Big Data

Introducción

Docente: Ph.d Laura Mónica Escobar V.

2024

Grupo de
Planeamiento en
Sistemas Eléctricos

Descripción y Contenidos

1. Breve descripción

Esta asignatura hace parte en la formación en ciencia de datos como área específica de aplicación que tiene como objetivo recopilar las técnicas asociadas al procesamiento de grandes volúmenes de datos. Esta área abarca sistemas operativos, arquitecturas cliente servidor, ecosistemas informáticos en la nube, lagos de datos, servicios y micro servicios, procesamiento en tiempo real y casi tiempo real.

2. Objetivos

OA1: Definir y explicar la terminología asociada a la ciencia de datos
OA2: Definir y explicar los conceptos y paradigmas asociados a big data
OA3: Explicar los modelos y ecosistemas asociados a big data
OA4: Implementar modelos y ecosistemas para el procesamiento de grandes volúmenes de información
OA6: Establecer espacios de discusión acerca del procesamiento de grandes volúmenes de información
OA7: Construir modelos de procesamiento de grandes volúmenes de datos que permitan resolver problemas reales

3. Resultados de aprendizaje de la Asignatura (RAA)

RA1: Explicar correctamente el concepto de ciencia de datos y las disciplinas asociadas a este
RA2: Explicar correctamente paradigmas asociados a big data
RA3: Explicar el concepto de ecosistema informático en el ámbito de big data
RA4: Comprender la diferencia entre almacenamiento, procesamiento y pos procesamiento de datos
RA5: Diferenciar las técnicas y tipos de procesamiento en big data
RA6: Seleccionar las técnicas y tipos de procesamiento adecuados para los procesos de almacenamiento y producción de datos
RA7: Solucionar problemas prácticos para datos de gran volumen utilizando herramientas de big data y ciencia de datos
RA8: Diferenciar los roles en las distintas instancias de construcción de productos de datos

4. Contenido

MÓDULO 1: Conceptos fundamentales de ciencia de datos. (6 horas)
MÓDULO 2: Conceptos fundamentales de big data (6 horas)
MÓDULO 3: Introducción a ecosistemas big data (12 horas)
MÓDULO 4: Introducción a bases de datos NoSQL (12 horas)
MÓDULO 5: Implementación de procesamiento de datos en entornos big data (16 horas)
MÓDULO 6: Arquitecturas big data y casos de estudio (4 horas)

7. Herramientas técnicas de soporte para la enseñanza

Clases magistrales, apoyados en video-beam. Software libre para configuración de servidores, herramientas de desarrollo y motores de bases de datos NoSQL. Se emplearán servidores de uso libre para las prácticas y talleres.

8. Trabajos en laboratorio y proyectos

- Laboratorio de configuración de servidores sobre arquitectura big data
- Laboratorio de configuración y uso de entorno de procesamiento de bases de datos NoSQL
- Laboratorio de herramientas de ecosistema big data para el procesamiento de datos
- Trabajo final de curso.

9. Métodos de aprendizaje

Se utilizará una metodología constructivista basada en proyectos. El estudiante deberá acceder a repositorios públicos de datos, procesar, caracterizar y construir modelos de procesamiento sobre grandes volúmenes de datos. Adicionalmente deberá hacer uso de estrategias de integración de herramientas, plataformas y servicios sobre los resultados obtenidos. Esta tarea se realizará de forma incremental planteando la necesidad de implementar distintos modelos de análisis debido a la variedad de escenarios donde se desarrollan los datos.

10. Métodos de evaluación

Se desarrollarán evaluaciones que permitan la verificación de cada uno de los resultados de aprendizaje planteados. Estas evaluaciones estarán distribuidas en 3 trabajos que se desarrollarán a lo largo del curso.

1. Evaluación 1: Examen escrito acerca de conceptos de ciencias de datos, big data, modelos de base de datos. 40%
2. Evaluación 2: Diseño e implementación de un modelo de procesamiento de datos, bases de datos NoSQL, uso de exosistemas de Big data y procesamiento de datos. 40%
3. Análisis y presentación ante el grupo de un articulo actualizado de la literatura especializada, relacionado con el implementación de Big data en la solución de problemas actuales. 20%

Introducción

En la actualidad la población a evolucionado a un punto en el cual el uso de la tecnología es indispensable.

Acceso a redes sociales, redes de comunicación, generación de ordenes a control remoto, educación virtual y uso de centros de almacenamiento de datos.

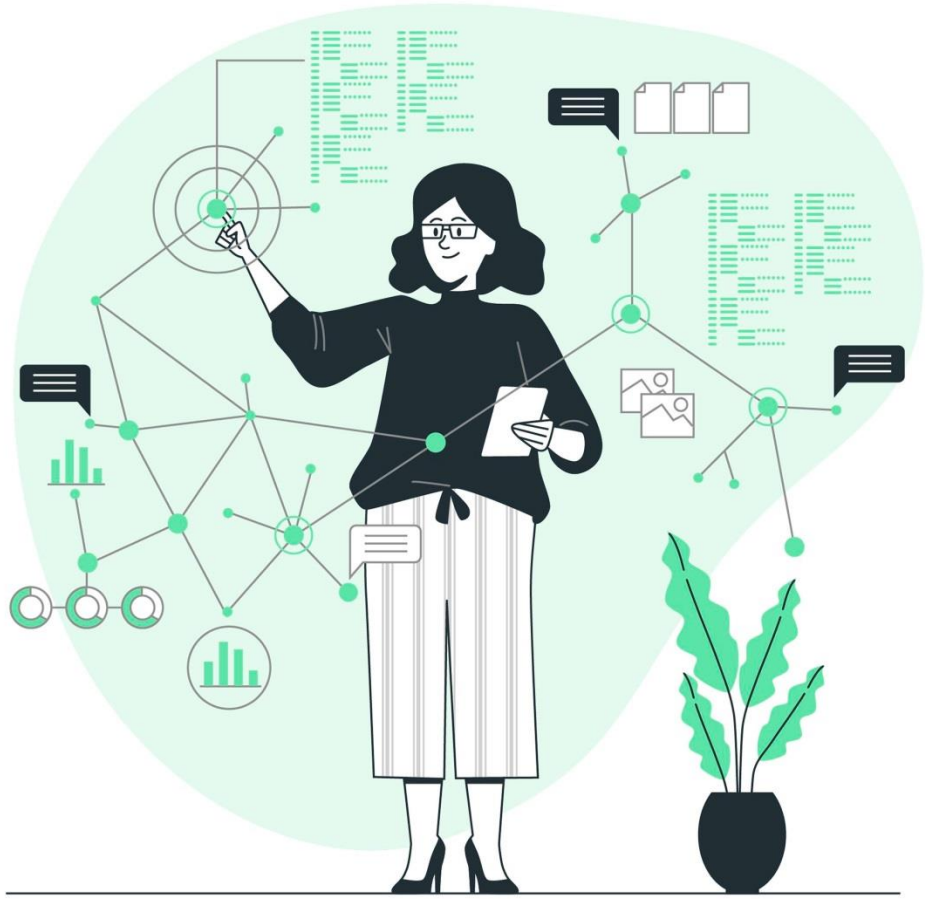
Esto genera un ingreso de grandes cantidades de datos, con diferentes estructuras y contextos.

Introducción

La gran entrada de información en formatos diferentes y que con el paso del tiempo aumentan en tamaño y complejidad genera una serie de preguntas:

- ***¿A que se debe su incremento en los últimos años?***
- ***¿Qué implica este cambio?***
- ***¿Por qué es necesario el Bid data?***

Comencemos desde el inicio, debemos buscar la razón por la cual se genero un cambio en la cantidad y tipo de datos que entrega un usuario promedio.



