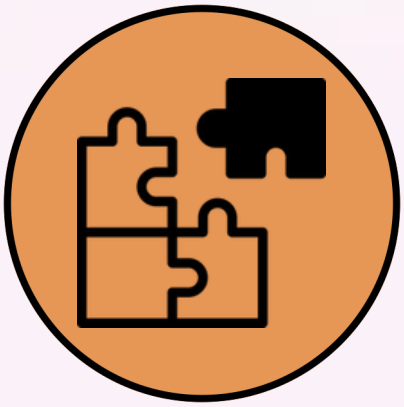
TIC

Conceptos básicos de ciencia de datos
LECCIÓN 1

Para la primera lección es necesario que los estudiantes debatan los conceptos previos sobre datos y que investiguen cuantos datos de audio y video se generan diariamente (un aproximado que puede estimarse por número de twits diarios, numero de videos que se suben al día a plataformas como youtube o alguna red social entre otros). Como ejercicio antes de la sesión, proponga a los estudiantes que intenten calcular cuantos discos duros de 1 Terabyte son necesarios para almacenar tal cantidad de información.

Conjunto de datos de mundiales de fútbol (https://github.com/openfootball/worldcup/blob/master/2018--russia/cup_finals.txt)

Desarrollo de la sesión (6 horas)



Inicie la sesión con la actividad 1 (diseñada para hacerse en 2 horas) Una vez terminadas las dos horas de sesión, inicie el segundo bloque de dos horas discuta con los estudiantes las tres preguntas planteadas, el objetivo de las preguntas es tener una discusión con los estudiantes acerca de cómo los sistemas de cómputo pueden acelerar el proceso de desarrollo de tareas de conteo, estadística y análisis de la información, enfatizando en cómo una tarea que para un humano puede ser compleja (contar varias veces en datos que no tienen una estructura muy clara) para un sistema de cómputo cuesta menos de un segundo. Hablar de cómo los sistemas de cómputo han evolucionado hasta la capacidad de hoy en día que permite ejecutar modelos de aprendizaje de máquina y análisis de muchos datos en poco tiempo.

Análisis de datos

Los datos son números asignados de manera objetiva a una propiedad o característica de un objeto o un evento. Cuando nos referimos a una forma objetiva, nos referimos a que cualquier observador con un instrumento adecuado puede asignar ese mismo número a una misma propiedad o a un mismo evento. Por ejemplo, queremos determinar la altura de un árbol. Para esto podemos hacerlo con diferentes sistemas de medición, por ejemplo, podemos utilizar el sistema métrico internacional, con el cual asignamos un número a la altura del árbol, (por ejemplo 5 metros, que está en el sistema métrico internacional). Sin embargo, también podemos medirlo en sistema imperial (la medición se realiza en pies o yardas) o incluso podemos medirlo de forma no estandarizada, por ejemplo, medir la altura con algún elemento de referencia (medir con nuestra propia mano 15 manos de altura) lo cual también es una forma de medir, imprecisa, pero se acercaría al número asignado a la propiedad altura del árbol.

Encontramos, entonces que:

Las mediciones y conteos o datos nos permiten conocer propiedades de objetos (altura, ancho, peso, masa, entre otros) y las características de eventos (duración, tiempo de inicio, fechas, trabajo, energía, velocidades, aceleraciones, entre otros).

Las mediciones y conteos o datos nos permiten conocer propiedades de objetos (altura, ancho, peso, masa, entre otros) y las características de eventos (duración, tiempo de inicio, fechas, trabajo, energía, velocidades, aceleraciones, entre otros).

Símbolo	Unidad	Magnitud
s	segundo	tiempo
m	metro	longitud
kg	kilogramo	masa
A	amperio	corriente eléctrica
K	kelvin	temperatura termodinámica
mol	mol	cantidad de sustancia
cd	candela	intensidad luminosa

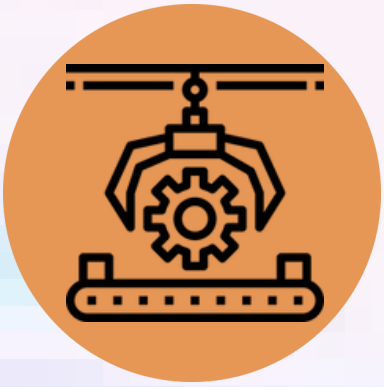
Tabla 1: Unidades base del sistema internacional de Unidades.

