Tema 1: Conceptos iniciales del Machine Learning

El aprendizaje automático, también conocido como machine learning en inglés, es una rama de la inteligencia artificial que ha cobrado una relevancia sin precedentes en los últimos años. Su capacidad para procesar grandes cantidades de datos y extraer patrones complejos ha revolucionado numerosos campos, desde la atención médica hasta la conducción autónoma de vehículos. En este primer tema, exploraremos en detalle los fundamentos del machine learning, los diferentes enfoques y algoritmos que utiliza, así como evolución histórica, recogiendo los principales avances que ha experimentado en el último siglo.

El concepto básico detrás del aprendizaje automático es enseñar a las computadoras a aprender de los datos disponibles y mejorar automáticamente su rendimiento en tareas específicas sin necesidad de una programación explícita. Este enfoque contrasta con el paradigma tradicional de la programación, donde los desarrolladores escriben reglas y algoritmos específicos para cada tarea que la computadora debe realizar. En cambio, en el aprendizaje automático, los algoritmos son alimentados con grandes conjuntos de datos, y a través de técnicas estadísticas y de optimización, son capaces de identificar patrones, relaciones y regularidades en los datos para realizar predicciones o tomar decisiones.

Existen varios tipos de aprendizaje automático, cada uno con sus propias características y aplicaciones, exploraremos cada uno de ellos en temas posteriores, pero vamos a dar un primer acercamiento.

Uno de los más comunes es el aprendizaje supervisado, donde el algoritmo es entrenado con ejemplos etiquetados, es decir, datos para los cuales ya conocemos la respuesta deseada. Por ejemplo, en un sistema de reconocimiento de imágenes, los datos de entrenamiento consistirían en imágenes junto con etiquetas que indican qué objeto o categoría se representa en cada imagen.

El algoritmo aprende a asociar características específicas de las imágenes con las etiquetas correspondientes, y luego puede generalizar este conocimiento para clasificar nuevas imágenes.

Otro tipo de aprendizaje es el no supervisado, donde el algoritmo se entrena con datos no etiquetados y debe encontrar patrones o estructuras ocultas en los datos por sí mismo. Este enfoque es útil en situaciones donde no tenemos etiquetas para entrenar al modelo o cuando queremos descubrir patrones emergentes en los datos que pueden no ser evidentes a simple vista.

Además, existe el aprendizaje semi-supervisado, que combina elementos de los dos enfoques anteriores, utilizando datos etiquetados y no etiquetados para el entrenamiento. Este enfoque es útil cuando tenemos acceso a grandes cantidades de datos no etiquetados, pero etiquetarlos manualmente sería costoso o poco práctico.

Otro tipo importante es el aprendizaje por refuerzo, donde el algoritmo aprende a través de la interacción con un entorno y recibe retroalimentación en forma de recompensas o castigos según las acciones que realiza. Este enfoque es común en la creación de agentes de inteligencia artificial para juegos o robótica, donde el agente debe aprender a tomar decisiones óptimas para maximizar una recompensa a largo plazo.

El campo del aprendizaje automático ha experimentado un rápido crecimiento en las últimas décadas, impulsado por avances en algoritmos (como habéis podido apreciar los hay muy diversos, para ajustarse a los diferentes problemas que podemos encontrarnos en la sociedad), hardware y disponibilidad de datos. Hoy en día, el machine learning está presente en una amplia gama de aplicaciones, desde sistemas de recomendación en plataformas de streaming hasta diagnósticos médicos asistidos por ordenador y sistemas de reconocimiento de voz en dispositivos móviles.

El aprendizaje automático es un campo fascinante y en constante evolución que está transformando la manera en que interactuamos con la tecnología y aborda problemas complejos en diversas áreas. En este tema vamos a explorar sus fundamentos y su evolución.

Definición y conceptos fundamentales

El Machine Learning, o aprendizaje automático en español, es un área que se centra en desarrollar algoritmos y modelos que permiten a las computadoras aprender patrones y tomar decisiones sin intervención humana directa. Desde la clasificación de correos electrónicos hasta la conducción autónoma de vehículos, el Machine Learning está presente en una amplia gama de aplicaciones que impactan nuestras vidas de manera significativa.

Vamos a comenzar explorando los conceptos abstractos que fundamentan el Machine Learning, como la definición de problemas de aprendizaje, la importancia del conjunto de datos y la noción de modelos. Luego, nos adentraremos en aspectos más técnicos, como los diferentes tipos de algoritmos de aprendizaje, la evaluación de modelos y la optimización de parámetros.

A lo largo de esta sección, descubriremos cómo estos conceptos se entrelazan para permitir que las máquinas aprendan de los datos y mejoren su desempeño con el tiempo, donde exploraremos desde los fundamentos más abstractos hasta los detalles más técnicos de esta apasionante disciplina.

Razonamiento y aprendizaje

La facultad de razonamiento es una capacidad que se atribuye exclusivamente a la especie humana, distinguiéndola del resto de seres vivos. Implica la habilidad de recordar información pertinente a los hechos que se desean analizar, así como de relacionar esos datos para inferir nuevo conocimiento en situaciones desconocidas.

Este proceso permite al ser humano clasificar objetos y situaciones, facilitando su reconocimiento y la respuesta adecuada ante ellos. Desde los inicios de la inteligencia artificial, se ha buscado replicar esta capacidad en los ordenadores. Los primeros intentos dieron lugar a los llamados Sistemas Basados en Conocimiento (SBC), diseñados para resolver problemas específicos dentro de un dominio de aplicación determinado, como el diagnóstico de enfermedades.

Dentro de los SBC, destacan los sistemas expertos, que se basan en reglas almacenadas en una base de conocimiento. Estos sistemas emulan el razonamiento humano en un área específica de conocimiento, utilizando un motor de inferencia para derivar conclusiones a partir de la información disponible.

Existen dos estrategias principales para la inferencia en sistemas expertos: encadenamiento hacia delante y encadenamiento hacia atrás. Mientras que el primero busca aplicar reglas basadas en antecedentes cumplidos, el segundo parte de los consecuentes deseados para determinar los antecedentes necesarios.

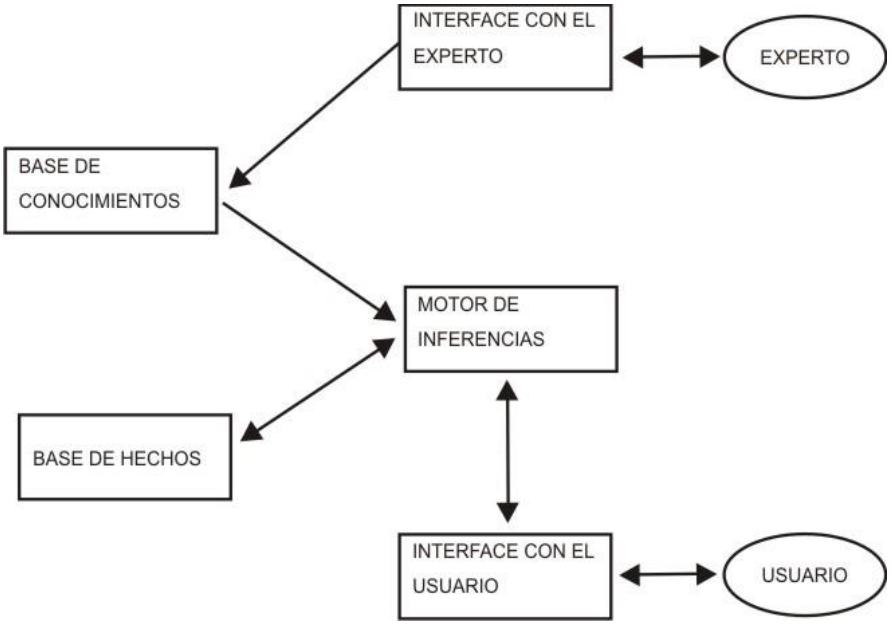
Además de los sistemas expertos, existen los sistemas difusos, que son eficaces para trabajar con estados que no son simplemente binarios, como en el caso de la temperatura corporal en el diagnóstico médico. Estos sistemas utilizan conjuntos difusos y lógica difusa para representar la incertidumbre y la imprecisión en los datos.

