

TALLER ANALITICA

Tito Pablo Neira Ávila
Doctor en gestión de la innovación tecnológica
Co-founder Neicon Consulting Group
Investor Sellit9

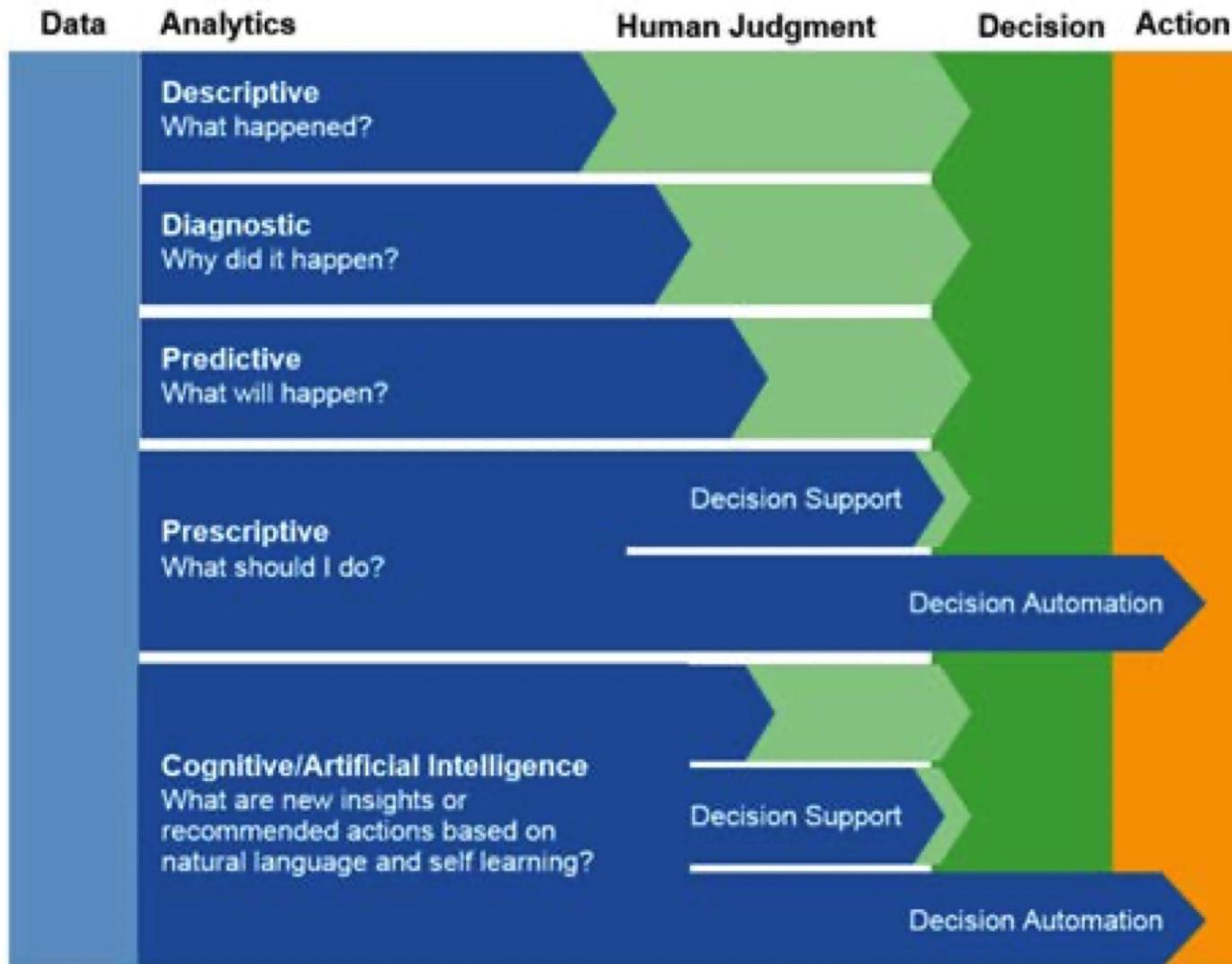


Figure 1. Types of analytics techniques (Gartner, 2017)

Problemas mal estructurados

- El problema por si mismo está mal definido, sin poder concordar con su definición por las partes involucradas.
- La situación considera diferentes involucrados con perspectivas diferentes acerca del problema.
- Hay mucha incertidumbre involucrada, y pocos o ningún dato, muchas veces no confiables.
- El éxito se mide en función de los acuerdos que se generen entre las partes involucradas.
- Este proceso se basa más en acuerdos y aprendizaje, que en soluciones técnicas.

Conjunto de datos y bases de datos.

- Un conjunto de datos es simplemente una colección de datos.
- Una base de datos es una colección de archivos relacionados que contienen registros personas, lugares o cosas. Las personas, lugares o cosas para las que almacenamos y mantenemos información se llaman entidades.
- La base de datos generalmente se organiza en una tabla bidimensional, donde las columnas corresponden a cada elemento individual de datos (llamados campos o atributos), y las filas representar registros de elementos de datos relacionados.
- Una característica clave de las bases de datos computarizadas es la capacidad para relacionar rápidamente un conjunto de archivos con otro.

Big Data

- Hoy en día, casi todos los datos se capturan digitalmente. Como resultado, los datos han ido creciendo a velocidad abrumadora, medida en terabytes (10^{12} bytes), petabytes (10^{15} bytes), exabytes (10^{18} bytes), e incluso en términos de dimensiones superiores.
- El término se utiliza para referirse a cantidades masivas de datos comerciales de una amplia variedad de fuentes, muchas de las cuales están disponibles en tiempo real, y muchos de los cuales son inciertos o impredecibles.
- Big Data tiene las siguientes características volumen, variedad, velocidad y veracidad.
- Muy a menudo, los macrodatos giran en torno al comportamiento y experiencias del cliente.

All of the data created in 2018 is equal to...

0001001000010010
10100110111010011101
0001001000010010
111001001111001001
0001001000010010

33 zettabytes Total data created worldwide in 2018



660 billion

standard Blu-ray discs (50 gigabytes)



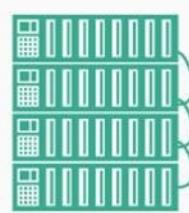
330 million

of the world's largest hard drive (currently 100 terabytes*)



33 million

human brains (1 petabyte**)



132,000

of the fastest supercomputer's storage space (250 petabytes***)



73 grams

of DNA (455 exabytes)

Size guide

x 1,000

x 1,000

x 1,000

x 1,000

x 1,000

Megabyte

Gigabyte

Terabyte

Petabyte

Exabyte

Zettabyte

* As at March 2019.