Introducción

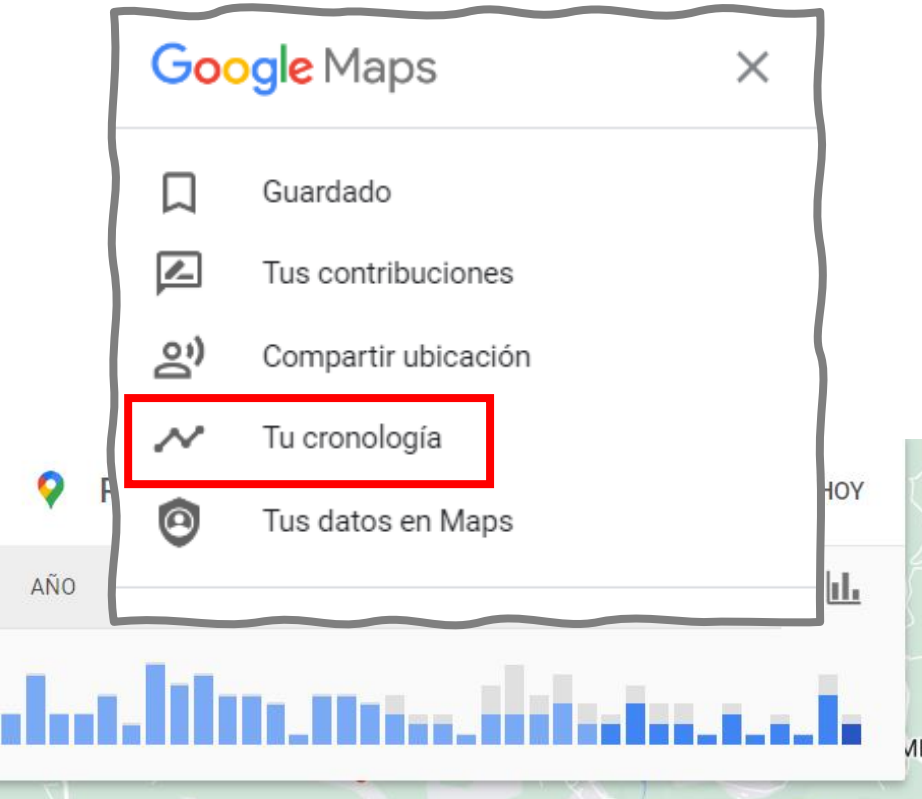
¿A que se debe el incremento en el volumen y complejidad de la información en los últimos años?

Con la aparición de los dispositivos electrónicos, dispositivos móviles, sensores, cámaras de mayor resolución, acceso a internet y acceso a software de forma libre, los usuarios encontraron la manera de comunicarse y compartir sus experiencias de forma mas completa y compleja para el entorno virtual.

Por lo tanto ahora tenemos imágenes con mayor resolución, un volumen mas grande de imágenes y una mayor necesidad por capacidad de memoria para su almacenamiento.

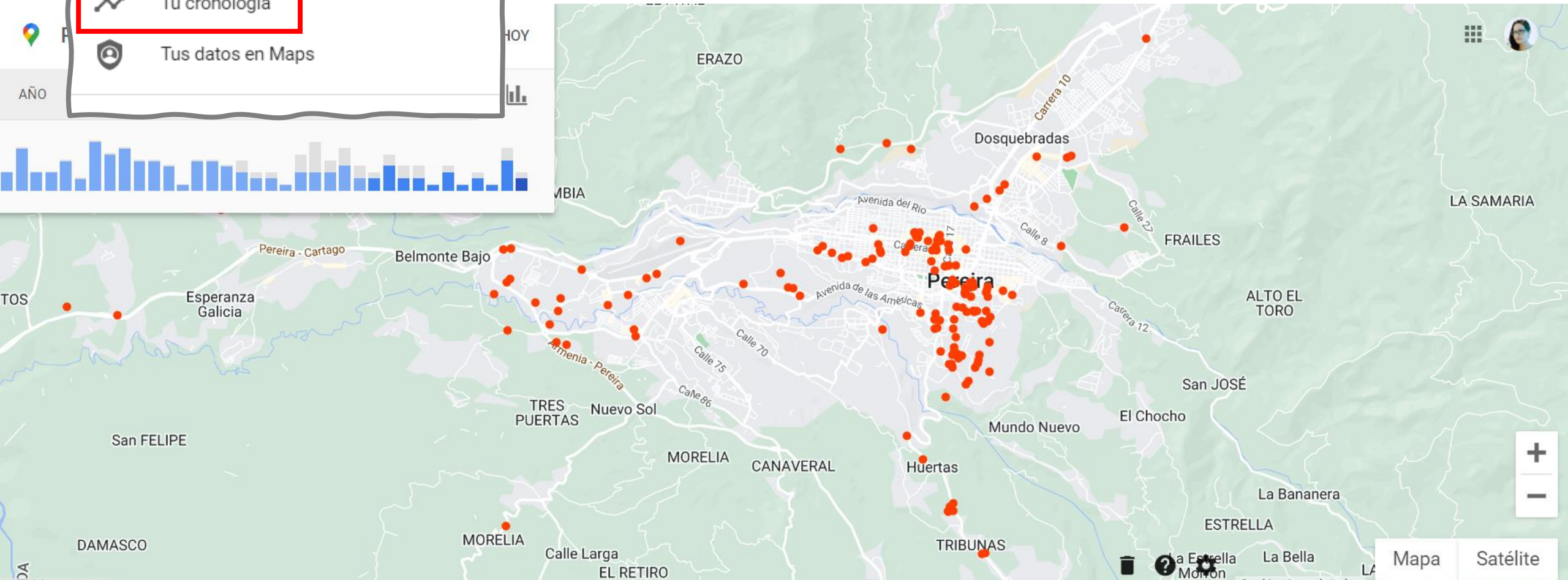
Podemos ver el cambio mas grande en el comportamiento generado por el avance de la tecnología, en la información entregada por los usuarios día a día.

De unas vacaciones de corto tiempo, se puede obtener una gran cantidad de información del usuario.



En muchas ocasiones el usuario acepta condiciones que hacen parte de ciertas aplicaciones que le permite a la empresa tener acceso a una gran cantidad de información.

Por ejemplo al aceptar que Google haga un seguimiento de su cronología, guardara la información relacionada con tu ubicación diariamente.



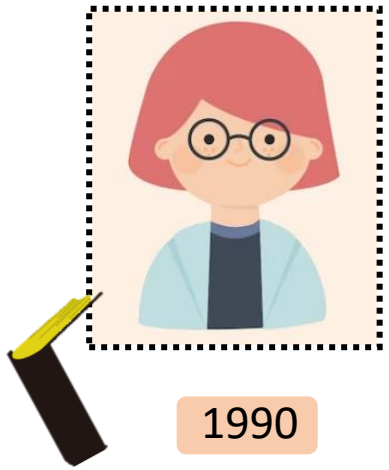
Con el acceso a internet, se facilita también conocer los hábitos y gustos de los usuarios y con esto organizar las tareas diarias, encontrar nuevas recetas, herramientas o productos que satisfagan nuestras necesidades.

Al punto tal que con una simple búsqueda en Google, llegara a nuestras cuentas privadas información de productos y servicios relacionados.

Introducción

¿Qué implica este cambio?

Ejemplo practico: Análisis comparativo entre la cantidad de datos encontrados por usuario en un periodo de 30 años.



1990

Nombre

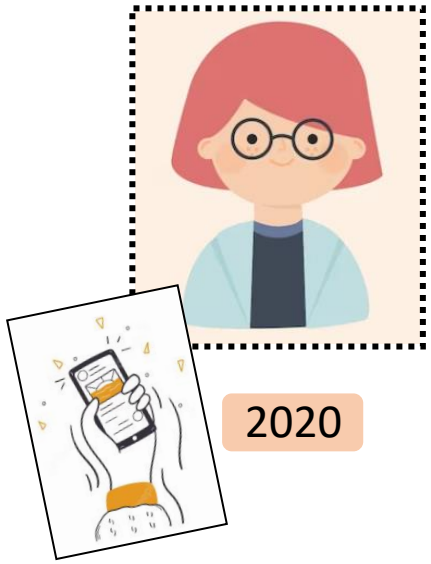
Teléfono fijo

Dirección

Cédula

Edad

Nacionalidad



2020

Nivel de educación

Pasaporte

Nombre

Teléfono fijo

Dirección

Cédula

Celular

Edad

Gustos

Correo electrónico

Historial de compras

Ubicación en tiempo real

Introducción

El ingreso de una mayor cantidad de información genero una nueva problemática: El análisis y almacenamiento de grandes cantidades de datos.