Los sistemas expertos son un tipo de sistema de inteligencia artificial que se basa en el conocimiento experto humano para resolver problemas en áreas específicas. Estos sistemas están diseñados para emular el razonamiento y la toma de decisiones de un experto humano en un campo particular, como la medicina, la ingeniería, la gestión empresarial, entre otros.

La base de conocimiento es una parte fundamental de un sistema experto. Consiste en una base de datos que contiene el conocimiento experto en forma de reglas, hechos y relaciones. Esta base de conocimiento se construye a partir de la experiencia y el conocimiento de expertos humanos en el campo específico.

Por otro lado, el motor de inferencia el componente que utiliza la base de conocimiento para realizar el razonamiento y llegar a conclusiones o soluciones. El motor de inferencia aplica algoritmos y técnicas de inferencia para deducir respuestas a partir de la información disponible en la base de conocimiento.

A medida que se adquiere nuevo conocimiento experto o se refinan las reglas existentes, la base de conocimiento puede actualizarse para mejorar el rendimiento del sistema experto. Esto permite que el sistema se mantenga relevante y preciso con el tiempo.



Tanto los sistemas expertos como los sistemas difusos son herramientas importantes en el ámbito de la inteligencia artificial, permitiendo modelar y resolver problemas complejos mediante el uso de reglas y conjuntos difusos. Estos sistemas han tratado de imitar la manera en la que los humanos razonamos, pero hay otro concepto propio de los humanos todavía más complejo de emular, y este es el aprendizaje.

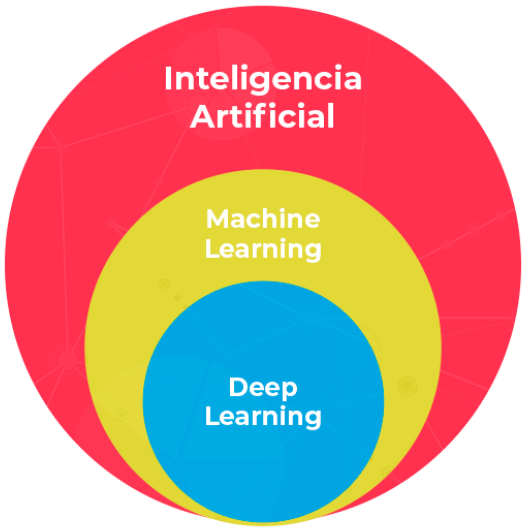
Desde el momento de nuestro nacimiento, nuestros sentidos nos sumergen en una realidad en constante cambio, a la cual debemos ajustarnos para garantizar nuestra supervivencia. El aprendizaje surge como una capacidad que nos permite, a partir de la información percibida del entorno, desarrollar o modificar hábitos, comportamientos y habilidades con el fin de adaptarnos de manera más efectiva a dicho entorno.

En el ámbito de la inteligencia artificial (IA), la capacidad de aprendizaje se ve significativamente limitada en comparación con la de los seres humanos. Tareas que para nosotros parecen simples, como reconocer el rostro de un conocido, representan un desafío enormemente complejo para una computadora. Incluso ante cambios como el uso de gafas de sol, el crecimiento de barba o la cobertura de la boca con una bufanda, nuestro cerebro cuenta con una poderosa capacidad para el reconocimiento de patrones.

Este fenómeno ha impulsado durante años el desarrollo de redes neuronales, con el objetivo de replicar la capacidad humana de reconocimiento. Las redes neuronales buscan emular el funcionamiento del cerebro humano, permitiendo así que las máquinas aprendan y adapten sus respuestas a través de la experiencia, aunque aún distan de alcanzar la versatilidad y la eficiencia del sistema cognitivo humano.

Definición de ML y componentes principales

El aprendizaje automático (ML) es una rama especializada dentro del campo de la inteligencia artificial que se centra en el desarrollo de técnicas y algoritmos que permiten a las computadoras aprender de los datos disponibles y mejorar su rendimiento en tareas específicas sin necesidad de ser programadas explícitamente para cada situación. Esta disciplina tiene como objetivo principal capacitar a las máquinas para que puedan identificar patrones complejos y extraer información útil de conjuntos de datos masivos y variados.



IA: Combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano.

Machine Learning: Rama de la Inteligencia artificial (IA) que estudia como dotar a las máquinas de capacidad de aprendizaje

Deep Learning: algoritmo automático jerárquico que emula el aprendizaje humano con el fin de obtener ciertos conocimientos.

El proceso de "aprender" en el contexto del aprendizaje automático implica que un algoritmo analiza los datos de entrada, identifica regularidades o patrones estadísticos en ellos y construye un modelo o representación interna de estas relaciones. Este modelo puede ser utilizado posteriormente para realizar predicciones o tomar decisiones sobre datos nuevos o no vistos anteriormente. Es importante destacar que este aprendizaje no se produce de manera estática, sino que los modelos de ML pueden mejorar y adaptarse con el tiempo a medida que se alimentan con más datos y reciben retroalimentación sobre su desempeño.

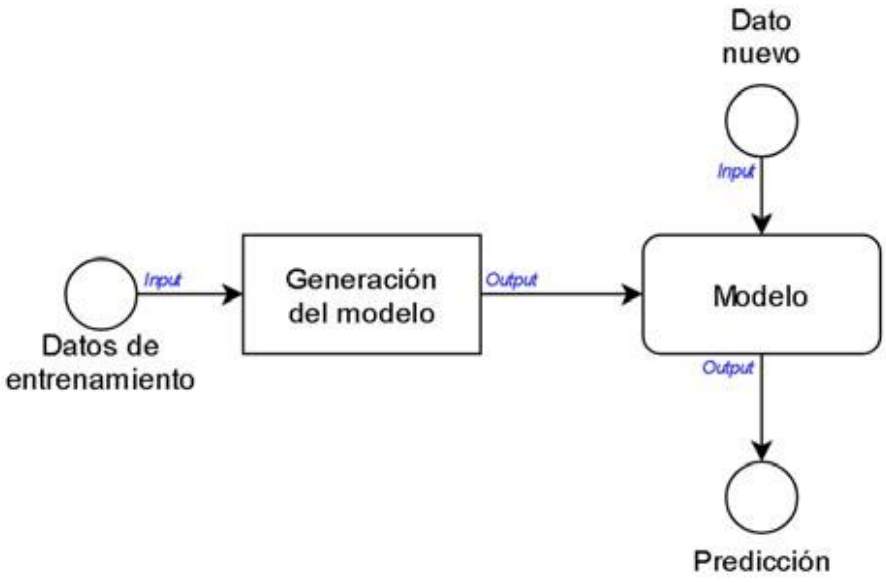
La característica "automática" del aprendizaje automático radica en la capacidad de estos sistemas para mejorar su rendimiento de forma autónoma con el tiempo, sin requerir una intervención humana directa para ajustar los algoritmos o modificar los modelos. Esta capacidad de automejora es fundamental para que los sistemas de ML puedan adaptarse a entornos cambiantes y mantener su relevancia y eficacia a lo largo del tiempo.

El siguiente diagrame diagrama ilustra el proceso de entrenamiento y predicción en la generación de un modelo de ML. En la parte izquierda del diagrama, se ingresan los datos iniciales que se utilizarán para entrenar el modelo. Estos datos pueden ser ejemplos etiquetados, características de entrada y salidas esperadas.

En la etapa de generación del modelo, los datos de entrenamiento se procesan para crear un modelo. El modelo puede ser una red neuronal, un árbol de decisión u otro algoritmo de aprendizaje automático.

El modelo representa el resultado del procesamiento y aprendizaje de los datos iniciales. El modelo captura patrones y relaciones en los datos para hacer predicciones. De esta manera, cuando se tiene un nuevo conjunto de datos (por ejemplo, una imagen sin etiquetar), se ingresa al modelo entrenado.

El modelo aplica el conocimiento aprendido para hacer una predicción basada en los datos nuevos. Por ejemplo, si el modelo fue entrenado para clasificar imágenes de gatos y perros, hará una predicción sobre si la nueva imagen es un gato o un perro.

