery of materials guided by theory and computation. *Npj Computational Materials*, *1*(1), 15007. https://doi.org/10.1038/npjcompumats.2015.7

Chibani, S., & Coudert, F. (2021). *Machine learning approaches for the prediction of materials properties*. *080701*(August 2020). https://doi.org/10.1063/5.0018384

Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S., & Levy, O. (2013). The high-throughput highway to computational materials design. *Nature Materials*, *12*(3), 191–201. https://doi.org/10.1038/nmat3568

DeCost, B. L., & Holm, E. A. (2017). Characterizing powder materials using keypoint-based computer vision methods. *Computational Materials Science*, *126*, 438–445.

https://doi.org/https://doi.org/10.1016/j.commatsci.2016.08.038
Fausett, L., & Fausett, L. V. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall. https://books.google.com.co/books?id=ONylQgAACAAJ
Fernandez, R., Okariz, A., Ibarretxe, J., Iturrondobeitia, M., & Guraya, T. (2014). Use of decision tree models based on evolutionary algorithms for the morphological classification of reinforcing nano-particle aggregates. *COMPUTATIONAL MATERIALS SCIENCE*, *92*, 102–113. https://doi.org/10.1016/j.commatsci.2014.05.038
Feynman, R. P. (1939). Forces in Molecules. *Phys. Rev.*, *56*(4), 340–343. https://doi.org/10.1103/PhysRev.56.340
François-lavet, S. C. V., Henderson, P., Islam, R., François-lavet, V., Pineau, J., & Bellemare, M. G. (2018). *An Introduction to Deep Reinforcement Learning*. https://doi.org/10.1561/2200000071.Vincent
Freund, Y., & Hill, M. (1996). *Experiments with a New Boosting Algorithm*.
Gibert, X., Patel, V. M., & Chellappa, R. (2017). Deep Multitask Learning for Railway Track Inspection. *Trans. Intell. Transport. Sys.*, *18*(1), 153–164. https://doi.org/10.1109/TITS.2016.2568758
Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. B., & Nørskov, J. K. (2006). Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature Materials*, *5*(11), 909–913. https://doi.org/10.1038/nmat1752
Hansen, K., Biegler, F., Fazli, S., Rupp, M., Sche, M., Lilienfeld, O. A. Von, Tkatchenko, A., & Mu, K. (2013). *Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies*.
Hautier, G., Jain, A., & Ping, S. (2012). *From the computer to the laboratory : materials discovery and design using first-principles calculations*. https://doi.org/10.1007/s10853-012-6424-0
Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science*, *6*(21). https://doi.org/10.1002/advs.201900808
Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601
Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley and Sons.
Jablonka, K. M., Ongari, D., Moosavi, S. M., & Smit, B. (2020). Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews*, *120*(16), 8066–8129. https://doi.org/10.1021/acs.chemrev.0c00004
John, G., & Langley, P. (2013). Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, *1*.
Kessler, S., Spearing, S., Atalla, M., & Cesnik, C. (2001). Structural health monitoring in composite materials using frequency response methods. *Proceedings of SPIE - The International Society for Optical Engineering*, *4336*. https://doi.org/10.1117/12.435552
Kohavi, R. (1997). The Power of Decision Tables. *Proceedings of European Conference on Machine Learning*. https://doi.org/10.1007/3-540-59286-5_57
Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1), 161–205. https://doi.org/10.1007/s10994-005-0466-3
Langley, P., Carbonell, J. G., & IBM. (2018). Machine Learning For Dummies. In *Journal of the American Society for Information Science* (Vol. 35, Issue 5). https://doi.org/10.1002/asi.4630350509

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, M.-X., Zhao, S.-F., Lu, Z., Hirata, A., Wen, P., Bai, H.-Y., Chen, M., Schroers, J., Liu, Y., & Wang, W.-H. (2019). High-temperature bulk metallic glasses developed by combinatorial methods. *Nature*, *569*(7754), 99–103. https://doi.org/10.1038/s41586-019-1145-z

Liu, H., Bimbo, J., Seneviratne, L., Althoefer, K., & Mary, Q. (2014). *Surface material recognition through haptic exploration using an intelligent contact sensing finger*. *October 2012*. https://doi.org/10.1109/IROS.2012.6385815

Liu, Y., Zhao, T., Ju, W., Shi, S., Shi, S., & Shi, S. (2017). Materials discovery and design using machine learning. *Journal of Materiomics*, *3*(3), 159–177. https://doi.org/10.1016/j.jmat.2017.08.002

Maimon, O., & Rokach, L. (2005). *Data Mining And Knowledge Discovery Handbook*.

Mason, L. (2002). The Alternating Decision Tree Learning Algorithm. *Proc. 16th International Conference on Machine Learning*, *99*.

Mohri, M. (2018). *Foundations of Machine Learning* (F. Bach (ed.); 2nd ed.). The MIT Press.