¿Si el usuario no se encuentra conscientemente entregando su información personal para su uso y analisis, como se obtiene?

- Que has buscado y a donde has viajado
- Información personal y relaciones personales
- Que sitios web visitas
- Las tiendas en donde compras
  Que marcas compras más
- Muestra que tan estable eres emocionalmente
- Que series y temas te gustan
- Que gustos tienes en música o en videos de entretenimiento.

Introducción

¿Por qué es necesario el Big data?

Por lo tanto, para mejorar los procesos de generación y prestación de servicios, las empresa deben trabajar con una mayor cantidad de información, muchas veces no estructurada, y en grandes volúmenes. Por lo tanto la entrada de la ciencia de datos y el Big data a este tipo de análisis, permite el uso de nuevas herramientas y la formación de nuevas estrategias.

Dentro de las estrategias de mejora, se encuentra el almacenamiento de información y la clasificación de los datos, por lo cual, dentro del curso se deben revisar ambos casos de forma detallada.

Introducción

Para comprender el proceso de análisis que lleva al uso de herramientas de Big data, para analizar y utilizar correctamente los datos, es necesario comprender que es y que implica

Definición de dato:

De forma muy general, un **dato** es la representación de una variable o característica que puede ser cuantitativa o cualitativa y que a su vez que indica un valor.

Sin embargo un dato no depende solo de su valor, también depende de un contexto o características que permitan ubicarlo en el espacio de análisis y comprender su comportamiento y posible uso.

Introducción

Para entender mejor el concepto de dato y como puede ser transformado en información útil, vamos a tomar un valor aleatorio:

28

El dato presentado por si solo no tiene significado, como el valor seleccionado. ¿ Que significa el valor 28? ¿28 vehículos comprados en Pereira? ¿28 teléfonos? ¿28 sacos de harina?

Existen muchas características diferentes que pueden estar relacionadas con este valor.

Es necesario entonces, generar un atributo o un contexto al valor seleccionado, para así conseguir que ese valor o ese dato que hemos extraído, se convierta en información útil para un análisis.

Introducción

Numero de ventas de **Legend of Zelda: Breath of the Wild**: 28 Millones

Ahora el valor tiene un contexto, me indica el numero de copias vendidas del juego Legend of Zelda: Breath of the Wild.

Sin embargo para un analisis que busca generar mejoras, esta información sigue siendo insuficiente. No tengo suficientes datos complementarios que se encuentren relacionados con el dato principal de analisis, que me permita conocer mas sobre él.

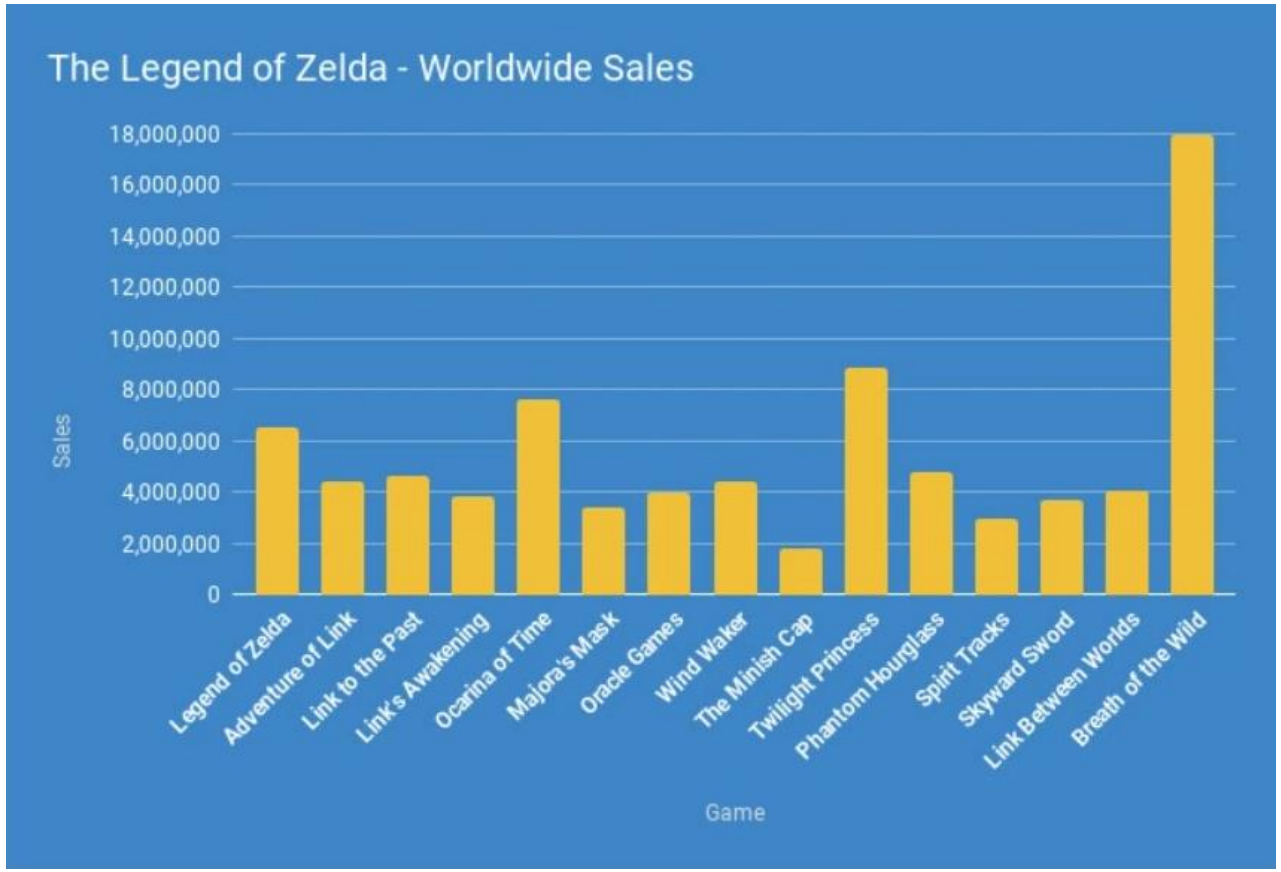
Adicionemos una fecha, para mejorar la información:

Numero de ventas de Legend of Zelda: Breath of the Wild: 28 Millones
Copias vendidas hasta noviembre del 2022.

Introducción

Si tenemos la posibilidad de obtener otros datos, como tipo de consola, ventas por país, etc. La información se convierte en una base suficiente para su clasificación y análisis.

Como ejemplo, podríamos realizar la comparación del numero de ventas obtenidas por la ultima entrega de **Legend of Zelda: Breath of the Wild**, revisar que tan bueno a sido su comportamiento en ventas y que relación tienen las mejoras de esta entrega con su desempeño a comparación de las anteriores. Como caso de estudio tenemos la comparación realizada en el 2020.



Introducción

Esta información es necesaria para realizar un analisis que permita generar mejoras o nuevos servicios. Por lo tanto, dependiento de los datos que tenemos disponibles es posible clasificar el problema de diferentes estados de tiempo y generar analisis distintos:

Pasado

Futuro

¿Que paso?

Reportes
Históricos

¿Porque paso?

Análisis forence
Data Mining

¿Qué esta pasando?

Análisis en tiempo real

¿Por qué esta pasando?