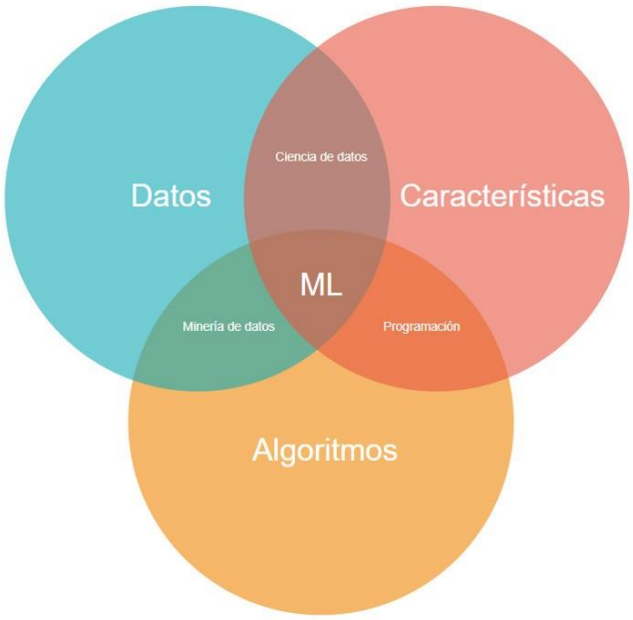


Como hemos podido apreciar en la construcción del modelo han intervenido algunos elementos fundamentales. De esta manera, tenemos que los tres elementos fundamentales del ML son los siguientes:

1. **Datos**: Representan el conjunto histórico de información que se utiliza como entrada para el algoritmo de aprendizaje. La cantidad y calidad de los datos son cruciales para el éxito del proceso de aprendizaje. Cuantos más datos tengamos disponibles y de mejor calidad sean, mejores serán las predicciones o decisiones que pueda realizar el algoritmo.

2. **Características**: Estos son los atributos o variables que el algoritmo de aprendizaje analiza para encontrar patrones y tomar decisiones. Las características pueden ser diversas y representan aspectos relevantes de los datos que se están considerando. Por ejemplo, en un conjunto de datos sobre automóviles, las características podrían incluir el color, el modelo, la potencia del motor, etc.

3. **Algoritmos**: Son los procedimientos matemáticos y estadísticos que se aplican a los datos y características para entrenar el modelo de aprendizaje automático. Estos algoritmos aprenden de los datos proporcionados, identificando patrones y relaciones entre las características para realizar predicciones o tomar decisiones. Existen diversos tipos de algoritmos de ML, cada uno con sus propias características y aplicaciones específicas, como los algoritmos de regresión, clasificación, agrupamiento, entre otros. La elección del algoritmo adecuado depende del tipo de problema y de los datos disponibles.



Como podréis observar el dato es fundamental dentro del proceso de construcción de un modelo de ML. El modelo aprende a partir de los datos, por ello hay diferentes factores alrededor de ellos que pueden ser determinantes para el ML.

En primer lugar, la calidad de los datos tiene un impacto directo en la calidad del modelo. Si los datos están incompletos, son incorrectos o están sesgados, es probable que el modelo produzca resultados erróneos o poco fiables. Por lo tanto, es crucial asegurarse de que los datos utilizados para entrenar el modelo sean precisos, completos y representativos del problema que se está tratando de resolver.

Por otro lado, un modelo de machine learning busca aprender patrones y relaciones a partir de los datos de entrenamiento para hacer predicciones o tomar decisiones sobre nuevos datos no vistos. Cuanto más variados y representativos sean los datos de entrenamiento, mejor será la capacidad del modelo para generalizar y realizar predicciones precisas sobre nuevos datos.

Los modelos de machine learning son capaces de identificar patrones y relaciones complejas en los datos que pueden no ser evidentes para los humanos. Sin embargo, para que los modelos sean efectivos en esta tarea, es necesario proporcionarles datos de alta calidad que contengan información relevante y útil para el problema que se está abordando.

Sin nos referimos a los datos hay un concepto que no podemos dejar de lado y que ha sido fundamental para que el ML se encuentre en el punto en el que se encuentra a día de hoy, y este es el Big Data. El Big Data y el Machine Learning están intrínsecamente relacionados, ya que el primero proporciona los datos necesarios para entrenar y desarrollar modelos de ML, mientras que el segundo proporciona las técnicas y algoritmos para extraer información valiosa y conocimientos de estos grandes conjuntos de datos. Juntos, el Big Data y el Machine Learning están transformando la forma en que las organizaciones operan, toman decisiones y brindan servicios personalizados a sus usuarios.

Big Data

El término "Big Data" hace referencia a la gestión y análisis de grandes volúmenes de datos, tanto estructurados como no estructurados, que son demasiado complejos o masivos para ser procesados mediante métodos tradicionales de almacenamiento y análisis de datos. Este concepto ha ganado relevancia debido al rápido crecimiento en la cantidad de datos generados en el mundo, especialmente con la expansión de internet, las redes sociales, los dispositivos móviles y los sensores inteligentes.

La generación masiva de datos se debe a una combinación de varios factores, el aumento en la conectividad global y la proliferación de dispositivos inteligentes, como teléfonos móviles, tabletas, computadoras portátiles, dispositivos IoT (Internet de las cosas) y sensores inteligentes, ha llevado a una explosión en la generación de datos. Cada interacción en línea, cada transacción, cada dispositivo conectado genera datos que se acumulan rápidamente.

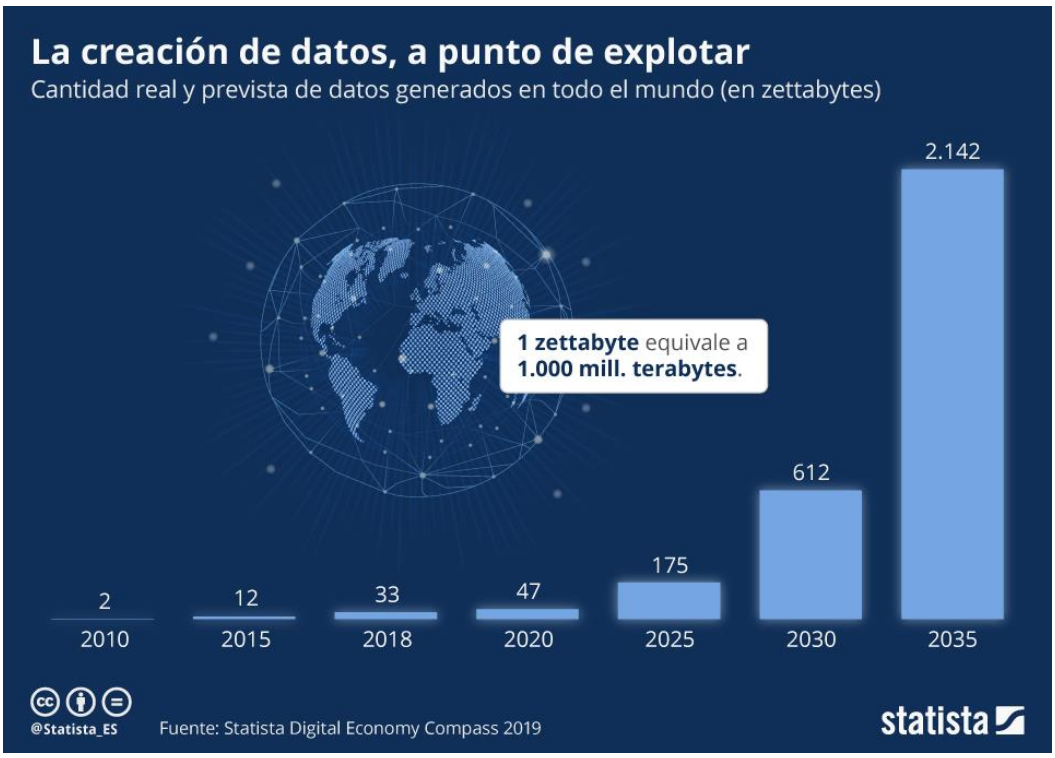
Por otro lado, las redes sociales, los blogs, los foros y otras plataformas en línea han permitido que los usuarios generen y compartan una cantidad enorme de contenido digital, que incluye publicaciones, fotos, videos, comentarios y más. Esto contribuye significativamente al volumen de datos generados en Internet.

Además, los avances en la tecnología de almacenamiento y procesamiento de datos han hecho que sea más fácil y económico almacenar grandes cantidades de datos. Esto ha llevado a que las organizaciones y los individuos guarden datos de manera más prolífica y durante períodos de tiempo más largos.