** Varies depending on calculation method.

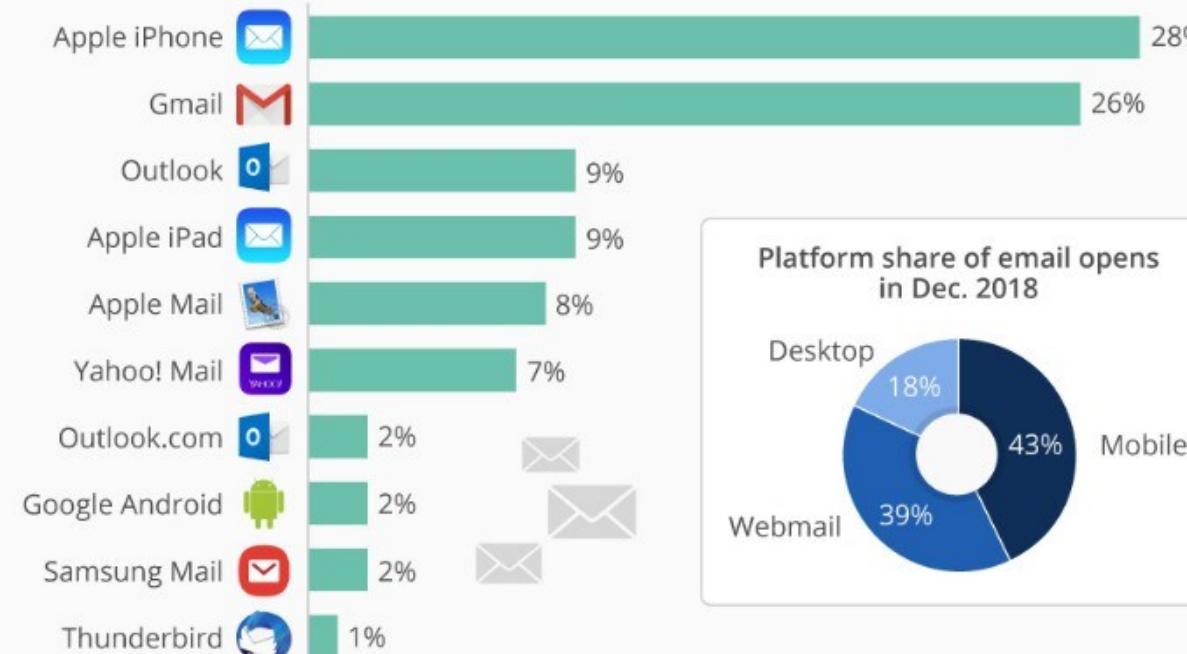
*** As at June 2018.

¿Qué pasa en Internet en un minuto?

2019 *This Is What Happens In An Internet Minute*

The World's Most Popular Email Clients

Top 10 email clients worldwide based on the share of email opens in March 2019*



* percentages based on 834 million email opens worldwide between March 1 and April 1, 2019

