INTRODUCCIÓN

En la actualidad, estamos presenciando una serie de transformaciones significativas en múltiples áreas como la medicina, la economía, la defensa y el comercio, entre otras, como resultado de la incorporación de la analítica de datos. Estas transformaciones han estado en crecimiento debido al desarrollo tecnológico y a la aparición de diversos algoritmos agrupados bajo el concepto de *Machine Learning* o aprendizaje automático, que se puede definir como una subdisciplina de la inteligencia artificial que se basa en abordar y solucionar problemas a partir de disciplinas numéricas como el razonamiento probabilístico, la investigación basada en la estadística, la recuperación de información y el reconocimiento de patrones; de esta forma, las máquinas o computadoras mediante la ejecución de los algoritmos, hacen tareas que comúnmente realizan los humanos [1].

El *Machine Learning* es considerado una de las ramas importantes de la inteligencia artificial en constante evolución debido al desarrollo de algoritmos computacionales diseñados para emular la inteligencia humana al aprender del entorno que los rodea [2]. Por otra parte, la inteligencia artificial como un campo más amplio dentro de la informática, se enfoca en el desarrollo de programas con la capacidad de realizar acciones inteligentes o demostrar inteligencia en diversas tareas. En este contexto, el *Machine Learning* o aprendizaje automático, representa un subcampo esencial de la inteligencia artificial [3].

Otro concepto fundamental, relacionado con los mencionados anteriormente, es el conocido como *Big Data,* que se refiere a la gestión y análisis de conjuntos

de datos extremadamente grandes. El auge de esta disciplina se debe en parte a la evolución de los métodos utilizados para analizar estos datos. Inicialmente, se empleaban enfoques estadísticos, posteriormente se incorporaron los algoritmos de Minería de Datos (*Data Mining*), y con el avance continuo de la ciencia, las ciencias de la computación y la creciente presencia de la inteligencia artificial se han aplicado cada vez más algoritmos de *Machine Learning*.

Este libro se estructura en capítulos que abordan los pasos esenciales para adquirir y preparar un conjunto de datos, lo que posteriormente permite al analista realizar una selección más efectiva de las características influyentes y aplicar algoritmos supervisados y/o de ensamble para predecir o pronosticar una variable de salida específica.

El Capítulo 2 proporciona una visión general del *Machine Learning*, abarcando su clasificación, algoritmos y herramientas relacionadas. El Capítulo 3 proporciona un panorama de los datos educativos y cómo han sido estudiados por diversas herramientas tecnológicas.

El Capítulo 4 se centra en el proceso inicial de instalación del entorno Jupyter (Python), junto con sus instrucciones y código básico, permitiendo al lector adentrarse en el mundo del Aprendizaje Automático. En el Capítulo 5, abarca la fuente de datos (variables y factores) así como la etapa de preparación de datos, que comprende la limpieza y el análisis de la base de datos deseada, utilizando herramientas estadísticas, tablas y gráficos.

El Capítulo 6 aborda las transformaciones de datos, cuyo objetivo es llevar la distribución de los datos hacia una forma cuasi gaussiana mediante técnicas como el reescalado, la estandarización y la normalización. El Capítulo 7 presenta diferentes métodos utilizados para obtener los mejores resultados posibles de un modelo predictivo, enfatizando la necesidad de realizar una cuidadosa selección de características que puedan influir de manera óptima en la variable de respuesta.

Es relevante mencionar que parte del contenido, como código y definiciones, puede ser similar a lo que se encuentra en blogs e internet, ya que este libro ha tenido en cuenta alguna información divulgada en estos medios. Algunos enlaces de consulta se encuentran en la bibliografía.

COMPLEMENTO EN CÓDIGO QR

En el interior del libro el lector encontrará diversos QR con material complementario.

1

EXPLORANDO EL APRENDIZAJE AUTOMÁTICO

El *Machine Learning* (ML) o aprendizaje automático se define como una subdisciplina de la inteligencia artificial que se apoya en enfoques numéricos como el razonamiento probabilístico, la estadística, la recuperación de información y el reconocimiento de patrones para abordar y resolver problemas. En esencia, mediante la ejecución de algoritmos, las máquinas realizan tareas que suelen ser llevadas a cabo por seres humanos [4]. Otra definición ampliamente aceptada por la comunidad científica describe el ML como el diseño y estudio de programas computacionales que utilizan experiencias pasadas para tomar decisiones futuras [5].