Es difícil predecir con precisión la cantidad exacta de datos en Internet en los próximos años, ya que está influenciada por una serie de factores, como el crecimiento de la conectividad, la adopción de dispositivos inteligentes, el aumento del uso de redes sociales, el avance de la tecnología de sensores, entre otros. Sin embargo, se espera que la cantidad de datos en Internet continúe aumentando exponencialmente en los próximos años.

Según el "Cisco Annual Internet Report" de 2021, se estima que el tráfico de datos IP global alcanzará los 748 exabytes por mes para 2025. Además, se espera que el número de dispositivos conectados a Internet aumente significativamente, con alrededor de 29.3 mil millones de dispositivos conectados en 2025, en comparación con los 14.7 mil millones en 2020.

Otro informe de IDC y Seagate predice que la cantidad total de datos generados en el mundo alcanzará los 175 zettabytes para 2025. Esto representa un crecimiento masivo en comparación con los 33 zettabytes de datos generados en 2018.



Estas cifras muestran una tendencia clara hacia un crecimiento exponencial en la cantidad de datos en Internet en los próximos años. Este aumento en la cantidad de datos presenta desafíos y oportunidades significativas en términos de almacenamiento, procesamiento, análisis y gestión de datos, así como en la aplicación de tecnologías como el Big Data y el Machine Learning para extraer información valiosa y conocimientos de estos grandes conjuntos de datos.

Para las empresas es realmente importante analizar estos datos por varias razones fundamentales, entre otras cosas, les proporciona información valiosa y conocimientos que pueden respaldar la toma de decisiones informadas en una variedad de áreas, desde estrategias de negocios hasta políticas públicas. Al comprender y analizar los datos, las organizaciones pueden identificar tendencias, patrones y relaciones que les permiten tomar decisiones más efectivas y fundamentadas.

El análisis de datos puede ayudar a las organizaciones a identificar oportunidades emergentes en el mercado, así como posibles amenazas o riesgos. Al examinar datos relevantes, las organizaciones pueden anticipar cambios en la demanda del mercado, identificar necesidades no satisfechas de los clientes, detectar posibles problemas operativos y tomar medidas proactivas para abordarlos.

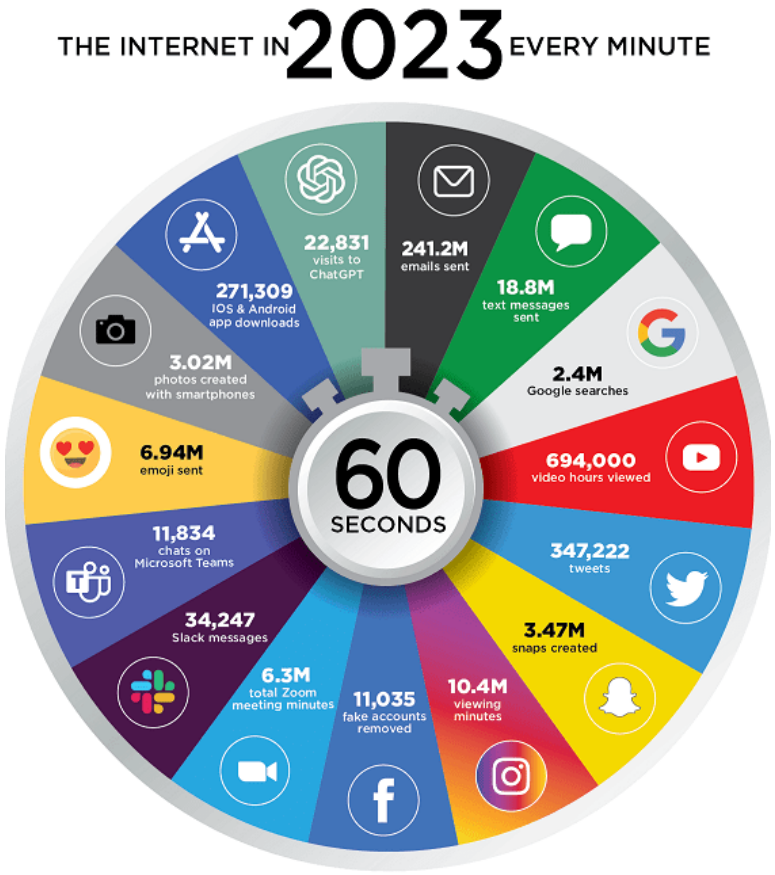
Por otro lado, el análisis de datos permite comprender mejor las preferencias, comportamientos y necesidades individuales de los clientes. Esto permite a las organizaciones ofrecer experiencias personalizadas y adaptadas a las necesidades específicas de cada cliente, lo que puede mejorar la satisfacción del cliente, aumentar la lealtad y generar mayores ingresos.

Aun así, no todo es de color de rosas en lo que respecta al Big Data este presenta una serie de características clave del Big Data que se describen mediante las "7 V" del Big Data y que suponen un gran desafío para las organizaciones. Aun así, de superarlas pueden obtener un retorno muy interesante. Las “7 V” son las siguientes:

1. **Volumen**: Se refiere a la cantidad masiva de datos que se generan y recopilan continuamente. Estos datos pueden provenir de diversas fuentes, como transacciones comerciales, redes sociales, sensores, registros de servidores, entre otros. Como hemos comentado previamente, el volumen de datos esta creciendo exponencialmente, pero no es la única característica que define el Big Data, hay más características interesantes y desafiantes, como las que veremos a continuación.

2. **Velocidad**: Hace referencia a la rapidez con la que los datos son generados, procesados y analizados. Con la tecnología actual, es posible capturar y procesar datos en tiempo real, lo que permite tomar decisiones más ágiles y basadas en información actualizada. Esto es necesario ya que se generan muchos datos cada minuto que pasa (como podéis apreciar en la siguiente imagen) y algunos de esos datos tienen “fecha de caducidad”.

Ya no es que haya muchos datos, si no que algunos de ellos deben de ser procesados de manera prácticamente instantánea, ya que pasado cierto tiempo ya no tienen validez, lo que es tendencia ahora puede dejar de serlo en muy poco tiempo. Es fundamental tener esto en cuenta y procesar los datos con la inmediatez que requieres. Por ello se han desarrollado muchas tecnologías enfocadas al procesamiento de datos en Streaming, una necesidad cada vez más real.



3. **Variedad**: La Variedad en el contexto del Big Data se refiere a la diversidad de tipos y formatos de datos que se generan y recopilan en la actualidad. A diferencia de los datos estructurados que se organizan en tablas con filas y columnas, como en una base de datos relacional tradicional, los datos del Big Data pueden ser estructurados, semiestructurados o no estructurados.

    Datos Estructurados: Son datos organizados en un formato predefinido y bien definido. Estos datos se almacenan en tablas con campos específicos y tipos de datos asignados a cada columna. Los datos estructurados son comunes en las bases de datos relacionales y son fáciles de consultar y analizar utilizando consultas SQL.

Datos Semiestructurados: Son datos que no se ajustan a un formato predefinido, pero tienen alguna estructura, como documentos XML, JSON, CSV, entre otros. Aunque los datos semiestructurados pueden contener etiquetas o metadatos que los organizan parcialmente, su estructura no es tan rigurosa como la de los datos estructurados.

Datos No Estructurados: Son datos que no tienen una estructura predefinida y no se pueden almacenar fácilmente en bases de datos relacionales. Esto incluye datos como texto libre, imágenes, audio, video, correos electrónicos, redes sociales, comentarios de clientes, entre otros. Los datos no estructurados representan la mayor parte del Big Data y pueden ser difíciles de analizar con enfoques tradicionales.

La variedad de datos en el Big Data presenta desafíos significativos en términos de almacenamiento, procesamiento y análisis. Los sistemas y herramientas de Big Data deben ser capaces de manejar y procesar diferentes tipos de datos de manera eficiente y efectiva. Además, el análisis de datos del Big Data debe ser lo suficientemente flexible para adaptarse a la diversidad de los datos y extraer información útil y conocimientos significativos de ellos.