Source: Litmus Email Analytics

statista

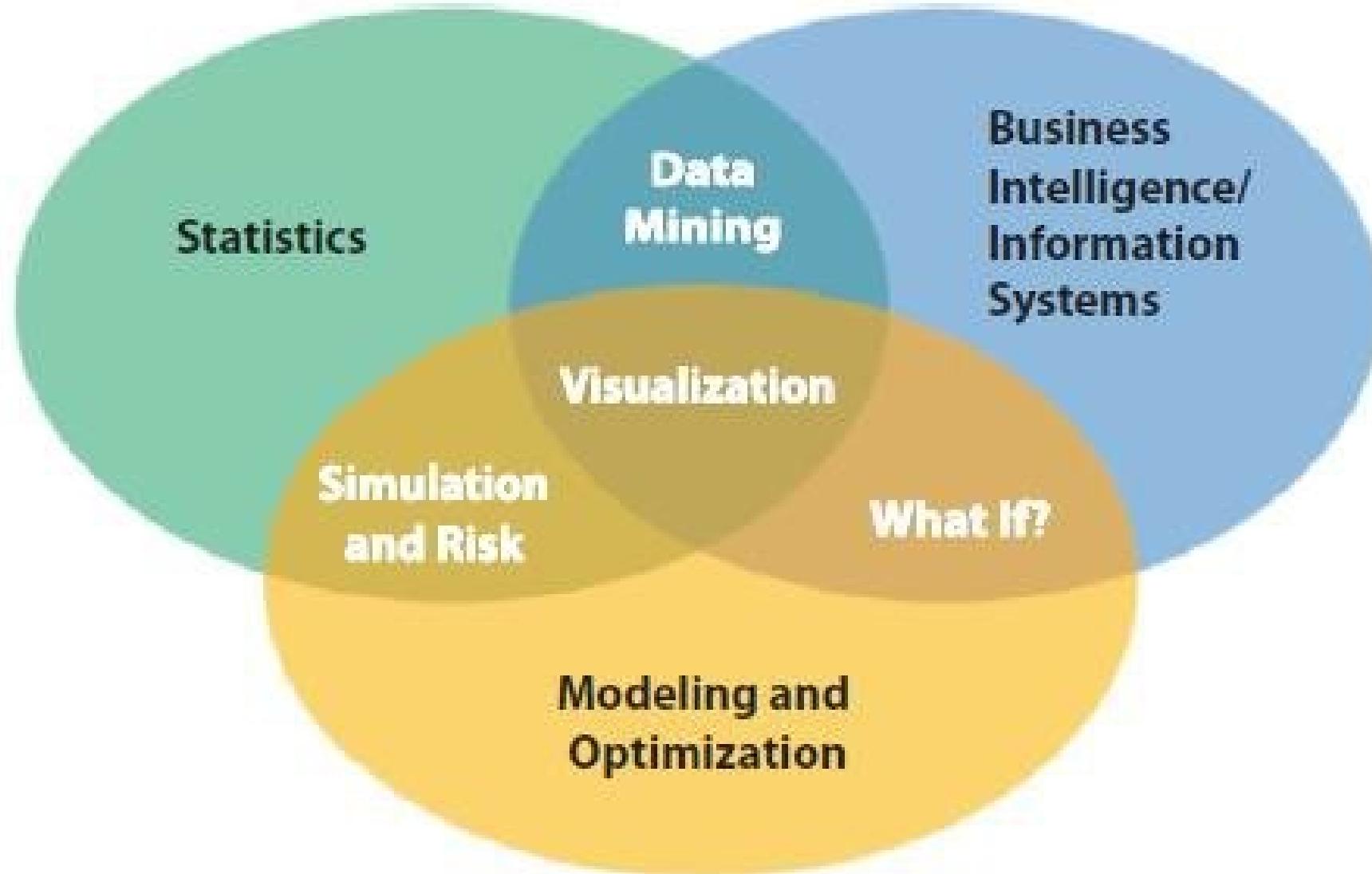


¿Qué es Analítica de los Negocios (Business Analytics)?

- De acuerdo con Evans (2017), es el uso de datos, tecnología de la información, análisis estadístico, métodos cuantitativos y modelos matemáticos o informáticos para ayudar a la administración a obtener una mejor comprensión de sus operaciones comerciales y tomar mejores decisiones basadas en hechos.
- Liberatore y Luo en su obra “The Analytics Movement” afirman que como término general, la analítica se refiere a la ciencia del análisis lógico. Como tal, se relaciona con el trabajo de muchas profesiones y disciplinas académicas.
- “Un proceso de transformación de datos en acciones mediante análisis y conocimientos en el contexto de la toma de decisiones organizacionales y la resolución de problemas ”.

Evans, James. (2017). Business Analytics, 2ed. Pearson Global Edition.

Matthew J. Liberatore, Wenhong Luo, (2010) The Analytics Movement: Implications for Operations Research. *Interfaces* 40(4):313-324.

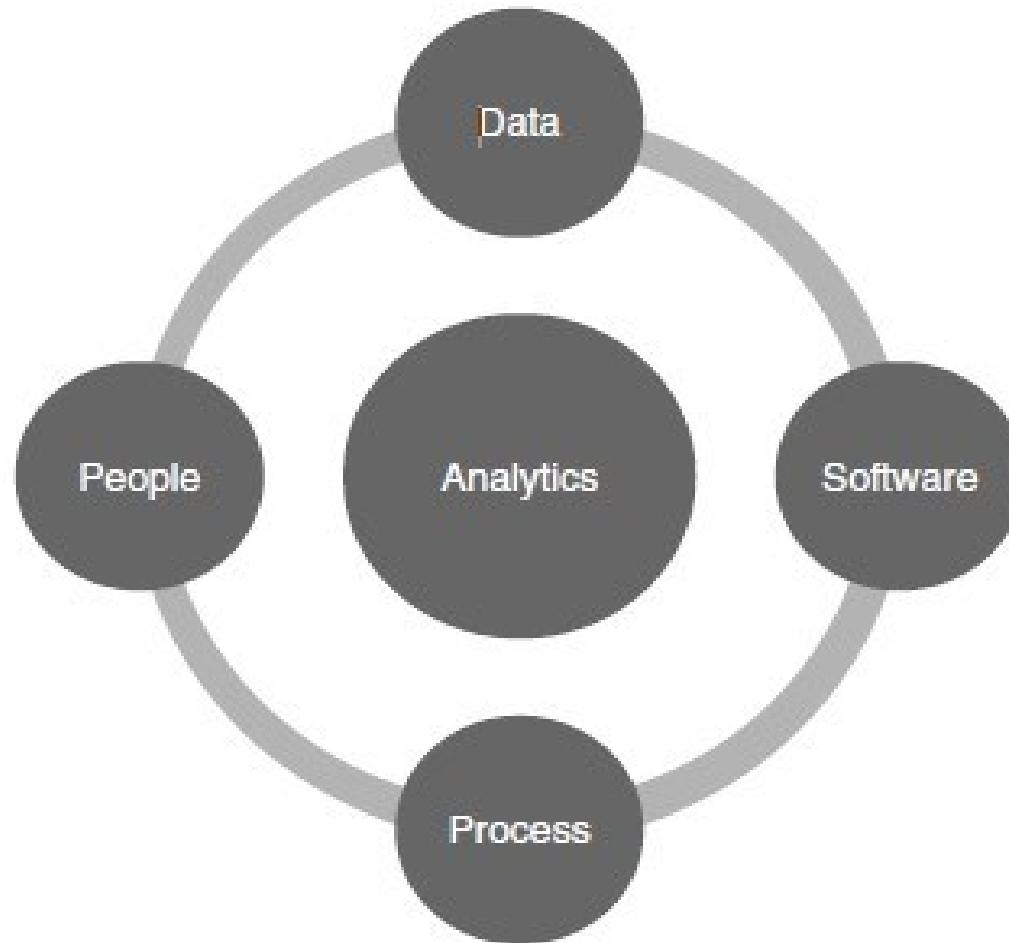


Perspectiva de la Analítica de los Negocios

Herramientas

- Puede verse como una integración de Inteligencia de Negocios, Sistemas de Información, Estadística y Modelado y Optimización como conocimientos y herramientas básicas y tradicionales.
- Por otro lado, la intercepción de estos conocimientos genera otros no tradicionales como:
 - Minería de datos. Que ayuda a entender las características y patrones de comportamiento de bases de datos grandes utilizando una variedad de herramientas estadísticas.
 - Análisis de Riesgo y Simulación: ayuda a determinar el impacto de la incertidumbre y su impacto potencial en la comportamiento de las diferentes variables bajo estudio.
 - Análisis “What-If (Y-Si): permite estudiar como una combinación específica de variables afecta los resultados del modelo estudiado.
 - Visualización: hace única la Analítica de Negocios, ya que provee una manera sencilla de comunicar y revelar comportamientos de datos.