Por otro lado, el término *Big Data* se refiere a la gestión de conjuntos de datos voluminosos que son difíciles de procesar con métodos tradicionales [1]. Su creciente relevancia se atribuye, en parte, a la evolución de los métodos utilizados para analizar estos datos. Inicialmente, se empleaban enfoques estadísticos, y posteriormente se adoptaron los algoritmos de minería de datos. Sin embargo, con el avance de la ciencia, las ciencias de la computación y la aparición de la inteligencia artificial (como se ilustra en la Figura 1), que busca emular las capacidades humanas para realizar tareas complejas, se han desarrollado tres campos principales: la robótica (que se enfoca en replicar movimientos humanos), el procesamiento de lenguaje natural (que se centra en texto y voz) y el *Machine Learning* (que se especializa en datos). Este último hace uso de algoritmos que pueden predecir, prevenir o mejorar procesos, incluso en el ámbito académico de los estudiantes [5], y se aplica en diversos sectores, incluyendo la educación [6].

La inteligencia artificial también es considerada un área de la informática, pero tiene que ver con el desarrollo de programas que sean inteligentes o puedan hacer cosas inteligentes, por tanto, el *Machine Learning* o aprendizaje automático es

un subcampo de esta que consiste básicamente en buscar un modelo o procedimiento que haga el mejor uso de los datos históricos compuestos por entradas y salidas para predecir hábilmente salidas nuevas a partir de entradas nuevas que no eran conocidas [6].

Los campos de la minería de datos y el aprendizaje automático están entrelazados. La minería de datos utiliza algoritmos de aprendizaje automático para interrogar grandes bases de datos y descubrir conocimientos ocultos en los datos, mientras que muchos algoritmos de aprendizaje automático emplean métodos de minería de datos para reprocesar los datos antes de aprender las tareas deseadas. Sin embargo, debe tenerse en cuenta que el aprendizaje automático no se limita a resolver problemas similares a las de las bases de datos, sino que también se extiende a la resolución de desafíos complejos de inteligencia artificial aprendiendo y adaptándose a una situación que cambia dinámicamente, como por ejemplo prácticas de oncología y radioterapia [7].

En conclusión, *Machine Learning* es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante y se les considera el caballo de batalla en la nueva era de los llamados *Big Data*. Las técnicas basadas en el aprendizaje automático se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, las finanzas, el entretenimiento y la biología computacional hasta las aplicaciones médicas y biomédicas [7].