Análisis de Data Mining en tiempo real

¿Qué podría pasar?

Análisis predictivo

¿Qué se debería hacer?

Análisis prescriptivo

Introducción

Con el uso de la información de forma ordenada es posible:

- Segmentación de clientes
- Descubrimiento de patrones
- Detección de fraude
- Análisis del mercado
- Modelo predictivos

Fundamentos de la ciencia de datos

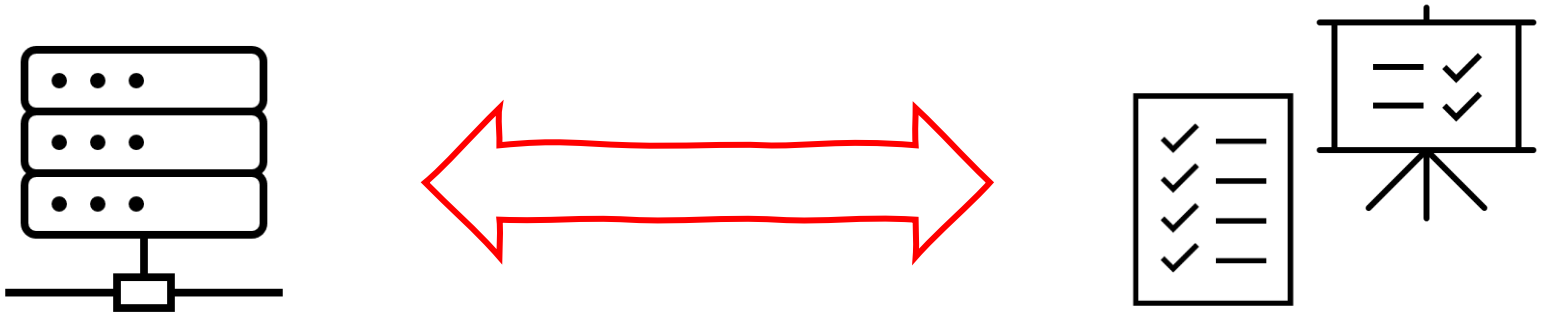
La ciencia de datos es un campo interdisciplinario que se enfoca en el estudio y la extracción de conocimiento a partir de los datos. Combina conocimientos de **estadística**, **matemáticas**, **informática**, **visualización de datos** y otras disciplinas para analizar grandes conjuntos de datos y extraer insights útiles que puedan ser utilizados para la toma de decisiones informadas. La ciencia de datos se utiliza en una amplia variedad de aplicaciones, desde la industria hasta la investigación académica.

Ejemplo practico: Las empresas pueden utilizar la ciencia de datos para analizar los datos de ventas y clientes con el objetivo de mejorar sus procesos empresariales y la experiencia del cliente.

- El proceso de ciencia de datos comienza con la identificación de un problema o pregunta de investigación.
- Se recolectan y preparan los datos para su análisis.
- Se aplican técnicas estadísticas y de aprendizaje automático para analizar los datos y descubrir patrones y tendencias.
- Finalmente, se presentan los resultados de manera clara y comprensible para los tomadores de decisiones.

Fundamentos de la ciencia de datos

Desde un entorno real, enfocado a casos de estudio tenemos:

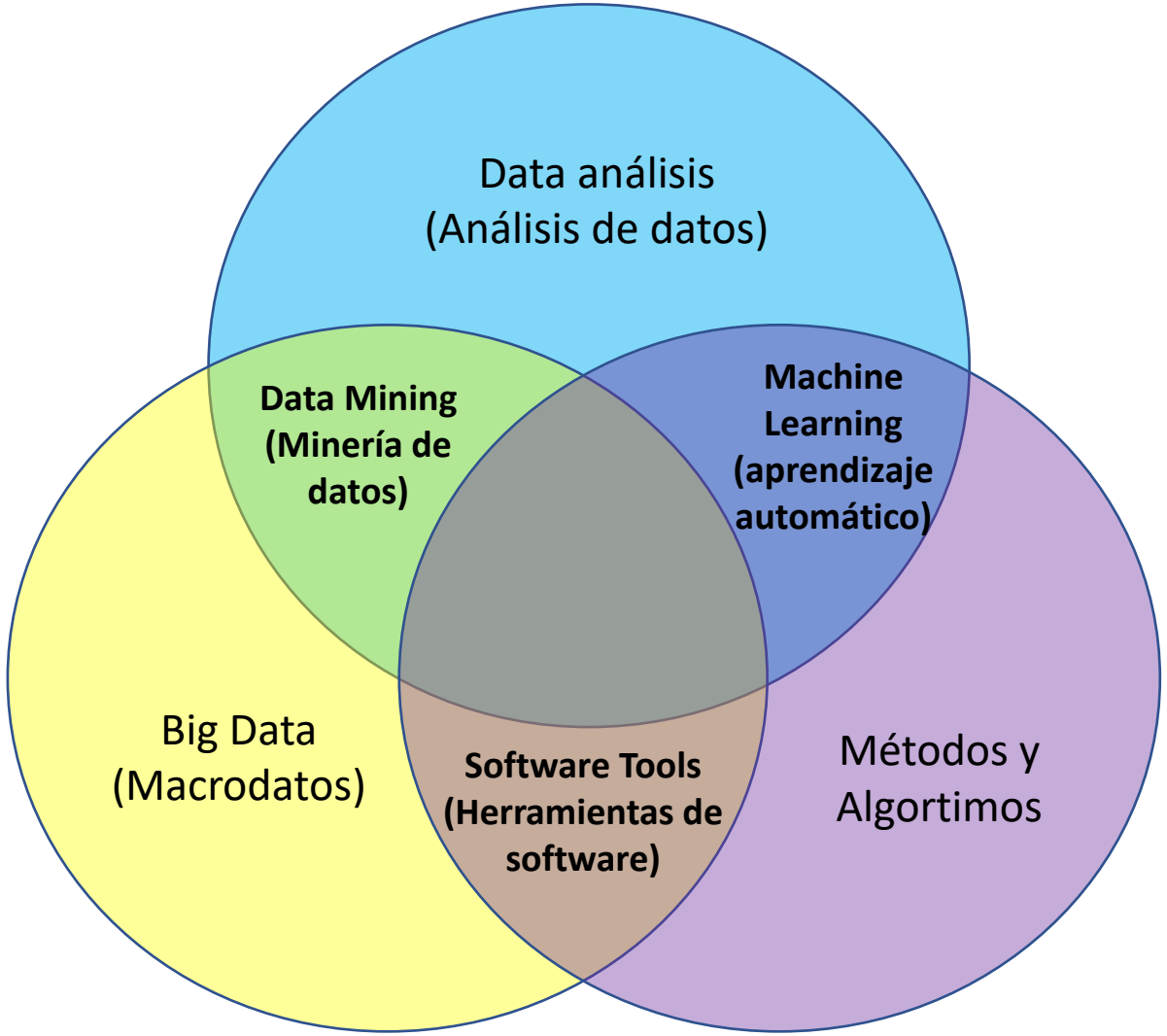


Datos

Objetivos

Nuestro objetivo es enlazar los datos de interés con los objetivos del negocio permitiendo resolver problemas o apoyar en la toma de decisiones. La ciencia de datos prospera en la intersección de conocimiento de la informática, las matemáticas y la experiencia estratégica. Los investigadores pueden utilizar la ciencia de datos para analizar grandes conjuntos de datos en el campo de la salud, la biología y otras disciplinas.