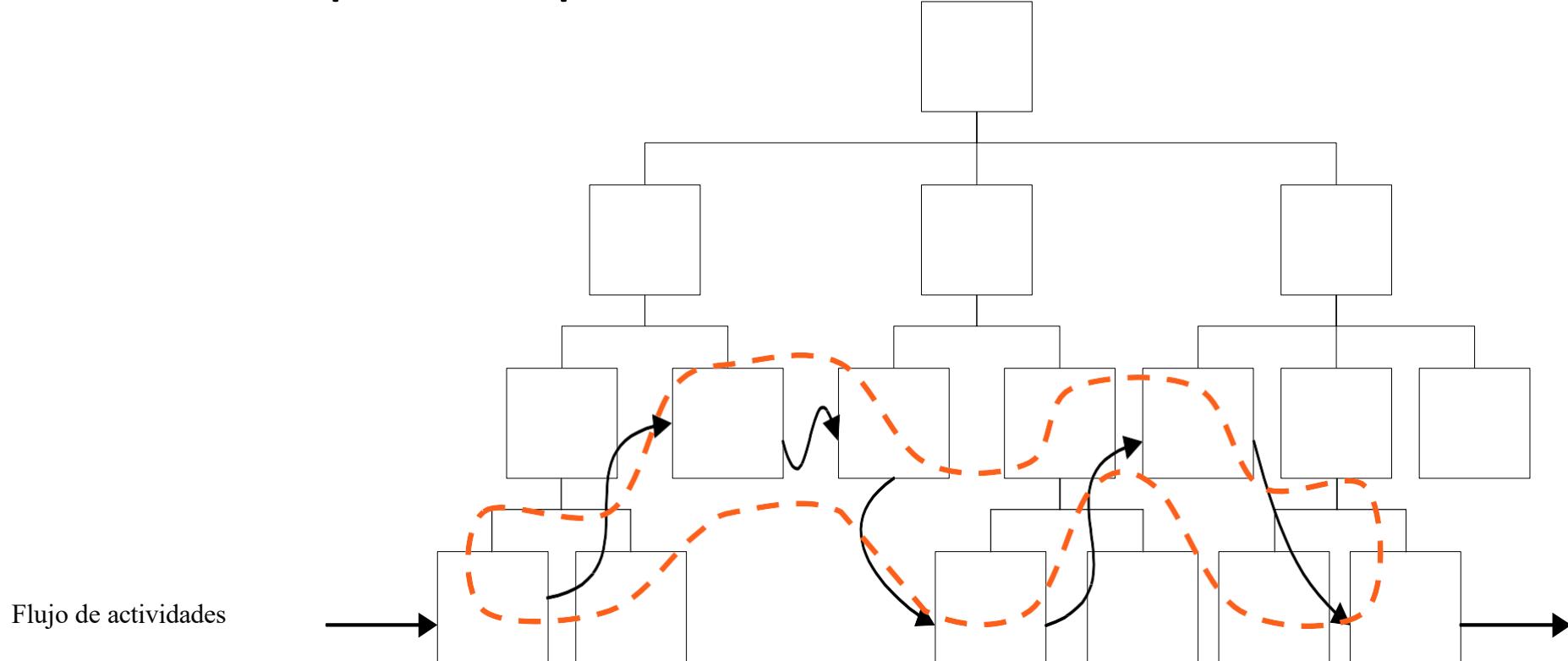
Impulsores de la Analítica



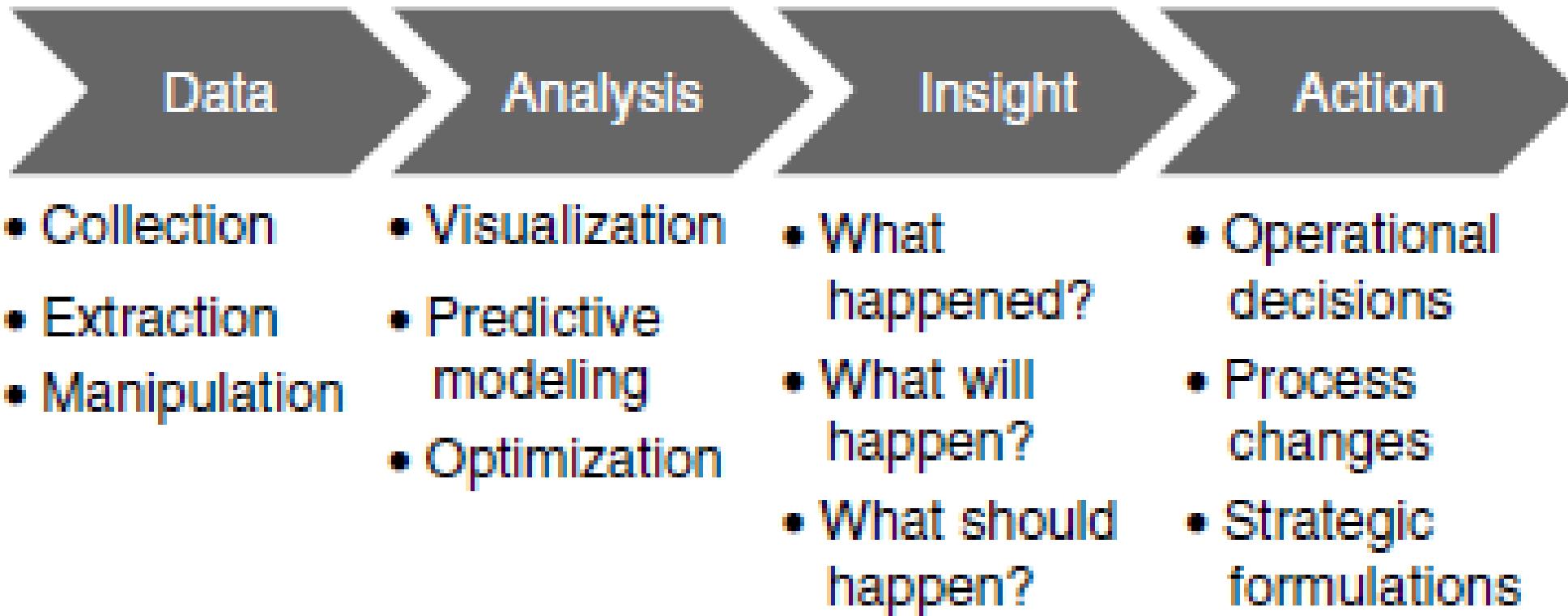
El enfoque de procesos

- Procesos:
 - Una colección de modelos identificados por el tipo de decisión y por una secuencia de tareas.
 - Esas tareas son las unidades mínimas identificables de análisis
 - Su arreglo óptimo es la variable de diseño crítica al determinar la eficiencia del enfoque seleccionado