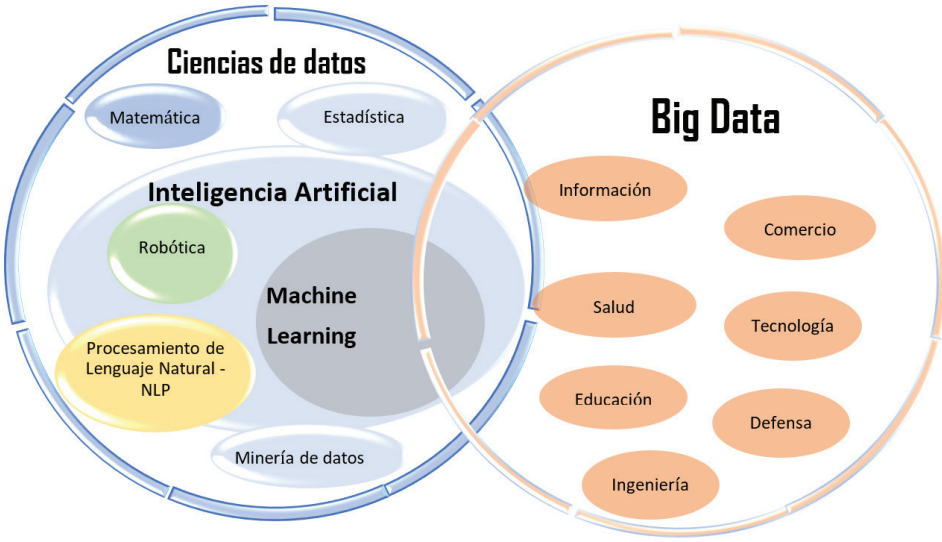

Figura 1. Tendencia en herramientas de análisis de datos

1.1 UNA INTRODUCCIÓN A LA CLASIFICACIÓN DE MACHINE LEARNING

El *Machine Learning* se convierte en una herramienta valiosa para comprender la clasificación y, posteriormente, poder seleccionar los algoritmos más adecuados en función del problema que se busca resolver. Comprender estos tipos de problemas en *Machine Learning* es esencial para seleccionar el enfoque adecuado y aplicar los algoritmos correspondientes según las necesidades específicas del problema que se esté abordando. En el ámbito del *Machine Learning*, se encuentran varios tipos de problemas clásicos:

- Clasificación: este tipo de problemas involucra múltiples características que influyen en una variable dependiente. Los datos se encuentran etiquetados, lo que significa que se asigna una clase o categoría a la variable dependiente, por ejemplo, "vive" o "no vive", "aprueba" o "no aprueba". El objetivo en este tipo de tarea es que el modelo pueda asignar una etiqueta a un nuevo dato no visto previamente por el modelo, por ejemplo, determinar si "aprueba" o "no aprueba".

- Regresión: en contraste con la clasificación, en los problemas de regresión se busca predecir un valor numérico en lugar de una etiqueta. Los datos se etiquetan con un valor real o flotante. Un ejemplo sería predecir el rendimiento académico de un estudiante, donde la respuesta es un número dentro de un rango, como entre 0 y 5.

- Agrupación: en este tipo de problemas, los datos no están etiquetados, es decir, no hay una variable dependiente definida. En su lugar, el objetivo es dividir el conjunto de datos original en grupos basados en la similitud y otras medidas de estructura inherente en los datos.

- Extracción de reglas: aquí, los datos se utilizan como base para extraer reglas proposicionales, a menudo conocidas como reglas if-then (si-entonces). Se aplican, por ejemplo, cuando se lanza un nuevo producto al mercado y se desea determinar si "si un cliente compra el producto X, entonces comprará el producto Y". Estas reglas pueden involucrar atributos de los datos que no necesariamente están relacionados con lo que se está prediciendo.

La forma más general para empezar a comprender el campo del *Machine Learning* es por medio de la Figura 2. A continuación, se procede a explicar de manera breve cuáles son cada uno de los campos o subáreas.