El ML juega aquí un papel crucial en el manejo y análisis de la variedad de datos en el Big Data. En el caso de las imágenes, el ML es esencial para realizar tareas como la clasificación de imágenes, detección de objetos, segmentación de imágenes y reconocimiento facial, entre otros. Esto permite convertir los millones de imágenes y videos que se generan en redes sociales en una variable categórica. Convirtiendo un dato no estructurado y difícil de trabajar en algo mucho más manipulable

En el caso de los textos, como documentos, correos electrónicos, redes sociales y comentarios de clientes, el ML es esencial para realizar tareas como la clasificación de textos, extracción de información, resumen de texto y análisis de sentimientos, entre otros. Esto es especialmente útil ya que de igual manera nos permite convertir un dato difícil de procesar en algo mucho más tangible.

4. **Veracidad**: La veracidad se refiere a la confiabilidad y precisión de los datos. Es fundamental asegurar la calidad de los datos para evitar errores y sesgos en el análisis, lo que podría conducir a decisiones incorrectas.

   La veracidad es un aspecto crítico del Big Data, ya que la precisión y la fiabilidad de los datos pueden afectar significativamente la validez de los insights y decisiones que se derivan del análisis de esos datos. Los datos deben ser completos, consistentes y libres de errores o inconsistencias para garantizar que los análisis y las conclusiones basadas en ellos sean confiables y precisas. Los errores en los datos pueden conducir a interpretaciones incorrectas y decisiones erróneas.

   Es importante también conocer el origen y la fuente de los datos para evaluar su veracidad. Los datos de fuentes confiables y verificadas tienen una mayor probabilidad de ser precisos y confiables en comparación con los datos de fuentes desconocidas o no verificadas. La transparencia en cuanto a la procedencia de los datos es fundamental para evaluar su confiabilidad.

5. **Valor**: El valor del Big Data radica en la capacidad de obtener información significativa y útil a partir de los datos. El análisis de grandes volúmenes de datos puede proporcionar insights valiosos para mejorar la toma de decisiones, optimizar procesos y descubrir nuevas oportunidades de negocio.

6. **Variabilidad**: La variabilidad se refiere a la diversidad en la estructura y el significado de los datos. Esto incluye cambios en la calidad de los datos a lo largo del tiempo, así como la necesidad de adaptar los análisis a diferentes contextos y escenarios.

   La calidad de los datos puede variar significativamente entre diferentes conjuntos de datos y fuentes de datos. Algunos datos pueden ser precisos, completos y confiables, mientras que otros pueden contener errores, duplicados o valores atípicos. La variabilidad en la calidad de los datos puede afectar la validez y la confiabilidad de los análisis realizados con esos datos.

7. **Visualización**: La visualización de datos es crucial para comprender y comunicar los insights obtenidos del análisis de Big Data. Las técnicas de visualización permiten representar de manera clara y concisa la información, facilitando su interpretación y toma de decisiones.

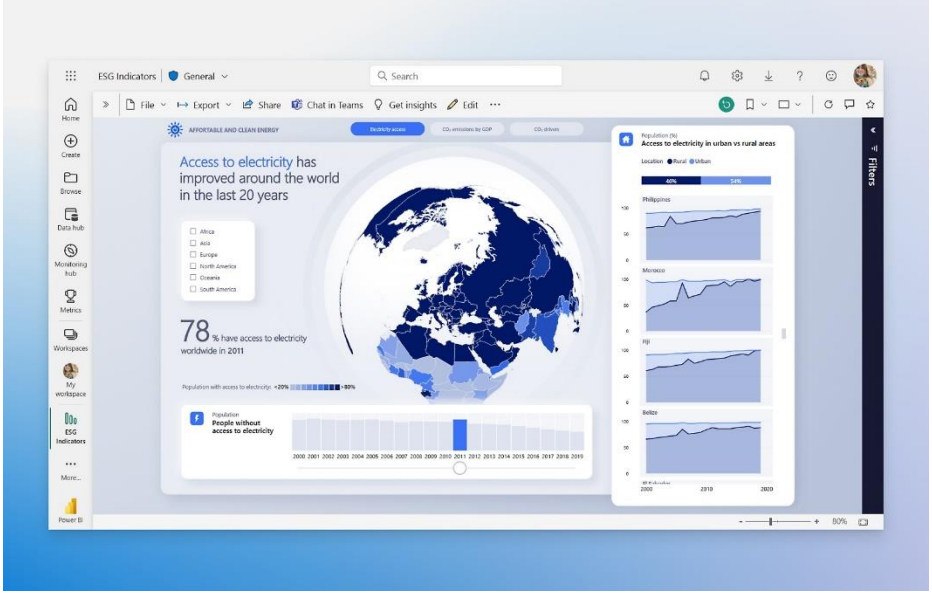
   La visualización permite a los analistas explorar datos complejos de manera intuitiva y comprender mejor su estructura y distribución. Gráficos como histogramas, diagramas de dispersión y parcelas de densidad pueden ayudar a identificar patrones y anomalías en los datos.

Además, facilita la comunicación de insights y hallazgos a una audiencia no técnica. Gráficos claros y efectivos pueden ayudar a transmitir información compleja de manera clara y concisa, lo que permite a los tomadores de decisiones comprender rápidamente las implicaciones de los datos.

Y, por último, pueden resaltar tendencias y patrones en los datos que pueden pasar desapercibidos en conjuntos de datos masivos. Gráficos de líneas, gráficos de barras y diagramas de caja pueden ayudar a identificar tendencias temporales, fluctuaciones estacionales y cambios en el comportamiento del usuario.

Asi pues, alrededor de la Visualizacion surgen otros conceptos relacionados como el BI, el cual se define como un conjunto de tecnologías, aplicaciones y prácticas que permiten a las organizaciones recopilar, almacenar, analizar y presentar datos para facilitar la toma de decisiones empresariales informadas. El BI se centra en transformar datos brutos en información significativa y conocimientos útiles para ayudar a las organizaciones a comprender mejor su desempeño, identificar tendencias, oportunidades y desafíos, y tomar decisiones estratégicas.

A continuación, puedes observar una imagen que muestra una de las herramientas de visualización y BI mas interesantes, Power BI, de Microsoft.



En resumen, el Big Data representa una oportunidad significativa para las organizaciones que buscan aprovechar el potencial de los datos para obtener insights valiosos y tomar decisiones más informadas. Sin embargo, también plantea desafíos en términos de almacenamiento, procesamiento, análisis y gestión de la enorme cantidad de datos disponibles.

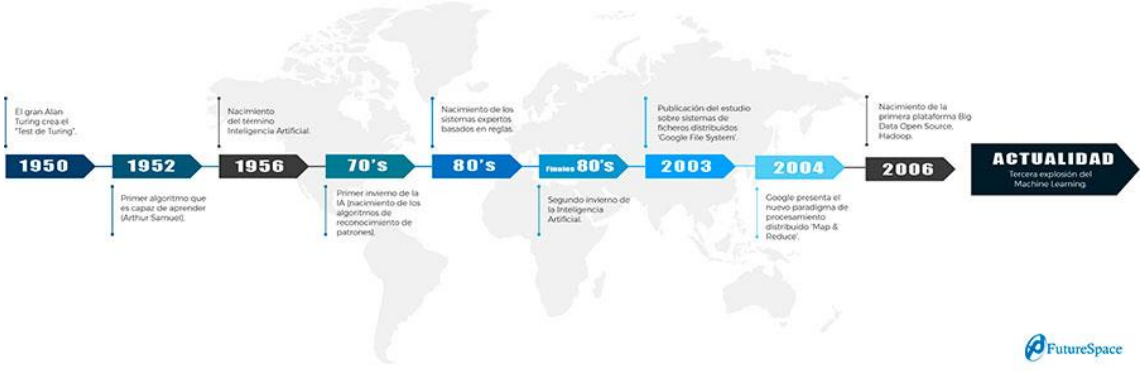
Historia y evolución

Por raro que parezca y aunque os cueste creerlo, la historia no comenzó en 2020, se remonta al siglo XVII. El aprendizaje automático es una poderosa herramienta para implementar tecnologías de inteligencia artificial.