El enfoque de procesos

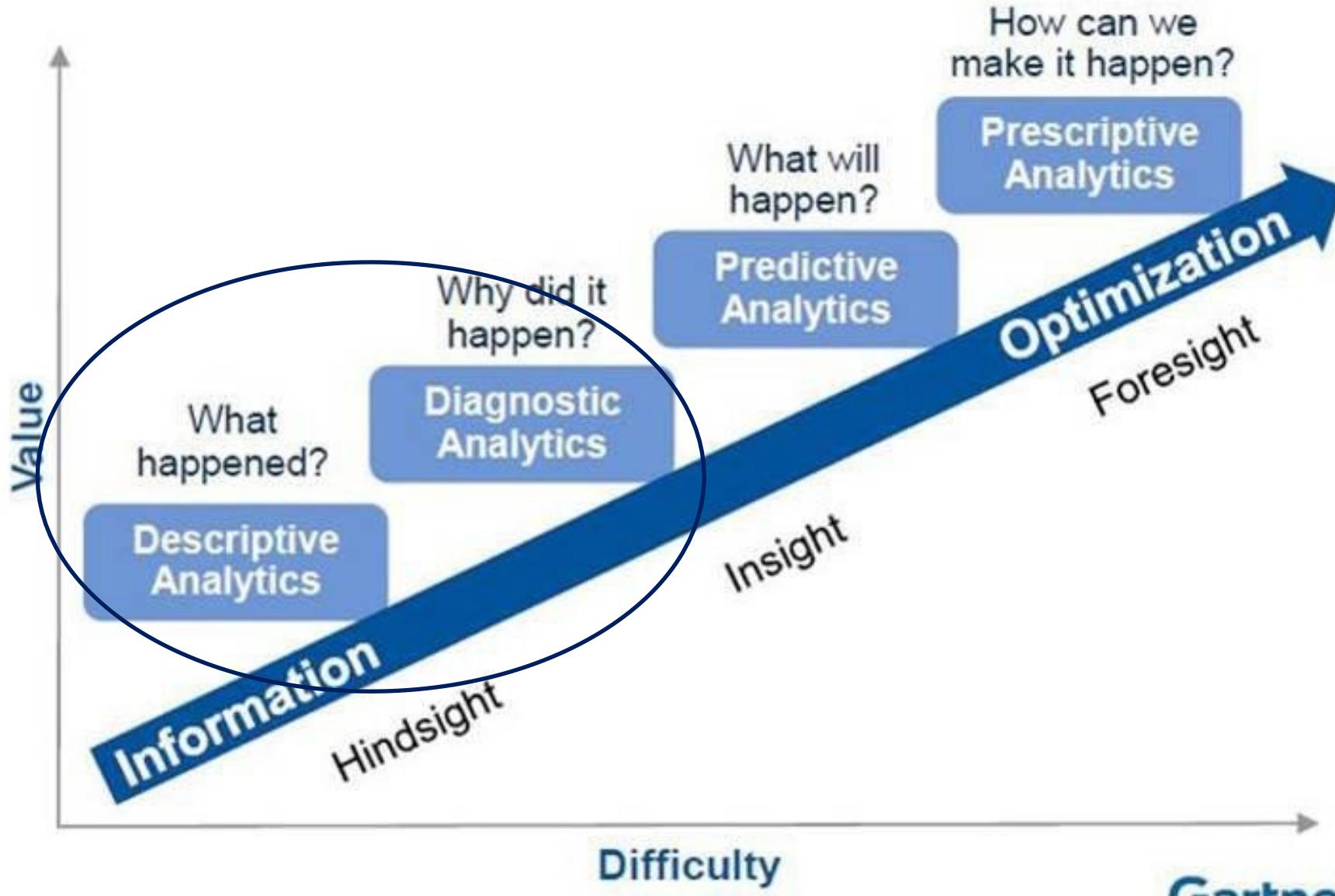


Enfoque de procesos de la Analítica de Negocios



Enfoques

- **Analítica descriptiva:** es el tipo de análisis más antiguo y más utilizado por las empresas. En los negocios, este tipo de análisis a menudo se conoce como inteligencia comercial, ya que proporciona el conocimiento necesario para hacer predicciones futuras, similar a lo que hacen las agencias de inteligencia para los gobiernos. Esta categoría de análisis incluye el análisis de datos del pasado mediante técnicas de agregación y minería de datos para determinar qué ha sucedido hasta ahora, que luego se puede utilizar para determinar qué es probable que suceda en el futuro.
- **Analítica predictiva:** es el arte de obtener información de los datos recopilados y utilizarla para predecir patrones y tendencias de comportamiento. Con la ayuda del análisis predictivo, puede predecir factores desconocidos, no solo en el futuro, sino también en el presente y el pasado.
- **Analítica prescriptiva:** es la tercera rama de las tres grandes ramas de datos. Este sistema de análisis es una suma de los dos anteriores, a saber, análisis descriptivo y análisis predictivo. Utiliza algoritmos de optimización y simulación para determinar opciones para el futuro. Responde a la pregunta: "¿Qué debemos hacer?"



Gartner

En conclusión

- ¿Qué define el éxito de una empresa? ¿Es el número de personas empleadas por la empresa? ¿Es el volumen de ventas de la empresa?
- ¿Es la fuerza de la base de clientes de la empresa? ¿La satisfacción de los empleados influye en el éxito de las operaciones comerciales?
- ¿Cómo influye la gestión en el éxito operativo general?
- Para contestar estas preguntas se necesita data e información, no simplemente historias o anécdotas. Los datos son necesarios para la supervivencia del negocio. Pero recolectar data no es suficiente, hay que procesarla para convertirla en información, que a su vez, después de analizada se convierte en conocimiento necesario para la organización y la toma de decisiones razonada.