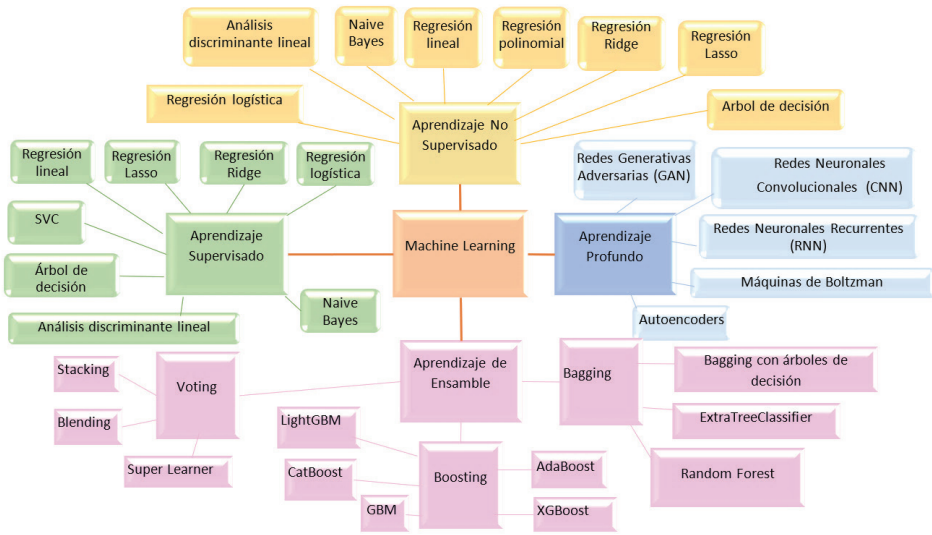

Figura 2. Líneas de Machine Learning

Dentro del campo de aprendizaje clásico, el *Machine Learning* se encuentra dividido en aprendizaje supervisado y no supervisado (Figura 3). El aprendizaje supervisado consiste en predecir los valores de un conjunto de datos de salida, a partir de un conjunto de datos de entrada. Se le llama supervisado porque conforme el modelo predice las salidas para datos de prueba, se calcula el error entre lo que predijo el algoritmo y el valor real. El objetivo es minimizar el error, ajustando la función de densidad de probabilidad que relaciona las entradas con las salidas. Es decir, son aquellos algoritmos donde el desarrollador actúa como un mediador para enseñar al algoritmo cuales son las respuestas a la que debe llegar, esto lo hace a través de unos datos de entrenamiento.

El segundo subtipo de la clasificación del aprendizaje automático o *Machine Learning* llamado aprendizaje no supervisado, ocurre cuando un algoritmo aprende de ejemplos simples sin ninguna respuesta asociada, dejando que el algoritmo determine los patrones de datos por sí mismo, sin un humano (no se le enseña cuales son las respuestas de salida) [9]. Es decir, organiza los datos según algunos grupos de características que puedan representar una clase. Cuando se tiene un dato nuevo, este tipo de algoritmo lo agrupará en un conjunto de datos que tenga las características más parecidas al dato inicial suministrado (no es una predicción). En esta categoría, los modelos de ML realizan principalmente agrupaciones o clasificaciones de los datos, según las variables de entrada y su diferenciación. Son modelos con métricas no muy precisas y que en ocasiones dependen de ciertas heurísticas para validar sus resultados [8].

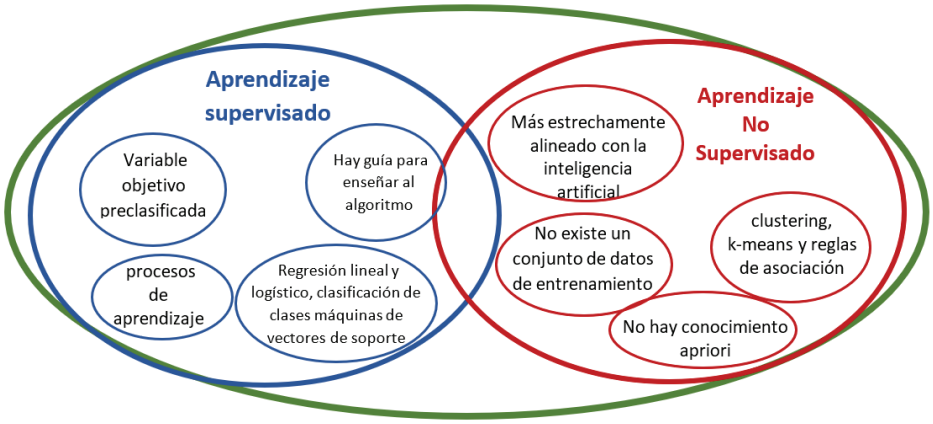

Figura 3. Clasificación del aprendizaje automático o Machine Learning

Existen dentro del campo diversos algoritmos utilizados para realizar predicciones sobre diversos valores de salida o clases, los cuales son clasificados según el tipo de variable de salida. Cuando la variable de salida es categórica se denominan algoritmos supervisados para tareas de clasificación. El otro grupo lo constituyen los algoritmos cuya variable de salida es numérica; se denominan algoritmos supervisados para tareas de regresión (Figura 4).

Aprendizaje supervisado	Algoritmos de clasificación	Regresión logística
		Análisis discriminante lineal
		KNN
		Naive Byes
		Arbol de decisión
		SVC
		Perceptrón
	Algoritmos de regresión	Regresión lineal
		Regresión polinomial
		Regresión Lasso
		Regresión Ridge
		KNN
		SVR
		Arbol de decisión

Figura 4. Métodos de Machine Learning, aprendizaje supervisado

En la otra rama conocida como aprendizaje no supervisado, se han desarrollado investigaciones con un enfoque diferente. Aquí, el objetivo no es la predicción de datos, sino más bien la exploración de relaciones entre los datos. En la actualidad, existen una variedad de algoritmos diseñados para abordar este tipo de tareas, entre los que se incluyen K-means, K-medoids, Fuzzy C-means, C-means, Hierarchical Clustering, Latent Dirichlet Allocation, Principal Component Analysis, Gaussian Mixture, Hidden Markov Model, Apriori, etc.