Generalmente los datos vienen representados por una unidad de medida para conocer las magnitudes de forma estandarizada.

Desde la revolución tecnológica a mediados de la década de 1970 con la invención del transistor hubo un auge de los sistemas electrónicos. Estos sistemas permiten adquirir datos y tomar medidas de forma automatizada, ayudando a que las líneas de producción y la industria en general se vea beneficiada de conocer cómo suceden los procesos a partir de los datos.

Por ejemplo,

El departamento contable de una empresa no podría realizar balances sin los datos de cuantas unidades se produjeron a partir de el valor y la cantidad de insumos adquiridos para la fabricación. Sin datos, una máquina embotelladora no podría determinar cuando una botella se encuentra llena para pasar a la siguiente en ser llenada.



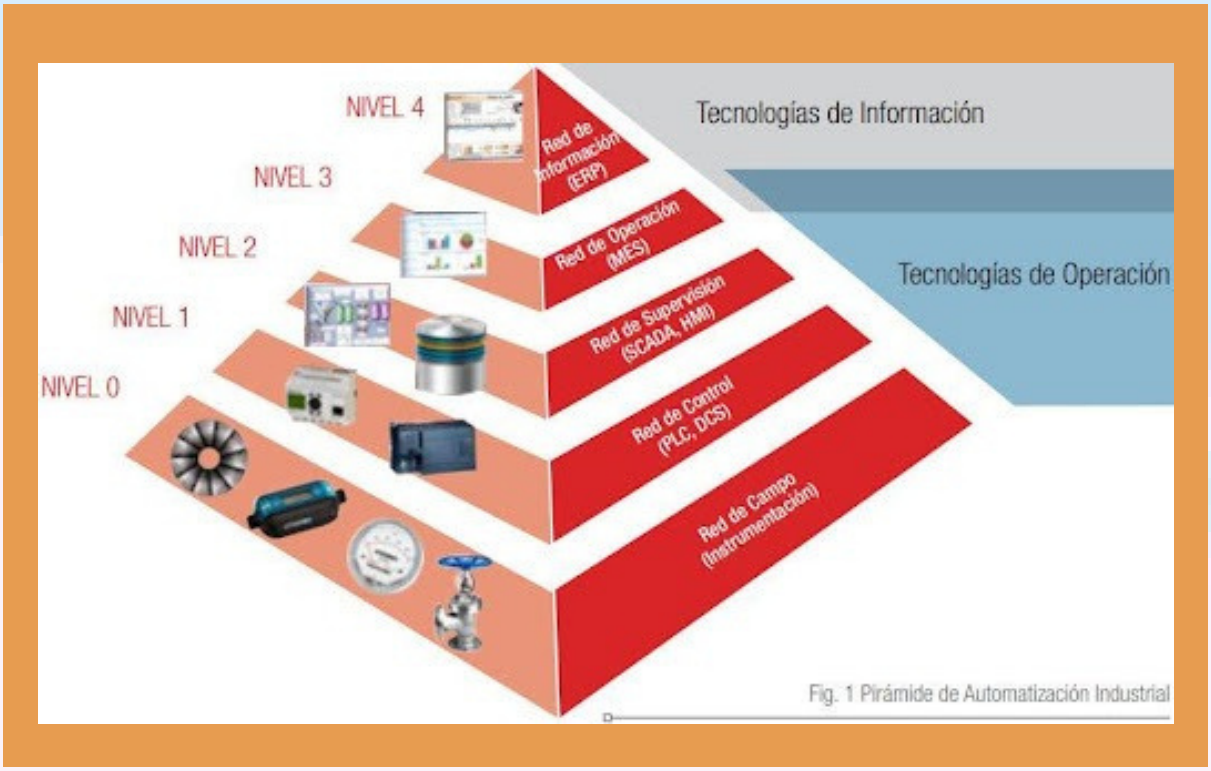
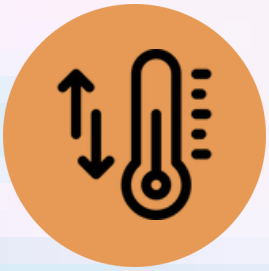


Figura 1: Pirámide de la automatización (tomada de la universidad veracruzana).