PROCESO DE PLANEACIÓN ESTRATÉGICA



Especificación de las hipótesis

Es una afirmación o proposición no probada sobre un fenómeno, el comportamiento de una o mas variables, o la relación entre dos o mas variables

Los resultados de la investigación aceptan o rechazan las hipótesis

Anticipan las respuestas posibles a las cuestiones planteadas por la investigación

Manifiestan lo que se está buscando

Pueden derivarse de investigaciones exploratorias o afirmaciones

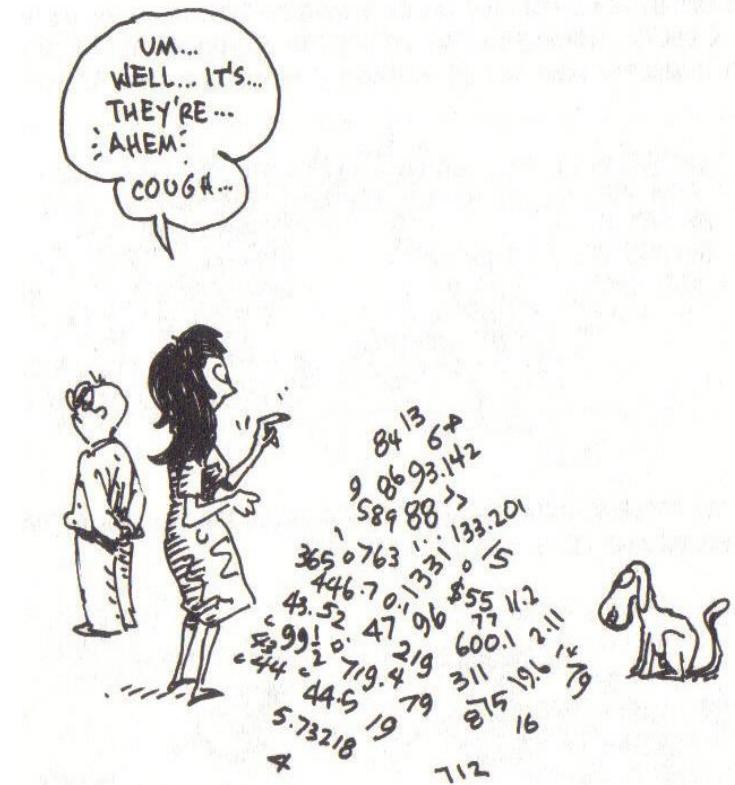


La estadística es una ciencia que facilita la solución de problemas en los cuales necesitamos conocer y relacionar características sobre el comportamiento de algún suceso económico, político, social, biológico o físico .



Los datos se editan y codifican para dar como resultado la información que responda las preguntas de la investigación.

Una forma de análisis consiste en resumir grandes cantidades de datos a fin de poder interpretar los resultados.



Definición, clasificación y medida de las variables de estudio

Variable:

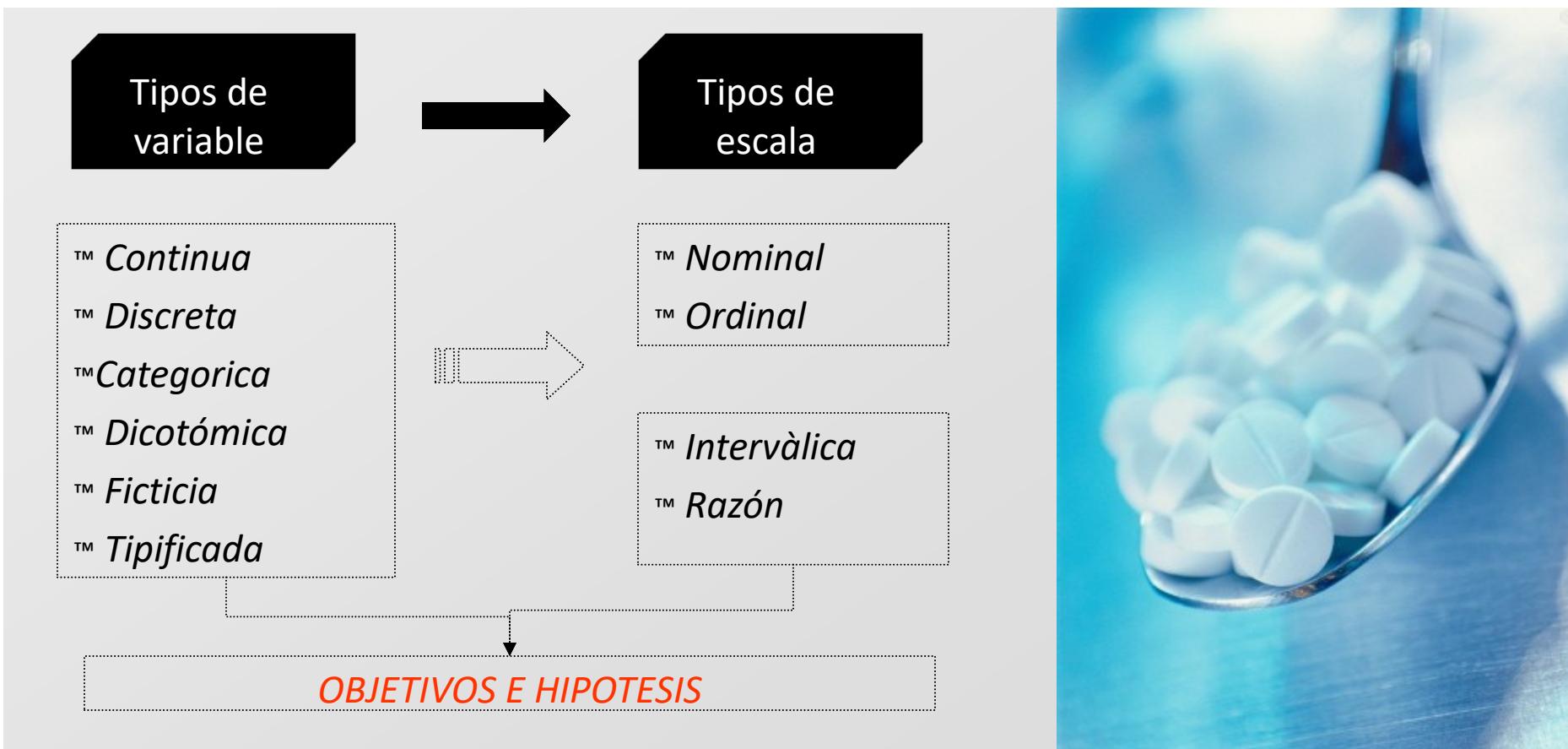
- *Es una magnitud cuyos valores son objeto de estudio*