Una de las ramas que son consideradas contemporáneas es la de los métodos o algoritmos de ensamble, también conocidos como métodos de conjunto, los cuales representan una vertiente crucial en el campo del *Machine Learning*. Estos métodos se fundamentan en la idea de combinar diversos algoritmos para trabajar en conjunto y así reducir los errores que cada uno de ellos podría cometer si operaran de manera individual. De esta manera, se logra elevar la calidad general del sistema por encima de lo que podría conseguir incluso el mejor de los algoritmos por sí solo. Dentro de esta categoría se incluyen técnicas como el Stacking, el Bagging y el Boosting, que desempeñan un papel esencial en la mejora del rendimiento y la precisión de los modelos de *Machine Learning* (Figura 5).

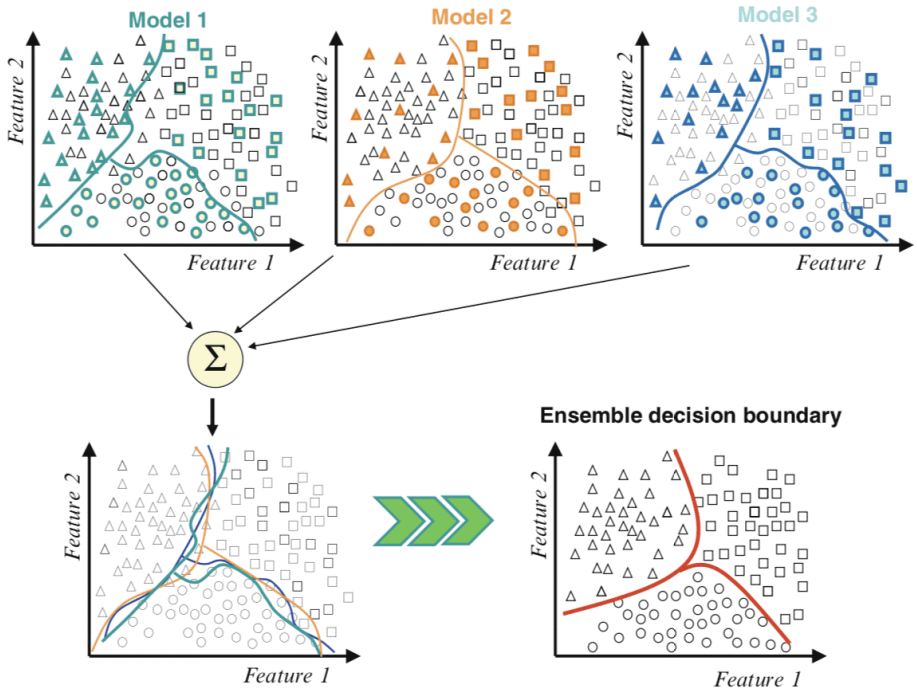

Figura 5. Forma de trabajo de los métodos de ensamble. Fuente: [41]

- *Bagging*: estos métodos, también conocido como "*Bootstrap Aggregating*", tienen como característica usar un mismo algoritmo al que se le pasan diferentes subgrupos aleatorios de datos. Al final se promedian las respuestas [10].

- *Boosting*: estos métodos se caracterizan porque cada uno de los algoritmos elegidos se entrenan secuencialmente. La entrada al primer algoritmo es un subconjunto de datos. El segundo algoritmo a usar tomará los datos que el algoritmo anterior no pudo procesar sumado a otra parte de los datos. Por tanto, lo que se busca es que cada nuevo algoritmo aprenda a corregir los errores del algoritmo anterior [11].

- *Voting:* estos métodos consisten básicamente en tener diversos algoritmos (decididos por el analista datos) a los que se les pasa el mismo grupo de datos. Los resultados de cada uno de los algoritmos se promedian para dar un resultado final [12].