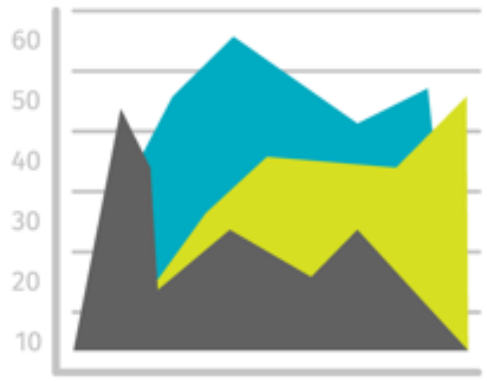
Fundamentos de la ciencia de datos

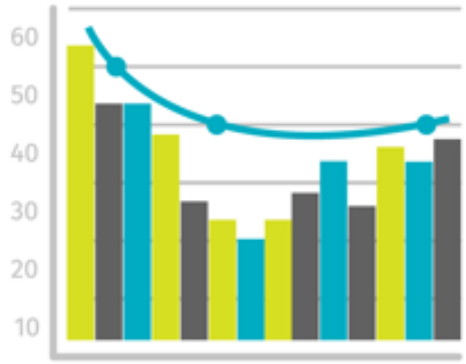


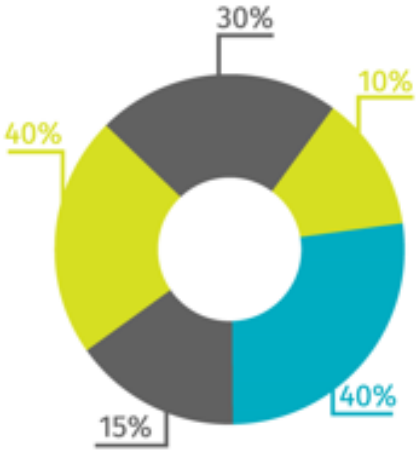
Fundamentos de la ciencia de datos

Por lo tanto el estudio debe apoyarse en los conocimientos que permiten la extracción, manejo y análisis de la información. Vamos a revisar con mas detalle lo que implica cada uno de estos puntos clave:

Conocimiento Estadístico: La estadística es una rama de las matemáticas que se enfoca en el análisis y la interpretación de datos. Es fundamental en la ciencia de datos, ya que proporciona las herramientas y técnicas para analizar los datos y hacer inferencias basadas en ellos. Los analistas de datos utilizan técnicas estadísticas para identificar **patrones**, hacer **predicciones** y tomar decisiones informadas.







Fundamentos de la ciencia de datos

Minería de datos: La minería de datos se refiere al proceso de descubrir patrones y tendencias en los datos a través del uso de técnicas estadísticas y de aprendizaje automático.

Los analistas de datos utilizan la minería de datos para descubrir insights ocultos en grandes cantidades de datos.

Esto les permite identificar patrones que pueden ser utilizados para mejorar los procesos empresariales, la toma de decisiones y el rendimiento general de la organización.

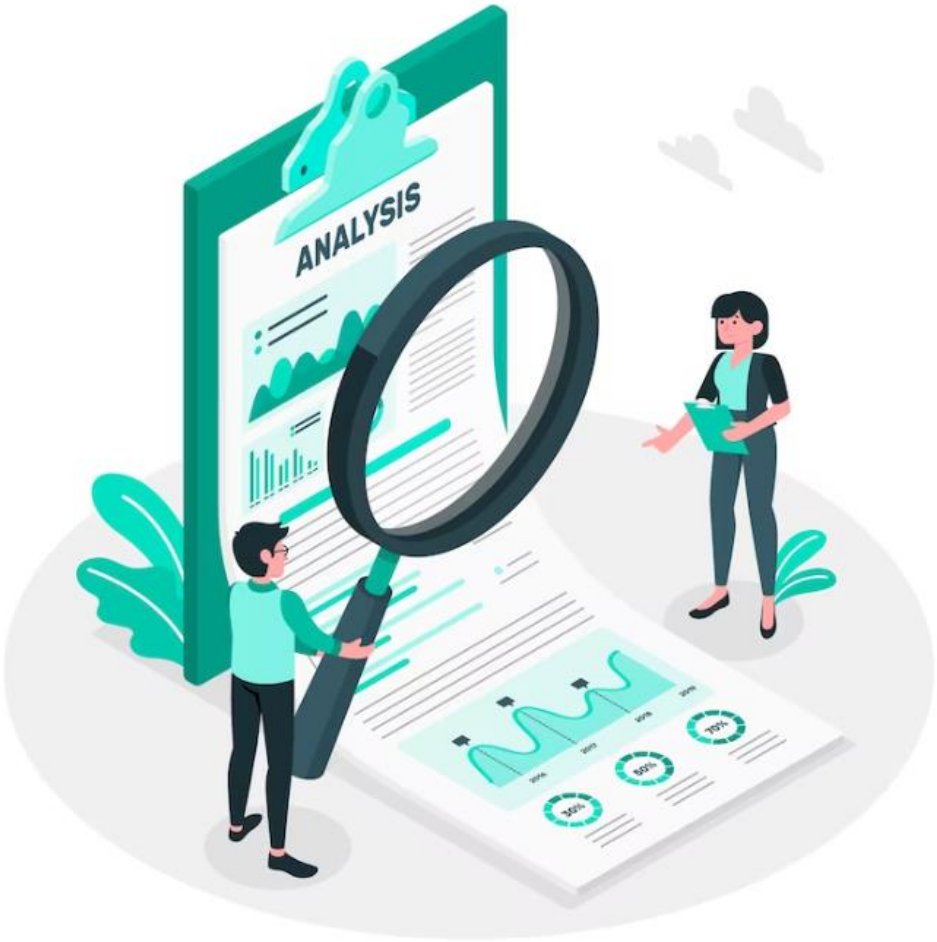
Fundamentos de la ciencia de datos

Visualización de datos: La visualización de datos es una técnica que se utiliza para representar gráficamente los datos y hacerlos más comprensibles.

Es importante en la ciencia de datos porque permite a los analistas y tomadores de decisiones entender los datos de manera más clara.

Las visualizaciones de datos pueden ser utilizadas para identificar patrones, tendencias y outliers en los datos.

Fundamentos de la ciencia de datos



Análisis de datos: El análisis de datos es el proceso de examinar los datos con el objetivo de descubrir patrones, tendencias, relaciones y otros insights que puedan ser utilizados para tomar decisiones informadas.

Los analistas de datos utilizan herramientas estadísticas y de aprendizaje automático para analizar los datos y descubrir insights que ayuden a la organización a tomar decisiones informadas.

Fundamentos de la ciencia de datos

Los fundamentos de la estadística y la probabilidad son esenciales para comprender y analizar datos en ciencia de datos y otros campos. Aquí te presento los conceptos básicos de cada uno:

Estadística

La estadística se ocupa de recopilar, analizar, interpretar, presentar y organizar datos. Se divide en dos grandes ramas:

Estadística Descriptiva: Se centra en describir y resumir conjuntos de datos a través de medidas como:

Medidas de tendencia central: Media, mediana y moda.
Medidas de dispersión: Rango, varianza, desviación estándar e intervalos intercuartílicos.
Gráficos y tablas: Histogramas, diagramas de caja, gráficos de barras, entre otros, para visualizar los datos.

Fundamentos de la ciencia de datos

La media, la mediana y la moda son ***medidas de tendencia central*** en estadística que se utilizan para identificar el valor central o típico dentro de un conjunto de datos:

Media (Promedio): La media es el promedio de un conjunto de números. Se calcula sumando todos los valores y luego dividiendo la suma total por la cantidad de valores. Es útil para obtener una idea general de la tendencia central de los datos, pero puede ser sensible a valores atípicos extremos.

Mediana: La mediana es el valor medio de un conjunto de números cuando están ordenados en secuencia. Si hay un número impar de observaciones, la mediana es el número que está en el medio. Si hay un número par de observaciones, la mediana es el promedio de los dos números del medio. La mediana es menos sensible a valores atípicos y puede dar una mejor idea del valor central en distribuciones muy sesgadas.

Moda: La moda es el valor o valores que aparecen con mayor frecuencia en un conjunto de datos. Un conjunto de datos puede tener una moda (unimodal), dos modas (bimodal) o múltiples modas (multimodal). Es útil para identificar valores comunes en un conjunto de datos, especialmente en datos categóricos.