En la figura 1 se muestra el concepto de la pirámide de la automatización en cuyos niveles se aprecia la gestión tecnológica y de datos que aparece según la complejidad del nivel.

En el nivel 0 se tienen los datos de campo o de instrumentos como pueden ser: velocidad de motores, temperaturas, medida de cantidades, y todo lo que proviene de sensores y actuadores en un proceso de producción industrial.



En el nivel 1 se presenta una jerarquía de agrupación de los datos, como el caso de los controladores o PLC, que son dispositivos que recogen los datos de los procesos y toman decisiones de mayor jerarquía, por ejemplo, escoger entre el proceso de fabricación bebidas gaseosas de 200 ml y botellas de 550 ml. Este cambio supone ajustes en ciertos parámetros, sin tener en cuenta cada detalle de cada sensor de forma independiente. Es decir, los procesos individuales de control quedan en el primer nivel de la pirámide.

En el nivel 2 se tienen las redes de supervisión y sistemas con interfaces humano máquina. Estas estaciones permiten agrupar datos productivos y de operaciones y mostrar mucha información de forma entendible por humanos. Es decir, en lugar de mostrar una tabla con miles de datos, se generan gráficos y análisis que permiten determinar rápidamente el estado de los procesos controlados.

En el nivel 3 se tienen las operaciones de fabricación, donde se toman ciertas decisiones administrativas, que beneficien los procesos de producción según los datos, como puede ser contratar mas personas, comprar máquinas, cambiar partes de los procesos. Todo basado en los datos de rendimiento y productividad (provenientes de los niveles anteriores).

En el nivel 4 tenemos los niveles de gestión del negocio, como es mantener los costos de compras, costos de ventas, diferentes puntos de venta y de fabricación. Es donde se toman las decisiones administrativas importantes, de cara a la rentabilidad de los negocios.

En la última década,

Se ha agregado un quinto nivel alineado con el desarrollo tecnológico asociado al big data (grandes cantidades de datos) y el aprendizaje de máquina, denominado inteligencia de negocios. En este nivel, se analizan grandes cantidades de datos generados de múltiples procesos, con el fin de tomar decisiones de alto nivel. Uno de los ejemplos apropiados es las máquinas de coca cola freestyle. Estos dispositivos representan un gran ejemplo del uso de big data en sistemas automatizados.

Las máquinas freestyle son dispensadores de bebidas que permiten a los usuarios mezclar bebidas gaseosas a partir de 15 sabores diferentes. En cada una de las máquinas se recolectaba la información de las ventas realizadas y de las preferencias de cada usuario para hacer las mezclas de bebidas. La empresa coca cola instaló más de 15.000 de estas máquinas en Estados Unidos y analizó como cambiaban los hábitos de consumo, determinando que la gente bebía mas los fines de semana y lo hacía menos en días festivos. Con estas máquinas se pudo determinar los hábitos de consumo en cada ubicación de cada tipo de bebida, por ejemplo, la Sprite se consumía mas en ciertos estados que la Fanta, y en el centro de Nueva York los dispensadores tenían más uso que en todo el país. Con lo que la empresa pudo instalar más y aumentar sus ventas. También encontraron que, en Georgia, la gente bebía más en horas de la noche en los restaurantes de 24 horas. 5.

Pero el dato más importante es la mezcla más popular en todo el país.

Se descubrió que la gente solía servir en sus vasos una mezcla de 33% Coca-Cola 33 % bebida de cereza y 34% Coca-Cola de vainilla. Con este dato se lanzó el sabor coca cola Cherry vainilla el cual en su primer año de lanzamiento vendió más de 2 millones de dólares. Todo esto a partir solamente de los datos de las máquinas freestyle. El punto de este caso de estudio es resaltar la importancia de los datos para la toma de decisiones de alto nivel y la inteligencia de negocios.