Algoritmos de ensamble o de conjunto	Bagging	Bagging con árboles de decisión
		Bosque aleatorio
		Arboles extremadamente aleatorios
	Boosting	AdaBoost
		GBM
		CatBoost
		Light GBM
	Voting	Blending
		Stacking

Figura 6. Clasificación de algoritmos de ensamble

Una rama contemporánea que surgió en la década de los años 90 es la de las redes neuronales profundas. Si bien las redes neuronales forman parte del *Machine Learning* clásico, en sus inicios eran modelos de una sola capa oculta en la red, conocidos como perceptrones monocapa. Con el tiempo, estos modelos evolucionaron significativamente, permitiendo que las redes neuronales actuales cuenten con múltiples capas ocultas, lo que se conoce como "perceptrón multicapa". Esta evolución ha brindado la capacidad de abordar problemas de gran complejidad.

El término "*Deep Learning*" se utiliza para describir esta corriente que engloba a las redes neuronales artificiales profundas. El adjetivo "profundo" se refiere a que los modelos de redes neuronales en esta vertiente no son simples, sino que

poseen múltiples capas ocultas, ya sea en menor o mayor cantidad. El *Deep Learning* ha demostrado ser altamente aplicable e impactante, y se enfoca principalmente en resolver problemas de regresión y clasificación [13]. Algunos autores pueden no considerar esta corriente como una rama contemporánea separada, ya que evolucionó dentro del campo clásico del *Machine Learning*, como se muestra en la Figura 7.

Las Redes Neuronales Artificiales (RNA) son estructuras abstractas que se inspiran en los mecanismos y la estructura física del aprendizaje y conocimiento biológico. Su diseño se basa en el funcionamiento de las neuronas biológicas [19]. Estas redes están compuestas por una serie de neuronas organizadas en capas interconectadas que comparten información. Para entender su funcionamiento, es útil compararlo con el funcionamiento de una neurona biológica, que consta de dendritas como entradas de información, un núcleo para procesar esta información y axones para transmitir la información a otras neuronas a través de sinapsis [18].