Fundamentos de la ciencia de datos

Estadística Inferencial

Permite hacer inferencias, predicciones o decisiones basadas en datos de muestras, generalizando para poblaciones más grandes:

Varianza: La varianza mide cuán dispersos están los valores de un conjunto de datos respecto a su media. Se calcula como el promedio de los cuadrados de las diferencias entre cada dato y la media del conjunto.

Desviación Estándar: La desviación estándar es la raíz cuadrada de la varianza. Proporciona una medida de dispersión de los datos en las mismas unidades que los datos originales, lo que facilita su interpretación.

Rango: El rango es la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Mide la extensión total de la variabilidad de los datos. El rango es una medida de dispersión muy básica y no proporciona información sobre cómo se distribuyen los valores entre el mínimo y el máximo.

Fundamentos de la ciencia de datos

Estimación de parámetros

Pruebas de Hipótesis: Las pruebas de hipótesis son un procedimiento estadístico que permite tomar decisiones sobre una población basándose en los datos de una muestra. Se comienza formulando dos hipótesis: la hipótesis nula (H0) y la hipótesis alternativa (1H1 o Ha). La hipótesis nula generalmente representa una afirmación de no efecto o estado de normalidad, mientras que la hipótesis alternativa representa lo que se está tratando de probar. Se calcula un estadístico de prueba y, basándose en este, se decide si rechazar o no la hipótesis nula, teniendo en cuenta un nivel de significancia (α), que es la probabilidad de rechazar la H0 cuando es verdadera (error tipo I).

Fundamentos de la ciencia de datos

Intervalos de Confianza: Un intervalo de confianza es un rango calculado a partir de datos de muestra que se estima con cierta confianza que incluye el valor real de un parámetro poblacional desconocido. El intervalo tiene un nivel de confianza asociado (comúnmente 95% o 99%), que refleja la certeza de que el intervalo contiene el parámetro. Por ejemplo, un intervalo de confianza del 95% para la media poblacional significa que si repetimos el experimento muchas veces, aproximadamente el 95% de los intervalos calculados incluirían la media real de la población.

Análisis de Regresión: El análisis de regresión es una técnica estadística utilizada para examinar la relación entre una variable dependiente y una o más variables independientes. El objetivo es modelar la relación entre estas variables para entender cómo el cambio en las variables independientes afecta la variable dependiente. El tipo más común de regresión es la regresión lineal, donde se ajusta una línea recta a los datos para modelar la relación, aunque hay muchos otros tipos de regresión para relaciones no lineales y complejas.

Fundamentos de la ciencia de datos

ANOVA (Análisis de Varianza): ANOVA es una técnica estadística utilizada para comparar las medias de tres o más grupos para determinar si al menos uno de los grupos difiere significativamente de los otros. Es útil cuando se quieren comparar las medias de grupos basados en una variable categórica independiente. ANOVA descompone la variabilidad observada en los datos en dos partes: variabilidad entre grupos y variabilidad dentro de los grupos, y luego evalúa la significancia de la diferencia entre grupos mediante un estadístico.

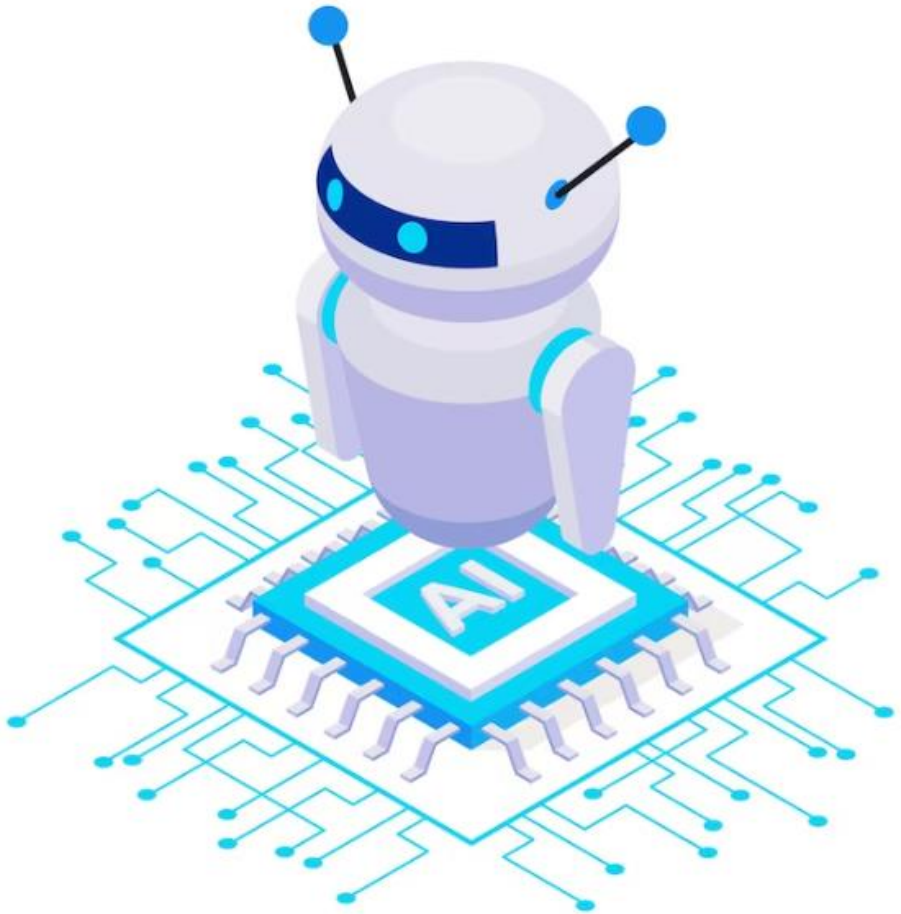
Actualización de hipótesis:

El Teorema de Bayes es un principio fundamental en la teoría de la probabilidad que describe cómo actualizar las probabilidades de las hipótesis a medida que se obtiene más evidencia. Fue formulado por el matemático inglés Thomas Bayes y publicado póstumamente. El teorema proporciona una manera de calcular la probabilidad de una hipótesis dado un conjunto de evidencia, utilizando el conocimiento previo o las creencias sobre esa hipótesis antes de obtener la evidencia.

Fundamentos de la ciencia de datos

Aprendizaje automático: El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos que permiten a las máquinas aprender de los datos y mejorar su rendimiento en tareas específicas.

Los algoritmos de aprendizaje automático son utilizados en una amplia variedad de aplicaciones, desde sistemas de recomendación en línea hasta el análisis de datos de sensores en la industria.



Fundamentos de la ciencia de datos

1. Adquisición de datos

El primer paso en el proceso de la ciencia de datos, es la adquisición de datos, es necesario determinar que datos se encuentran disponibles.

Cuando se trata de encontrar una base de datos para nuestro problema, es necesario buscar información en diferentes tipos lugares, y hacer uso de los datos relevantes para el objetivo y su análisis.

No es recomendable dejar de lado datos sin una revisión previo de la información, aun que representen una pequeña cantidad, ya que puede llevar a conclusiones incorrectas.

Fundamentos de la ciencia de datos

La herramienta de elección mas utilizada para trabajar con datos estructurados es SQL, o los almacenes de datos NoSQL que proporcionan una API que permite a los usuarios acceder a los datos. Adicionalmente la mayoría de los sistemas de bases de datos vienen con un entorno de aplicación grafica.