Debido a su capacidad de aprender y tomar decisiones, el aprendizaje automático suele denominarse IA, aunque técnicamente es una subdivisión de la tecnología de IA. Pero esto no ha sido siempre así, hasta finales de la década de 1970, el aprendizaje automático era solo un componente más del progreso de la IA. Luego se desvió y evolucionó por sí mismo, convirtiéndose en un facilitador vital en muchas áreas tecnológicas de vanguardia de la actualidad. El concepto ha evolucionado mucho como veremos a continuación, hasta el punto de que los científicos trabajan actualmente en enfoques de aprendizaje automático cuántico.

El Machine Learning (ML) ha experimentado una notable evolución desde sus primeros pasos en los años 50 hasta su prominencia actual en la industria de la tecnología de la información. Este viaje histórico nos lleva a través de hitos significativos, avances tecnológicos y la contribución de visionarios que sentaron las bases para su desarrollo y aplicación en una amplia gama de campos.

Aunque parezca que el ML es propio de la informática y el sector IT, contando con una gran influencia técnica, esto no empezó siendo así. El modelo de interacción entre células cerebrales que sustenta el aprendizaje automático moderno se deriva de la neurociencia. En 1949, el psicólogo Donald Hebb publicó The Organization of Behavior (La organización del comportamiento), en el que propuso la idea del aprendizaje “endógeno” o “autogenerado”. Sin embargo, tuvieron que pasar siglos e inventos locos como el telar para almacenar datos para que tuviéramos una comprensión tan profunda del aprendizaje automático como la que tenía Hebb en 1949.



Inicios pioneros

Los inicios del ML se sitúan en la década de 1950, marcándolo como un período crucial de descubrimientos y avances que sentaron las bases para la revolución tecnológica que estamos presenciando en la actualidad. Uno de los hitos más destacados de este tiempo fue el desarrollo del primer programa de aprendizaje informático por parte de Arthur Samuel. El programa Checkers de Arthur Samuel, creado para jugar en el IBM 701, se mostró al público por primera vez en televisión el 24 de febrero de 1956. Robert Nealey, un autodenominado maestro de las damas jugó el juego en un ordenador IBM 7094 en 1962. El ordenador ganó. El programa Samuel Checkers perdió otras partidas contra Nealey. Sin embargo, fue considerado un hito para la inteligencia artificial y proporcionó al público un ejemplo de las capacidades de un ordenador electrónico a principios de la década de 1960.

Lo sorprendente de este programa radicaba en su capacidad para mejorar su rendimiento a medida que adquiría experiencia. Utilizando técnicas de aprendizaje automático, el programa podía identificar patrones y estrategias ganadoras a partir de la práctica repetida, demostrando así la capacidad de las computadoras para aprender y mejorar con el tiempo.

Para el público analfabeto en tecnología de 1962, este fue un evento significativo. Sentó las bases para que las máquinas hicieran otras tareas inteligentes mejor que los humanos. Y la gente empezó a pensar: ¿superarán los ordenadores a los humanos en inteligencia? Al fin y al cabo, los ordenadores solo existían desde hacía unos años y el campo de la inteligencia artificial estaba aún en pañales…

Paralelamente, en el ámbito teórico, el artículo científico “A Logical Calculus of the Ideas Immanent in Nervous Activity”, publicado en 1948 por Walter Pitts y Warren McCulloch, sentó las bases para el desarrollo de las redes neuronales artificiales.

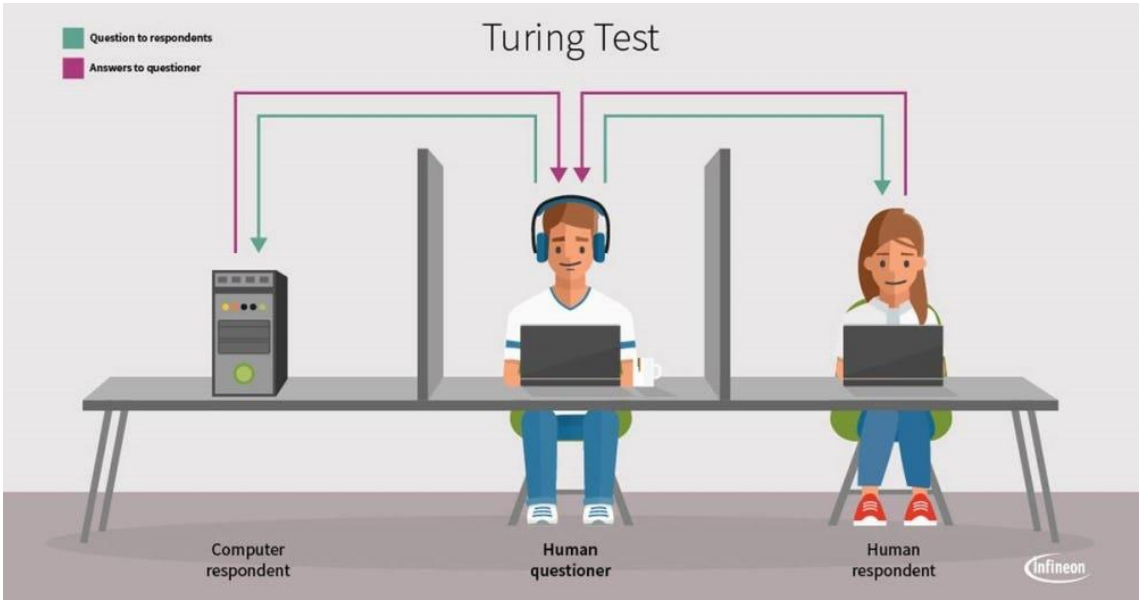
Este trabajo proporcionó el primer modelo matemático de las redes neuronales, inspirándose en los principios de funcionamiento del cerebro humano. Aunque en su momento recibió poca atención, este artículo es considerado ahora como el punto de partida de la disciplina moderna del aprendizaje automático, allanando el camino para avances posteriores como el aprendizaje profundo y el aprendizaje automático cuántico.

El modelo propuesto por McCulloch y Pitts, que más tarde se conocería como teoría Hebbiana, postulado de Hebb o regla de Hebb, sugiere que cuando una célula nerviosa activa repetidamente a otra, se produce algún tipo de cambio en la eficacia de su conexión. Este proceso, descrito como "aprendizaje hebbiano", implica que la fuerza de la conexión entre dos neuronas aumenta si una activa a la otra con frecuencia.

Hebb introdujo el concepto de "conjuntos celulares", grupos de neuronas que funcionan como una unidad de procesamiento, cuya combinación de conexiones se ajusta en respuesta a los estímulos. Este modelo ha influido en la comprensión psicológica del procesamiento de la información en la mente y ha abierto el camino para el desarrollo de sistemas computacionales inspirados en los procesos cerebrales naturales, como el aprendizaje automático.

Aunque la comunicación sináptica en el cerebro se basa principalmente en señales químicas, las modernas redes neuronales artificiales aún se basan en los principios eléctricos sobre los cuales se fundamenta la teoría Hebbiana.

Otro hito fundamental en la historia del Machine Learning es el famoso “Test de Turing”, propuesto por Alan Turing en 1950. Este test, diseñado para evaluar la capacidad de una máquina para exhibir un comportamiento inteligente equivalente al de un humano, marcó un antes y un después en el campo de la inteligencia artificial. Si una máquina puede engañar a un interrogador humano haciéndole creer que también es humano, según Turing, puede considerarse que posee inteligencia artificial.