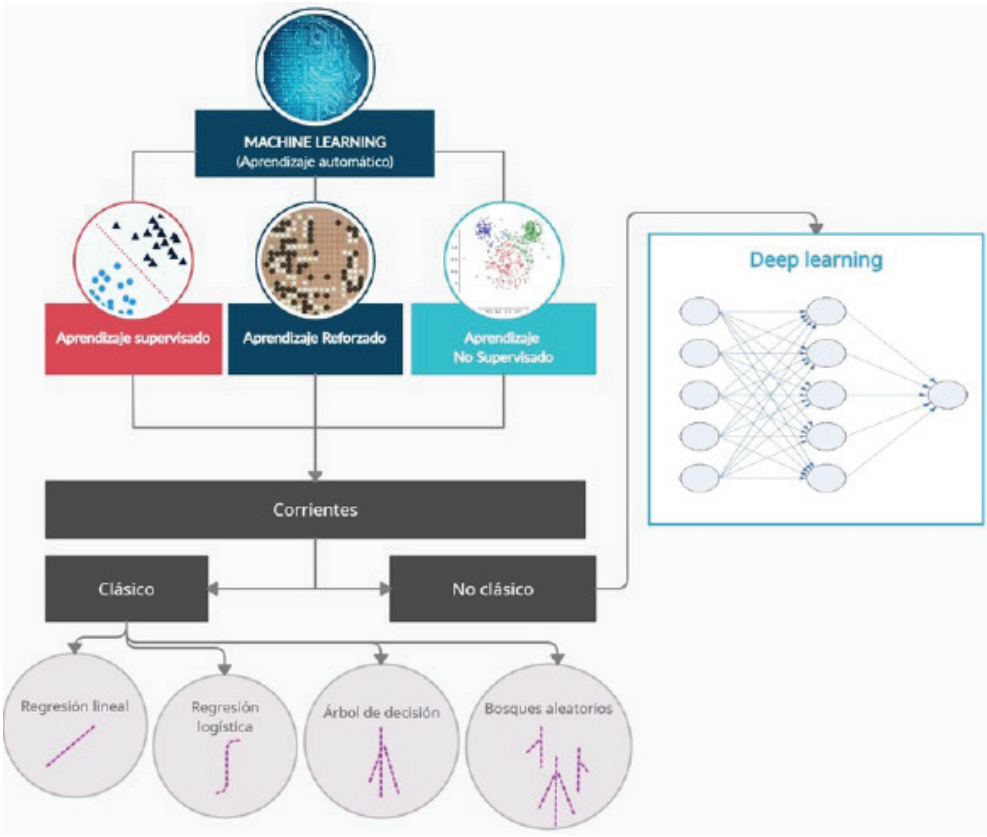

Figura 7. Clasificación de las redes como aprendizaje supervisado clásico. Fuente: [14]–[18]

En el contexto de las redes neuronales artificiales, cada nodo o neurona en la red es una entidad de procesamiento que ajusta los valores de los datos que fluyen a través de ella, funcionando como una función matemática. Estas redes se definen como un conjunto de entidades de procesamiento altamente interconectadas, tomando inspiración del sistema nervioso central biológico, en particular del cerebro. Cada neurona acepta entradas y produce salidas [20].

Es importante destacar que la capacidad de una red neuronal se relaciona con la cantidad de capas que posee. En general, cuanto mayor sea el número de capas, más información puede manejar y abordar funciones complejas [15].Estas redes neuronales artificiales se ubican dentro del campo del *Machine Learning* y se utilizan en una variedad de aplicaciones para el procesamiento de datos y la toma de decisiones.

Las redes neuronales artificiales o algoritmos de aprendizaje profundo o Deep Learning, son modelos inspirados en la estructura y / o función de las redes neuronales biológicas que buscan patrones (regresión y clasificación). Una neurona es una función con un montón de entradas y una salida. Su tarea es tomar todos los números de su entrada, realizar una función sobre ellos y enviar el resultado a la salida.

El aprendizaje profundo suele ser clasificado en dos ramas similar a como se clasifican los algoritmos clásicos de aprendizaje automático atendiendo si la red conoce o no la clase de la variable de salida: aprendizaje supervisado y aprendizaje no supervisado. La Figura 8 muestra un panorama de algunos de los algoritmos que son clasificados dentro del subcampo de aprendizaje profundo que tienen relaciones con las redes neuronales.

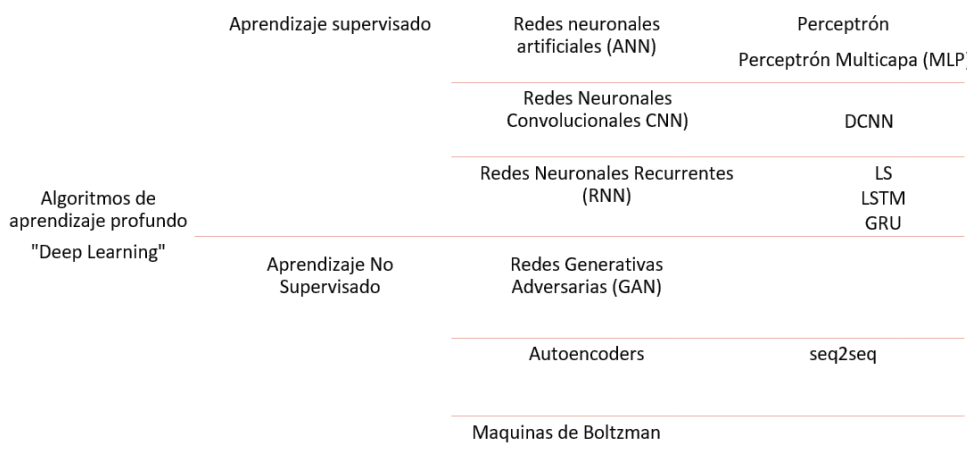

Figura 8. Clasificación de algoritmos de aprendizaje profundo

1.2 PROCESO DE MACHINE LEARNING

Según [21], este proceso guarda una notable similitud con el proceso de Data Mining o minería de datos. Es importante recordar que el *Machine Learning* se desarrolló en cierta medida a partir de esta área dentro de la ciencia de datos, que posteriormente se fortaleció al incorporar contribuciones de diversos campos, tales como la ingeniería, las matemáticas, la estadística, el aprendizaje computacional, la visualización y las bases de datos, entre otros. Estos pasos se resumen de la siguiente manera: comprensión del problema, establecimiento de un criterio de evaluación, evaluación de la solución actual, preparación de los datos, construcción del modelo, análisis de los errores y finalmente, integración del modelo en un sistema (Figura 9).