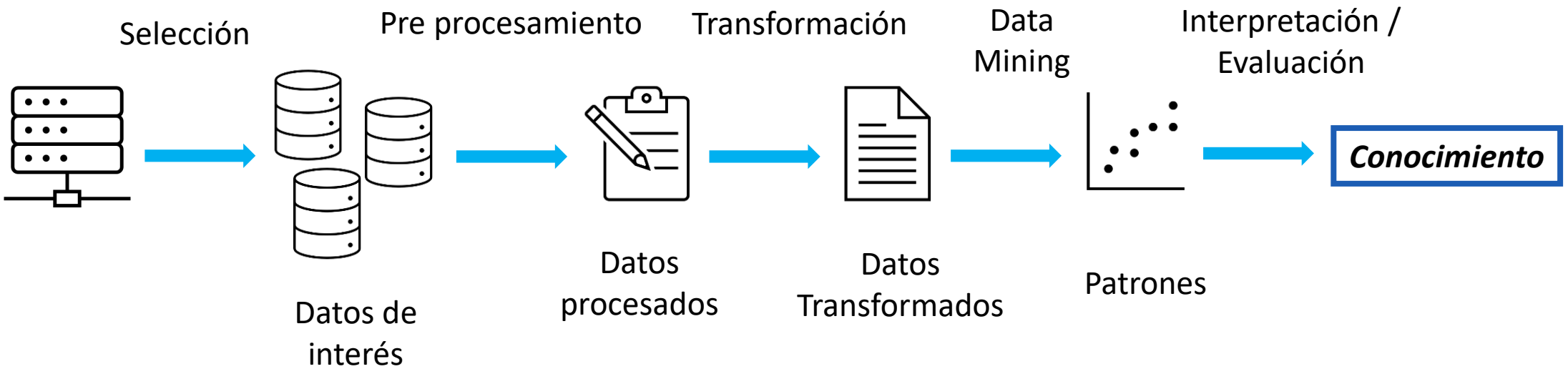
Existe una gran cantidad de datos que pueden ser obtenida a través e organizaciones, desde el centro de almacenamiento de la empresa, investigaciones, sitios web, etc.

Fundamentos de la ciencia de datos

2. Manipulación ordenamiento de los datos

Proceso de extracción del conocimiento.

Con el fin de iniciar este proceso, es necesario un análisis exploratorio que puede ser implementado de forma automática, de bases de datos de gran tamaño. Con el fin de obtener datos, organizar datos, identificar patrones validos, potencialmente útiles e inteligibles.



Fundamentos de la ciencia de datos

Paso a: Comprender las bases del problema y establecer los objetivos.

Es fundamental tener claras las características, los limites y los objetivos que tiene el problema original, para reunir toda la información importante y útil para el análisis.

Paso b: Crear un set de datos objetivo

Una vez establecido el problema y el objetivo del análisis, se debe determinar cual es la variable de decisión. Debido a que los datos se encuentran acumulados en una base de datos, documentos, imágenes, archivos, etc. Es necesario ordenar y homogenizar los formatos para lograr su procesamiento y su análisis.

Fundamentos de la ciencia de datos

Paso c: Pre procesamiento y limpieza de datos.

Preprocesamiento

- Eliminar datos duplicados, e inconsistencias
- Eliminar el ruido y datos aislados
- Adicionar información faltante
- Datos incompletos, atributos que carecen de valores, sin relación, sin interés.
- Con ruido, contienen errores
- Discrepancias en códigos o nombres
- Grandes volúmenes de filas y columnas.

El objetivo es mejorar la calidad de los datos antes de aplicar los modelos y algoritmos, y mejorar los resultados que entregara la minería de datos.

Fundamentos de la ciencia de datos

Limpieza de datos.

- Corregir datos erróneos
- Filtrar datos incorrectos
- Reducir el alto nivel de detalle
- Detección y resolución de discrepancias.

La calidad de datos se consigue cuando se cumple:

- **Integridad:** Cumple requisitos de entereza y validez
- **Consistencia:** Corrección de contradicciones
- **Uniformidad:** No debe presentar irregularidades
- **Densidad:** Valores omitidos sobre el numero de valores totales
- **Unicidad:** No presentar valores duplicados ni registros inconsistentes.

Fundamentos de la ciencia de datos

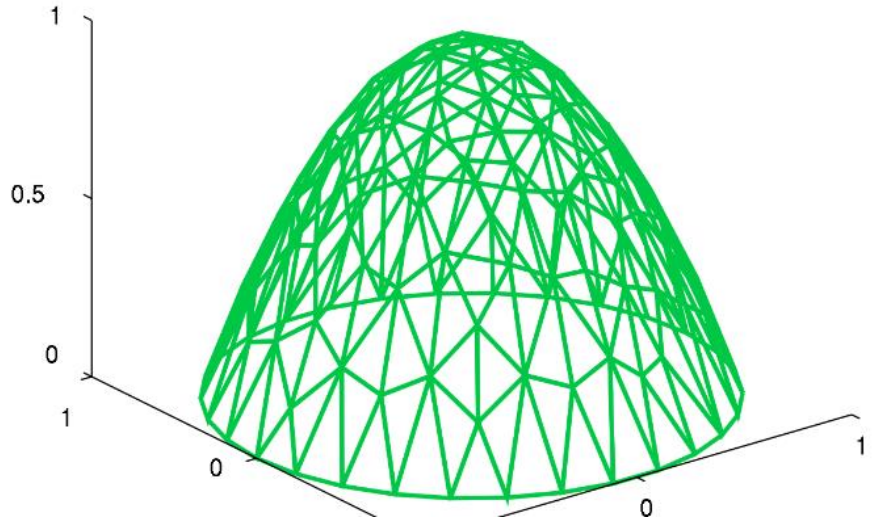
Transformación de datos

Discretización: es el proceso por el cual, se convierten variables continuas en variables categóricas.

Ejemplo practico: Puedo tener acceso a un dato relacionado con la velocidad que alcanza un vehículo en carretera, sin embargo, mi objetivo es comprar su desempeño con otros modelos de vehículos, por lo tanto podemos generar un rango desde la velocidad mínima a la máxima alcanzada y con esto comparar los resultados.

Con esto estamos discretizando los datos, y así los convertimos en variables categóricas.

Esto es necesario especialmente, cuando se trabajan con algoritmos que solo permiten la entrada de este tipo de variables.



Fundamentos de la ciencia de datos

Normalización de datos

La normalización busca conseguir que las variables estén expresadas en una escala similar, para que de esta forma todas tengan un peso comparable y podamos realizar un análisis estadístico mas coherente.

El escalado implica cambiar el rango de valores para que entre un valor especifico como, por ejemplo rangos de 0 a 1. Esto se hace para evitar que ciertas entidades, con valores grandes dominen los resultados.

Una formulación que permite escalar los datos en un rango que permita su análisis, es la normalización 0 – 1:

$$Z_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Data Values	Normalized
13	0
16	0.0517
19	0.1034
22	0.1552
23	0.1724
38	0.4310
47	0.5862
56	0.7414
58	0.7759
63	0.8621
65	0.8966
70	0.9828
71	1

Fundamentos de la ciencia de datos

Tratamiento de datos faltantes

Un valor faltante, perdido o desconocido, es un atributo que no esta almacenado. En la librería Pandas de Python, se presenta como **None** o **Nan**.

Como organizar la base de datos cuando se tienen este tipo de datos:

- Eliminación de filas que tiene en su mayoría datos faltantes.
- Eliminar columnas, cuando el atributo no representa valores para la mayoría de las observaciones.
- Imputar su valor, en situaciones intermedias, llenar el valor según medidas de resumen (media, moda, mdiana, etc.). No es recomendado debido a que puede presentar problemas.