- *En un estudio social puede referirse a un individuo, grupo de personas y/o organizaciones.*



- ™ *Comportamientos*
- ™ *Atributos*
- ™ *Actitudes y opiniones*
- ™ *Motivaciones necesidades*





Técnicas de Análisis de Datos

™*descriptiva*
↓
™*Estudian el comportamiento de una sola variable*

™*Inferencial*
↓
™*Estudian la relación entre (de asociación o dependencia) entre dos variables*

™*Multivariada*
↓
™*Analiza la relación simultánea entre el comportamiento de más de dos variables*

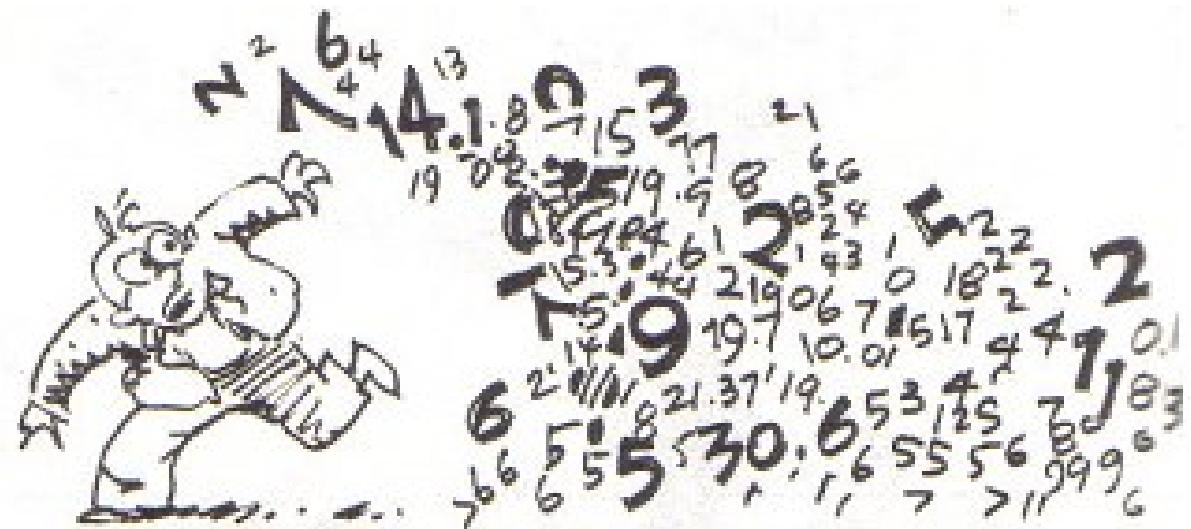
ANALISIS INTEGRAL

Análisis de los datos

→ ™*Una dependiente múltiples independientes*
→ ™*Múltiples dependientes múltiples independientes*

- Análisis Descriptivo:
 - Univariado:
 - Número de casos, suma, máximo, mínimo.
 - Tendencia central: Media, Mediana y Moda.
 - Dispersión: Rango, Varianza y Desviación estándar.
 - Bivariado:
 - Tabulaciones cruzadas: Análisis de Tablas de contingencia.
 - Multivariado

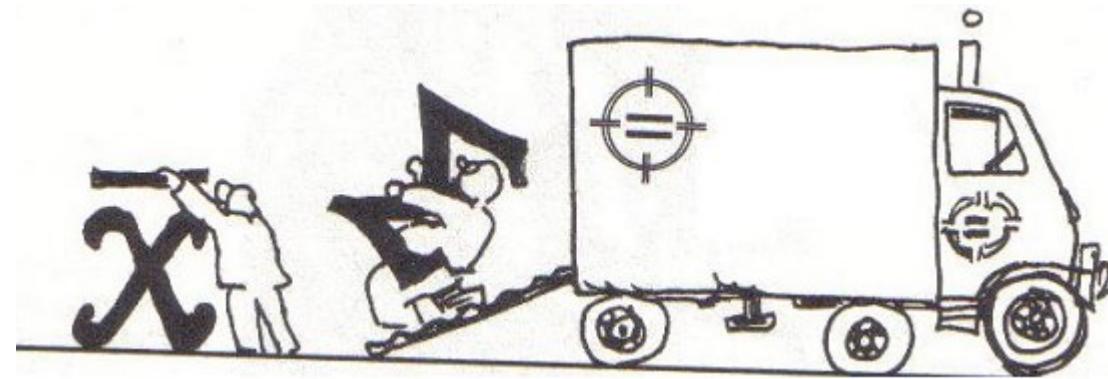
descriptiva



La estadística descriptiva formula reglas y procedimientos para la presentación de una masa de datos en una forma más útil y significativa, además establece normas para la representación gráfica de los datos.

Existen dos medidas de interés para cualquier conjunto de datos: las de tendencia central y las de variabilidad.

Medidas de tendencia central



La tendencia central de un conjunto de datos es la disposición de estos para agruparse ya sea alrededor del centro o de ciertos valores numéricos.

Supongamos que estamos tomando una serie de n observaciones... entonces escribimos

$x_1, x_2, x_3, \dots, x_n$



Usando la letra griega

$$\Sigma$$

(sigma mayúscula)

representamos la sumatoria de las observaciones así:

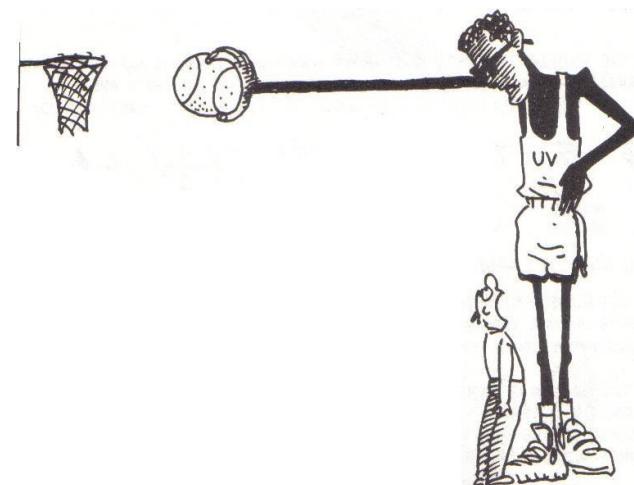
$$\sum_{i=1}^n x_i$$

- **Media aritmética**

Es la suma de todos los valores de una variable dividida entre el número total de observaciones; ella mide el valor alrededor del cual tienden a agruparse los datos.