Las fases de un proceso de *Machine Learning* se pueden resumir de la siguiente manera:

- Fase I: comprender el problema: es esencial comenzar por comprender en profundidad el problema que se desea resolver, así como los datos.
- Este paso puede requerir un tiempo significativo, especialmente si se trata de un problema en un campo poco conocido.
- Fase II: definir un criterio de evaluación: esta etapa se centra en la medición del error. Normalmente, se utilizan métricas como la magnitud cruzada para problemas de clasificación y el error cuadrático medio para problemas de regresión. Sin embargo, también es recomendable considerar otras medidas, como la exhaustividad y la precisión.
- Fase III: evaluación de la solución vigente: en muchos casos, el problema que se aborda con *Machine Learning* ya se está resolviendo de otra manera. Por lo tanto, se evalúa el rendimiento de la solución actual utilizando el esquema de evaluación definido. Luego, se compara con el rendimiento del modelo de *Machine Learning* para determinar si vale la pena implementar el nuevo modelo o seguir con la solución existente.
- Fase IV: preparar los datos: la preparación de los datos es una fase que requiere un esfuerzo considerable. Algunos desafíos comunes en esta etapa incluyen el manejo de datos faltantes, la combinación de datos de diferentes fuentes y la adaptación de los datos a un formato adecuado para su procesamiento. También, es importante normalizar los datos de entrada para facilitar el aprendizaje de los algoritmos.

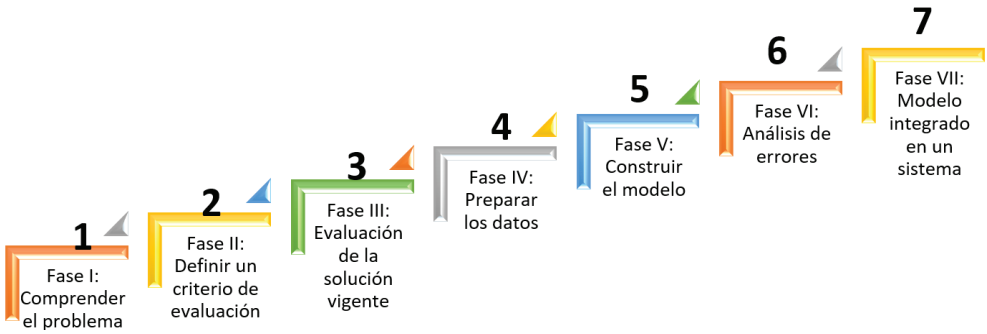

Figura 9. Proceso de Machine Learning

- Fase V: construir el modelo: una vez que los datos están preparados, el esfuerzo se reduce significativamente, ya que existen librerías de *Machine Learning* disponibles públicamente. Durante esta fase, se selecciona la técnica de *Machine Learning* adecuada para abordar el problema.

- Fase VI: análisis de errores: en esta etapa, se analizan los errores cometidos por el modelo. Esto permite identificar posibles mejoras, como el uso de un modelo más robusto, la aplicación de un modelo más sencillo, la necesidad de más datos o características, y una comprensión más profunda del problema.

- Fase VII: modelo integrado en un sistema: finalmente, una vez que se han abordado los errores y se ha comparado el rendimiento con la solución actual, se procede a implementar el modelo de *Machine Learning* en el sistema. Esta fase requiere un esfuerzo significativo, ya que implica la integración del modelo en el sistema existente, la automatización de los pasos de ajuste de datos y la comunicación de errores detectados en tiempo real.

Las fases mencionadas representan un enfoque para abordar problemas mediante el *Machine Learning* y son esenciales para lograr resultados efectivos en proyectos de este tipo.