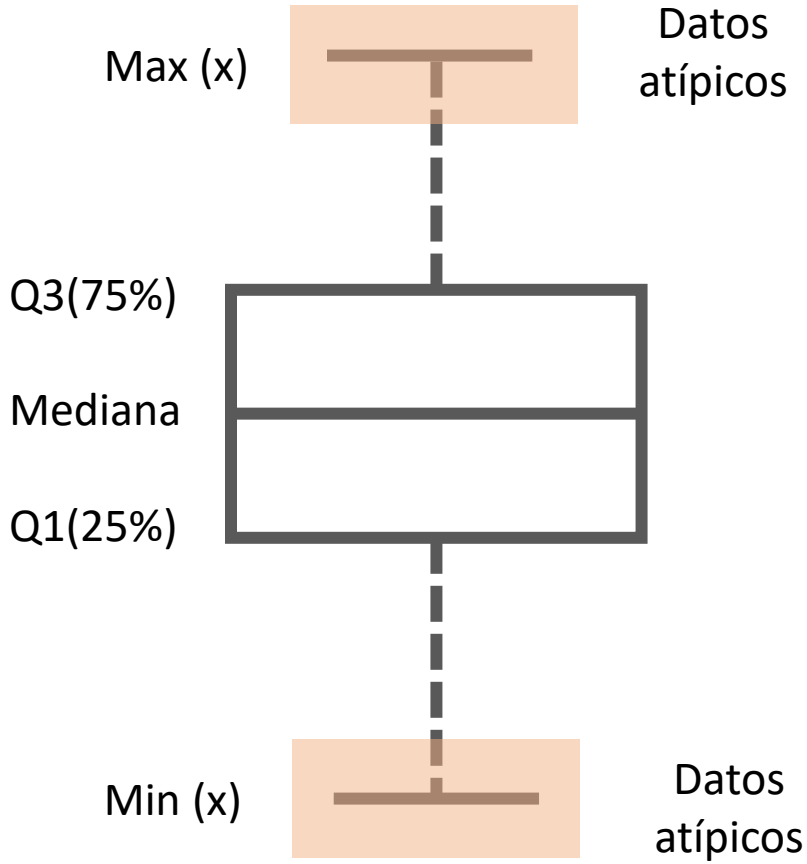
	y	x
1	0.047	NA
2	-1.806	Inf
3	0.045	NaN
4	0.377	-0.295
5	-0.516	0.853
6	-0.631	1.191

Valores fuera de rango

Otra definición, outlier, es un valor atípico en un conjunto de datos que pueden ser inconsistentes con el comportamiento general de la base de datos. Es necesario detectar los valores atípicos y determinar que impacto tienen en el análisis estadístico del problema.

Existen herramientas que permiten visualizar este tipo de información:

Como posibles acciones es posible, ignorar los datos, filtrar o reemplazar los datos, Sesgar los datos, Reemplazar el valor o discretizar.



Fundamentos de la ciencia de datos

Paso d: Minería de datos

Con el uso de los algoritmos, es posible extraer nueva información de nuestro set de datos. Pasos de la minería de datos:

- Selección la tarea
- Selección del algoritmo

Paso e: Interpretación de patrones

En esta etapa se interpretan los resultados obtenidos; así como la exposición de manera clara de éstos para que sean comprensibles. Se recomienda el uso de técnicas de visualización.

3. Aplicación de técnicas de análisis

Machine Learning

Una vez nuestros datos están organizados y hemos tomado las decisiones necesarias para limpiar los valores innecesarios o que presenten inconsistencias, es necesario apoyarnos del Machine Learning. Representa el aprendizaje automático que logra que los modelos implementados funcionen y realicen predicciones en función del objetivo del problema.

Fundamentos de la ciencia de datos

Dependiendo de la información disponible existen dos modelos disponibles:

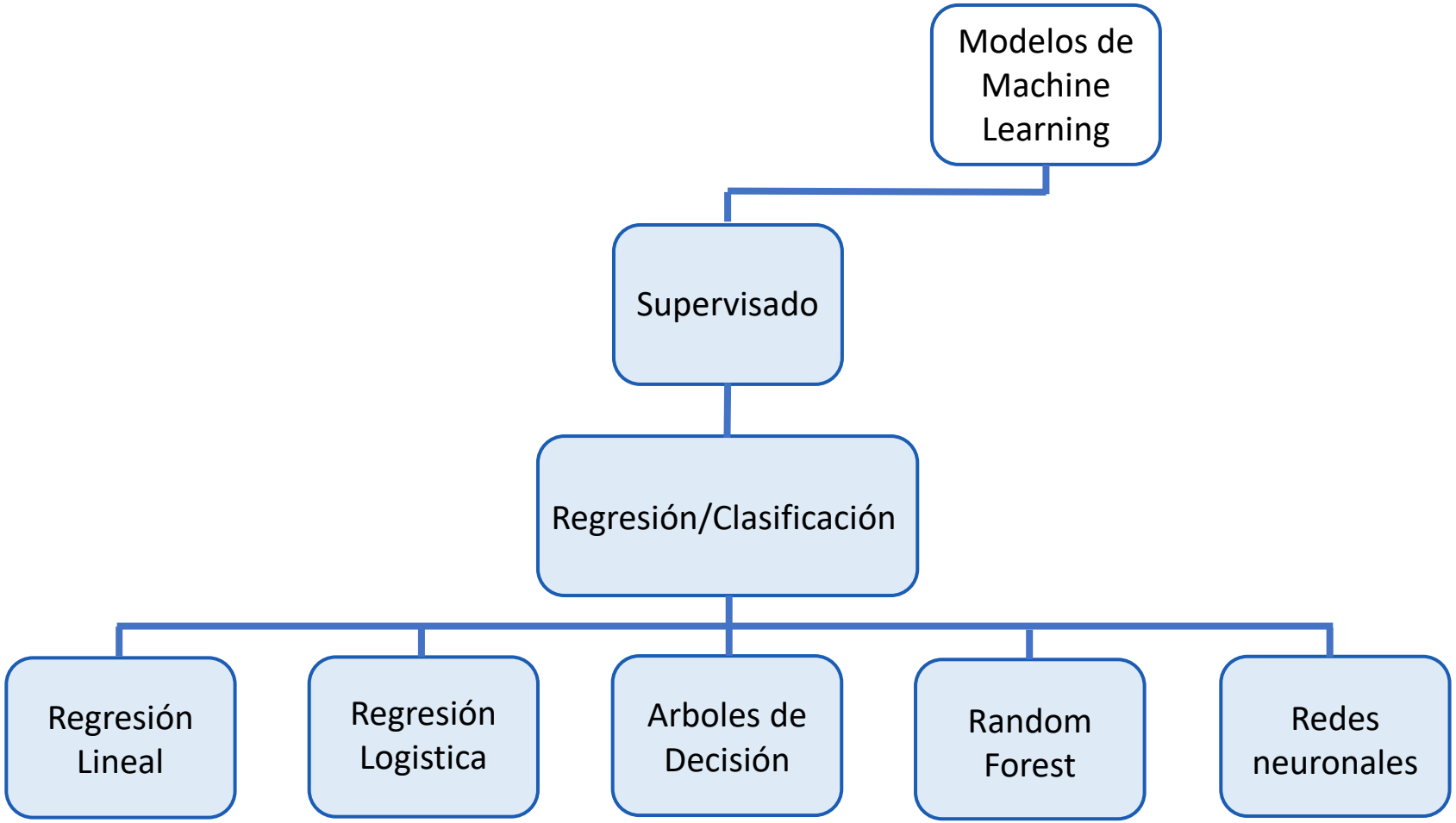
Aprendizaje supervisado

Requiere de una variable objetivo, la variable objetivo es el resultado de interés para nuestro problema, por ejemplo cuanto es el valor predicho del costo de un vehículo dentro de 1 mes,. El objetivo es predecir la respuesta futura o comprender la mejor relación entre variable objetivo y las variables predictivas resultado de inferencias.

El aprendizaje supervisado busca patrones en datos históricos relacionando los campos con un objetivo especifico. Existen dos tipos de campos dentro de este tipo de aprendizaje: Predicción de un valor y predicción de categoría:.

- **Regresión:** trata de predecir un número.
- **Clasificación:** trata de predecir una categoría

Fundamentos de la ciencia de datos



Aprendizaje no supervisado

Para cada observación, tenemos un vector de medidas que no lleva asociada una respuesta. No tenemos una variable que predecir, ni que pueda supervisar nuestro análisis.

El objetivo es entender los datos y organizarlos en grupos.