Morgan, D., & Jacobs, R. (2020). Opportunities and Challenges for Machine Learning in Materials Science. *Annual Review of Materials Research*, *50*, 71–103. https://doi.org/10.1146/annurev-matsci-070218-010015

Nigsch, F., Bender, A., Buuren, B. Van, Tissen, J., Nigsch, E., & Mitchell, J. B. O. (2006). *Melting Point Prediction Employing k -Nearest Neighbor Algorithms and Genetic Parameter Optimization*. 2412–2422.

Nosengo, N. (2016). Can artificial intelligence create the next wonder material? *Nature*, *533*(7601), 22–25. https://doi.org/10.1038/533022a

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*, 314–319. https://doi.org/https://doi.org/10.1016/j.commatsci.2012.10.028

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/https://doi.org/10.1016/j.ins.2014.01.015

Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N., & Kozinsky, B. (2016). AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science*, *111*, 218–230. https://doi.org/https://doi.org/10.1016/j.commatsci.2015.09.013

Quinlan, J. R. (1992). *Learning With Continuous Classes*.

Riveret, R., Gao, Y., Governatori, G., Rotolo, A., Pitt, J., & Sartor, G. (2019). A probabilistic argumentation framework for reinforcement learning agents: Towards a mentalistic approach to agent profiles. In *Autonomous Agents and Multi-Agent Systems* (Vol. 33, Issues 1–2). Springer US. https://doi.org/10.1007/s10458-019-09404-2

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630. https://doi.org/10.1109/TPAMI.2006.211

Rouhiainen, L. (2018). Inteligencia artificial 101 cosas que debes saber. *Alienta Editorial*, 352. https://planetadelibrosar0.cdnstatics.com/libros_contenido_extra/40/39307_Inteligencia_artificial.pdf

Russell, S. J., & Norvig, P. (2019). *Artificial Intelligence A Modern Approach* (4th ed., Vol. 1).

Sajid, S., Haleem, A., Bahl, S., Javaid, M., Goyal, T., & Mittal, M. (2021). Data science applications for predictive maintenance and materials science in context to Industry 4.0. *Materials Today: Proceedings*, *45*(xxxx), 4898–4905. https://doi.org/10.1016/j.matpr.2021.01.357

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, *16*(3), 235–240. https://doi.org/10.1007/BF00993309

Sánchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, *361*, 360–365.

Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., & Fazzio, A. (2019). From DFT to machine learning: Recent approaches to materials science - A review. *JPhys Materials*, *2*(3). https://doi.org/10.1088/2515-7639/ab084b

Shakhnarovich, G., & Darrell, T. (2005). Nearest-Neighbor Methods in Learning and Vision: Theory and Practice. In *The MIT Press*.

Shiraiwa, T., Miyazawa, Y., & Enoki, M. (2018). *Prediction of Fatigue Strength in Steels by Linear Regression and Neural Network*. *60*(2), 189–198.

Solomatine, D. P., & Xue, Y. (2004). *M5 Model Trees and Neural Networks : Application to Flood Forecasting in the Upper Reach of the Huai River in China*. *9*(6). https://doi.org/10.1061/(ASCE)1084-0699(2004)9

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Sumner, M., Frank, E., & Hall, M. (2005). Speeding Up Logistic Model Tree Induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. https://doi.org/10.1007/11564126_72

Thomas, C. R., George, S., Horst, A. M., Ji, Z., Miller, R. J., Peralta-Videa, J. R., Xia, T., Pokhrel, S., Mädler, L., Gardea-Torresdey, J. L., Holden, P. A., Keller, A. A., Lenihan, H. S., Nel, A. E., & Zink, J. I. (2011). Nanomaterials in the environment: from materials to high-throughput screening to organisms. *ACS Nano*, *5*(1), 13–20. https://doi.org/10.1021/nn1034857

Vapnik, V. (2000). The Nature of Statistical Learning Theory. In *Statistics for Engineering and Information Science* (Vol. 8, pp. 1–15). https://doi.org/10.1007/978-1-4757-3264-1_1

Virkus, S., & Garoufallou, E. (2019). Data science from a library and information science perspective. *Data Technologies and Applications*, *53*, 442–441. https://doi.org/10.1108/DTA-05-2019-0076

Wang, Y., & Witten, I. (1997). Induction of model trees for predicting continuous classes. *Induction of Model Trees for Predicting Continuous Classes*.

Ward, C. (2012). *Materials Genome Initiative for Global Competitiveness* (Issue June).

Weher, E. (1977). An introduction to linear regression and correlation. (A series of books in psychology.). *Biometrical Journal*, *19*(1), 83–84. https://doi.org/https://doi.org/10.1002/bimj.4710190121

Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019). Machine learning in materials science. *Wiley Interdisciplinary Reviews*, *1*(3), 338–358. https://doi.org/10.1002/inf2.12028

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science. https://books.google.co.uk/books?id=1SylCgAAQBAJ