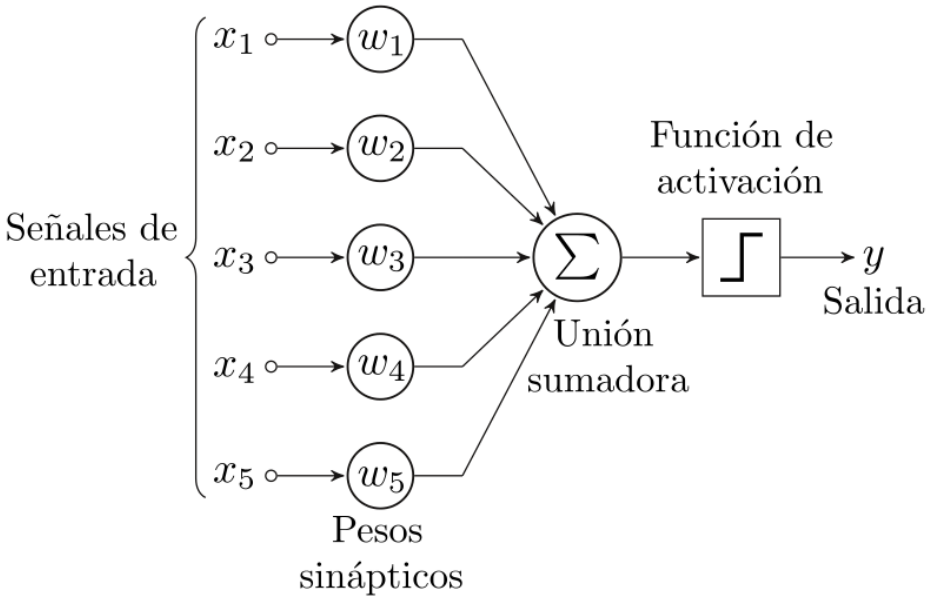
El procedimiento del Test de Turing original implica tres terminales físicamente separados. Uno de estos terminales está controlado por una computadora, mientras que los otros dos son utilizados por seres humanos. Durante el experimento, uno de los humanos asume el papel de interrogador, mientras que el segundo humano y la computadora son los encuestados. El interrogador formula preguntas dentro de un área de estudio específica, en un formato y contexto determinados. Tras un número predeterminado de preguntas, el interrogador debe decidir cuál de los encuestados es real y cuál es artificial.

El ordenador se considera que posee "inteligencia artificial" si el interrogador toma la decisión correcta en la mitad de las ejecuciones del test o menos. Este examen, conceptualizado por Turing en la década de 1950, marcó el inicio del campo del aprendizaje de máquinas. Su trabajo sobre este tema se plasmó en el artículo "Computing Machinery and Intelligence", publicado en 1950.

Estos hitos históricos marcaron el inicio del campo del Machine Learning, que desde entonces ha experimentado un crecimiento exponencial y ha revolucionado numerosos campos, desde la tecnología hasta la medicina y la industria. La capacidad de las máquinas para aprender y mejorar de forma autónoma sigue siendo uno de los mayores logros de la ciencia y la tecnología modernas.

Auge de las Redes Neuronales

En la misma época, Frank Rosenblatt presentó el concepto revolucionario del Perceptrón, una rudimentaria forma de red neuronal inspirada en el funcionamiento del cerebro humano. A través del Perceptrón, se exploraron las capacidades de las máquinas para tomar decisiones simples y resolver problemas complejos mediante conexiones de nodos.



En julio de 1958, la Oficina de Investigación Naval de los Estados Unidos reveló un invento notable: el perceptrón. Utilizando un IBM 704, un ordenador del tamaño de una habitación que pesaba 5 toneladas se alimentó con una serie de tarjetas perforadas. Después de 50 intentos, logró aprender a identificar las tarjetas con marcas a la izquierda de las marcas a la derecha. Según su creador, Frank Rosenblatt, este hito demostró las capacidades del "perceptrón", siendo "la primera máquina capaz de generar un pensamiento original". Rosenblatt observó en 1958: "Las historias sobre la creación de máquinas con cualidades humanas han sido durante mucho tiempo una provincia fascinante en el ámbito de la ciencia ficción. Sin embargo, estamos a punto de asistir al nacimiento de una máquina de este tipo: una máquina capaz de percibir, reconocer e identificar su entorno sin ningún tipo de entrenamiento o control humano". Aunque su visión era acertada, perfeccionarla tomó casi media década.

Avances en Reconocimiento de Patrones

En 1967, se desarrolló el algoritmo del KNN (K Nearest Neighbours), que permitía a las computadoras realizar una detección rudimentaria de patrones. Este algoritmo comparaba un nuevo objeto con los datos existentes y lo clasificaba como el vecino más cercano, es decir, el elemento más similar en la memoria. Este hito sentó las bases para el reconocimiento de patrones, una pieza fundamental en el desarrollo de la IA.

En los años 60, los científicos de los Laboratorios Bell intentaron enseñar a las máquinas a leer textos en inglés, lo que inspiró el término "aprendizaje profundo". El algoritmo de reconocimiento de patrones se atribuye a Fix y Hodges, quienes detallaron su técnica no paramétrica para la clasificación de patrones en 1951.

Este esfuerzo pionero fue una de las primeras tentativas de aplicar técnicas de aprendizaje automático para tareas lingüísticas y de comprensión de texto. Aunque estos primeros esfuerzos fueron limitados en comparación con las capacidades actuales del aprendizaje profundo, sentaron las bases para

investigaciones futuras en el campo del procesamiento del lenguaje natural y la comprensión de texto por parte de las máquinas.

Los científicos de los Laboratorios Bell estaban interesados en resolver problemas desafiantes en el campo de la inteligencia artificial y el procesamiento del lenguaje natural, y su trabajo en la enseñanza a las máquinas para leer textos en inglés fue parte de este esfuerzo más amplio. Aunque estos primeros intentos pueden haber sido rudimentarios en comparación con las tecnologías modernas, marcaron el inicio de la investigación en el uso de algoritmos de aprendizaje automático para tareas lingüísticas y de comprensión de texto, sentando así las bases para los avances futuros en el campo.

Durante la década de los 70, el campo de la IA enfrentó dificultades debido a las altas expectativas de los inversores y los pocos avances logrados. Sin embargo, el desarrollo del algoritmo "Nearest Neighbor" en 1967 marcó un avance significativo en el reconocimiento de patrones, incluso con aplicaciones prácticas en la planificación de rutas para vendedores ambulantes.

En los años 80, surgieron los sistemas expertos basados en reglas, lo que renovó el interés en el aprendizaje automático. Sin embargo, a finales de la década de los 80, comenzó un segundo y más prolongado "invierno" en la inteligencia artificial, que no se recuperaría completamente hasta bien entrados los 2000.

Innovaciones en la Década de 1980 y 1990

En 1981, Gerald Dejong introdujo el concepto de Aprendizaje Basado en Explicaciones (EBL), sentando así las bases para el desarrollo del aprendizaje supervisado moderno. Este enfoque revolucionario permitió a las computadoras analizar datos de entrenamiento y generar reglas generales para la toma de decisiones. Dejong fue pionero en esta metodología, que se convirtió en un precursor fundamental del aprendizaje supervisado, ya que los ejemplos de entrenamiento complementaban el conocimiento previo del mundo.

En esencia, el programa examina los datos de entrenamiento para discernir patrones y eliminar la información redundante, creando reglas amplias que se pueden aplicar a nuevas instancias. Por ejemplo, en el contexto del ajedrez, si se instruye al software para enfocarse en la reina, descartará automáticamente todas las piezas que no sean relevantes para esa tarea específica. Esta innovadora aproximación marcó un hito significativo en la aplicación de la inteligencia artificial tanto en ámbitos comerciales como académicos durante la década de 1980.

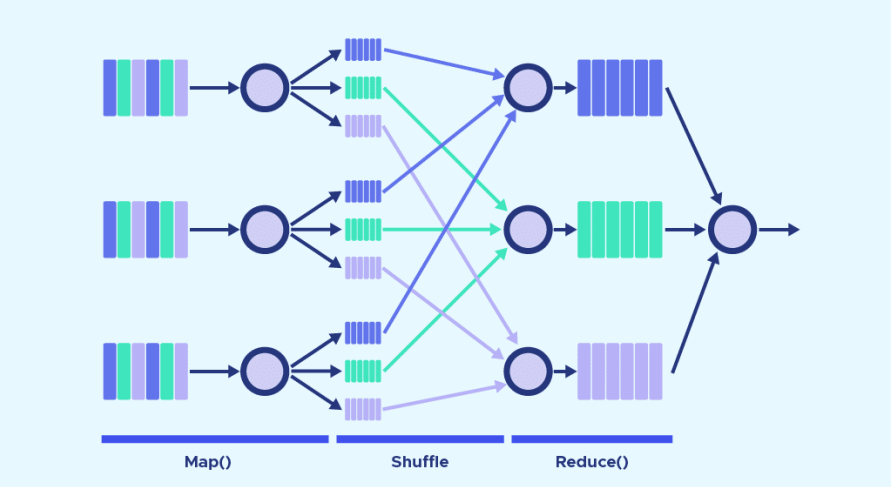
La era del Big Data y el auge del Machine Learning

En la década de 1990, se produjo una convergencia entre la informática y la estadística, lo que impulsó el desarrollo de enfoques probabilísticos en el campo de la Inteligencia Artificial. Este avance resultó fundamental para manejar grandes volúmenes de datos y propició un cambio significativo en el campo del aprendizaje automático. Durante esta época, los científicos comenzaron a aplicar el aprendizaje automático en diversas áreas como la minería de datos, el análisis de texto, la adaptación de software, entre otros. Esta década marcó el surgimiento de aplicaciones prácticas del aprendizaje automático en la vida cotidiana y en la industria. Los científicos desarrollaron programas informáticos capaces de analizar enormes conjuntos de datos y extraer conclusiones o aprender de los resultados obtenidos. Fue en este contexto cuando se acuñó el término "Aprendizaje Automático", para describir la capacidad de los programas informáticos para aprender y mejorar por sí mismos sin intervención humana directa.