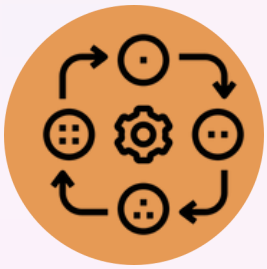
De la misma forma han aumentado su productividad y ventas empresas como amazon, Netflix, apple, y muchísimas más, con los datos de sus dispositivos y sus aplicaciones o con datos de ventas y el análisis de los hábitos de consumo de las personas.

Otro ejemplo es el análisis de hábitos de compra a través del big data. La recopilación de datos masivos provenientes de diversas fuentes, como transacciones en línea y redes sociales, permite a las empresas identificar patrones y tendencias. Esta información posibilita la personalización de estrategias de marketing, la mejora de la experiencia del cliente y la anticipación de las demandas del mercado. Tal es el caso de Target (cadena de supermercados), esto sucedió en Minneapolis, donde un padre se enteró por Target que su hija adolescente estaba embarazada antes que él mismo.

Luego de comprar en Target, la chica comenzó a recibir publicidad en casa de su padre anunciando productos para bebés: pañales, ropa, cunas y otros artículos específicos. El padre, indignado por el intento de la tienda de "incentivar" el embarazo adolescente, se quejó con la gerencia. Días después, el mismo hombre llamó para disculparse, pues descubrió que su hija sí estaba embarazada. Esto no es raro en la era de la recolección de datos computarizados por parte de las tiendas minoristas. Target asigna a cada cliente un "ID de invitado" vinculado a su nombre, tarjeta de crédito, correo electrónico y cualquier otra información que puedan recopilar. Usando el historial de compras, Target crea un perfil sorprendentemente preciso para personalizar la publicidad, empleando todos los datos disponibles de las personas tanto en sus bases de datos como en sus hábitos de compra y cookies del explorador. Esto también puede abrir la puerta a un debate sobre la ética en el uso de los datos de las personas. Por este motivo también es importante comprender a qué aplicativos le damos acceso a nuestros datos y qué hacen las grandes empresas con los datos que recopilan en nuestras redes sociales, teléfonos, televisores y otros dispositivos que toman datos de nuestro comportamiento.

Fases del análisis de datos

Todo proceso de análisis de datos puede tener múltiples fases, dependiendo de la necesidad de procesar los datos. En este curso se aprenderá qué proceso llevan los datos en cada fase y la necesidad de cada una de ellas, dependiendo de cómo los datos estén representados. En primer lugar, se tiene la recopilación de datos, el cual es el proceso que consiste en obtener los datos necesarios para el análisis. En la actividad 1 de la lección partimos desde algunos datos previamente recopilados, en donde, alguien revisó los registros históricos de diferentes fuentes y recopiló la información en un archivo de texto que puede o no estar ordenado. Sin embargo, contiene todos los datos.



En segundo lugar, se tiene la fase de limpieza de los datos, algo similar a lo hecho en la actividad 1 en donde, se toman los datos recopilados de algún formato y se transcriben o copian a un segundo formato ordenado de acuerdo con la necesidad del análisis. En nuestro caso limpiamos los datos y solo nos quedamos con los goles marcados en los partidos finales y con los países participantes.

En tercer lugar, se tienen la integración de datos. Esta etapa sucede porque los datos pueden provenir de diferentes fuentes y se necesita crear un análisis que permita visualizar y entender fenómenos. Los datos combinados serán almacenados en una única base de datos.

En cuarto lugar, se tienen las transformaciones de datos, en las que los puntos de datos individuales se convierten y se procesan utilizando el conocimiento de lo que se está analizando o "lógica de negocio". En este proceso intervienen expertos en la lógica de negocio, que ayudan a los analistas de datos a determinar las transformaciones que los datos deben tener. Los datos transformados se convierten en información.

En quinto lugar...

Se tienen las etapas de análisis y modelado. Una vez los datos fueron transformados, se procede con visualizaciones, reconocimiento de patrones, búsqueda de tendencias, análisis de correlación entre las variables, y análisis estadísticos que permiten conocer lo que los datos están modelando y entender los detalles del fenómeno observado.

En la etapa de análisis y modelado se tienen cinco tipos de análisis que son análisis descriptivo, exploratorio, inferencial, predictivo y de segmentación.