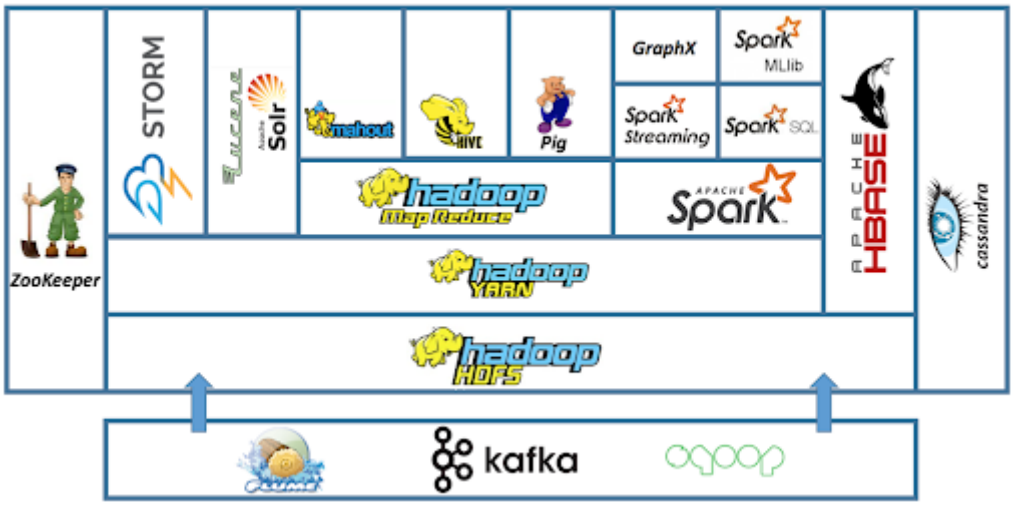
La revolución del Deep Learning

A partir del cambio de siglo, la noción de "Deep Learning", introducida por Geoffrey Hinton, transformó radicalmente el ámbito del Aprendizaje Automático al presentar estructuras de redes neuronales profundas. Este avance posibilitó que las computadoras llevaran a cabo tareas complejas como el reconocimiento de imágenes y el procesamiento de lenguaje natural con una precisión nunca antes vista.

En el año 2000, cuando el ML aún se recupera de un prolongado periodo de estancamiento, un nuevo protagonista está a punto de hacer su entrada en escena. Este protagonista comienza a tomar forma en 2003 con la publicación de un estudio sobre un sistema de archivos distribuidos conocido como 'Google File System' (GFS). Su definición se solidifica en 2004 cuando Google presenta un innovador paradigma de procesamiento distribuido denominado 'Map & Reduce'.



Mientras este protagonista emergente crece a pasos agigantados, su mentor, Google, continúa respaldándolo en su desarrollo al crear 'Cloud Bigtable', un servicio de bases de datos NoSQL para Big Data. Avanzamos hasta 2006, cuando los ingenieros de Apache materializan los paradigmas de Google en la primera plataforma de Big Data de código abierto, conocida como Hadoop.



De manera casi imperceptible, la capacidad de cómputo experimenta un crecimiento exponencial y se genera una abundancia de datos disponible. El principal protagonista, el ML, aprovecha esta situación de manera excepcional, logrando un desarrollo espectacular gracias al Big Data.

El viaje que emprenderá desde entonces estará marcado por éxitos y una transformación radical en su enfoque. Pasará de estar orientado principalmente al conocimiento ('knowledge-driven') a basarse completamente en los datos ('data-driven'). Este cambio redefine por completo la dirección y el alcance del ML hasta la actualidad.

Avances contemporáneos y aplicaciones prácticas

El nuevo milenio ha sido testigo de un notable auge en el campo de la programación adaptativa, donde el aprendizaje automático ha desempeñado un papel crucial durante mucho tiempo. Estos sistemas tienen la capacidad de identificar patrones, aprender de la experiencia y mejorar de forma autónoma en función de la retroalimentación recibida del entorno.

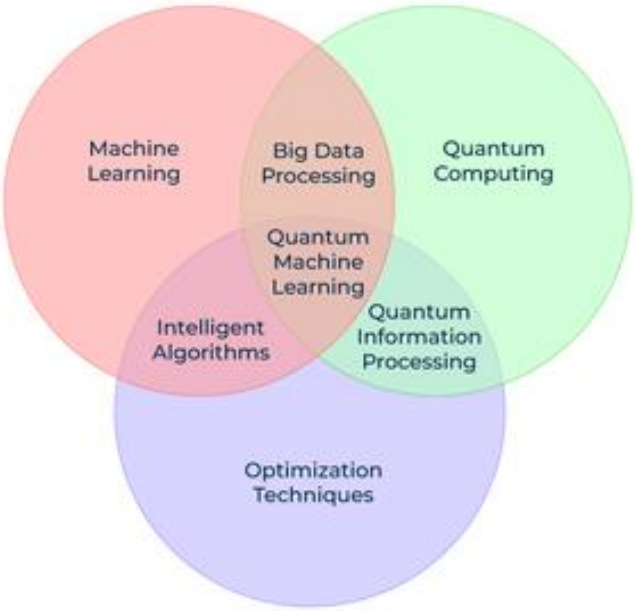
El aprendizaje profundo es un ejemplo destacado de programación adaptativa, donde los algoritmos pueden discernir y reconocer objetos en imágenes y vídeos, tecnología que subyace en iniciativas como las tiendas Amazon GO, donde los clientes pueden pagar al salir sin necesidad de hacer cola.

En la última década, importantes gigantes tecnológicos como IBM, Google, Facebook, Amazon y Microsoft han liderado el desarrollo de aplicaciones innovadoras de Machine Learning (ML). Desde sistemas de inteligencia artificial capaces de competir en juegos de preguntas y respuestas hasta avanzados algoritmos de reconocimiento facial, el ML ha transformado diversos aspectos de nuestra vida cotidiana.

Actualmente, nos encontramos inmersos en lo que podría denominarse como la tercera explosión del Machine Learning. Los avances tecnológicos han permitido que el aprendizaje automático encuentre aplicaciones en un amplio espectro de sectores empresariales, llegando incluso a generar mercados enteros y produciendo cambios significativos en las estrategias tanto de pequeñas como de grandes empresas.

Mirando hacia el futuro, el aprendizaje automático cuántico (QML) surge como un área de investigación prometedora que explora la interacción entre la computación cuántica y los métodos de aprendizaje automático convencionales. La aplicación de los ordenadores cuánticos en el aprendizaje automático tiene el potencial de acelerar significativamente los procesos y mejorar la generalización de los modelos con menos datos.

Es crucial explorar y comprender cómo aprovechar esta ventaja cuántica para impulsar aún más el campo del aprendizaje automático hacia el futuro.



A pesar de los avances espectaculares, el ML también plantea desafíos éticos y sociales que exploraremos en temas posteriores, como el uso responsable de la tecnología y la preocupación por el desarrollo de armas autónomas. Es fundamental abordar estas cuestiones para garantizar que el ML continúe beneficiando a la humanidad de manera positiva.

La evolución del Machine Learning desde sus modestos comienzos hasta su estado actual como una fuerza impulsora de la innovación tecnológica es un testimonio del ingenio humano y la capacidad de adaptación de la ciencia. A medida que exploramos las posibilidades infinitas del ML, es importante recordar su impacto transformador y trabajar juntos para aprovechar su potencial en beneficio de la sociedad.