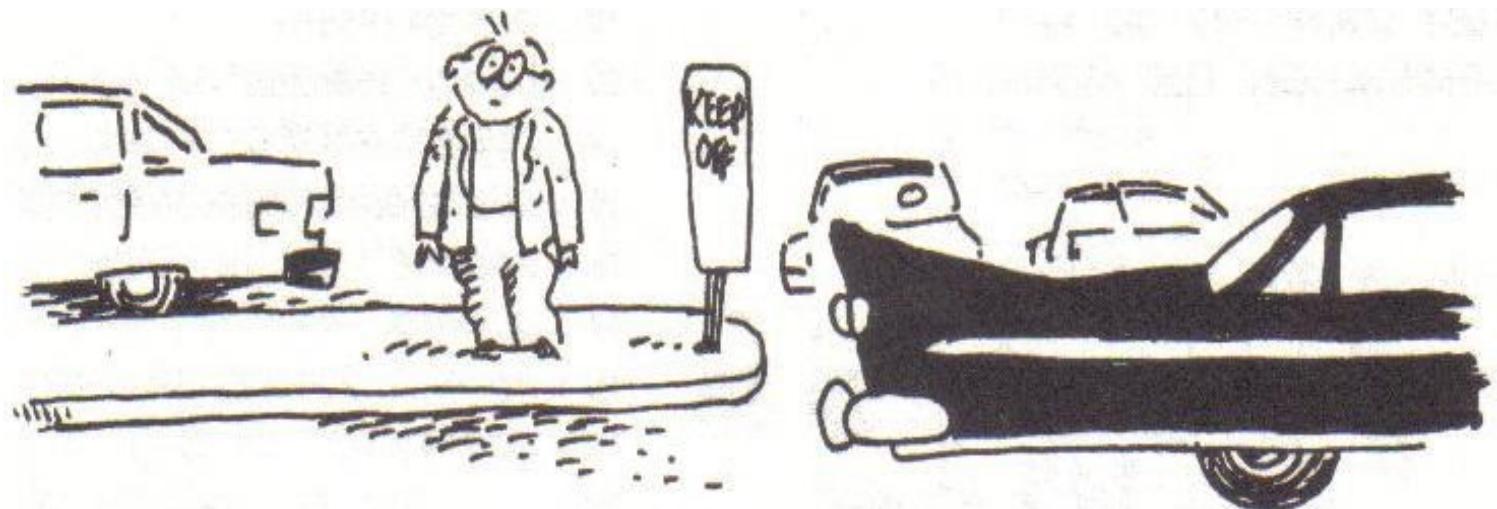
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad o \quad \sum_{i=1}^n \frac{x_i}{n}$$

Es apropiado usar la media cuando los resultados son simétricos, pero en otros casos cuando hay observaciones extremas puede ser muy engañoso.



- **Mediana**

Es otra clase de centro, es el punto medio de los datos, como la línea media que divide una carretera



3 5 7 7 38
↑
la mediana



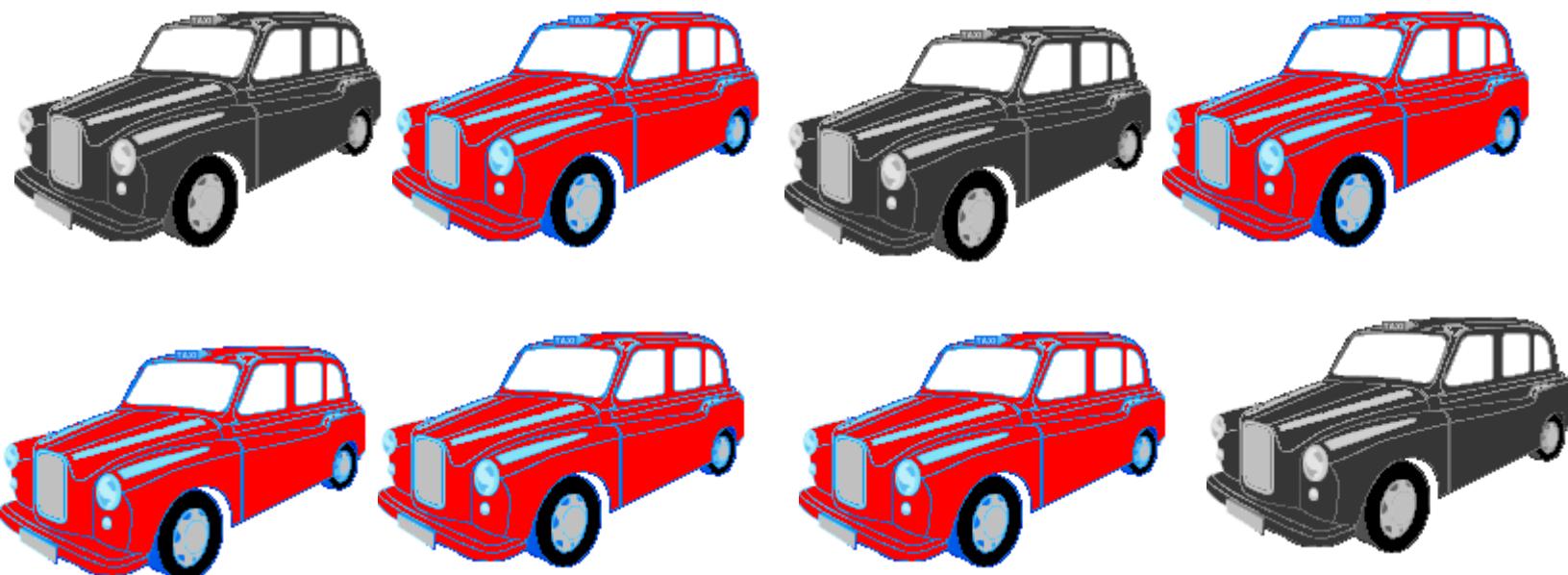
$$3 \quad 5 \quad 7 \quad 7$$

entonces
espacio en el medio

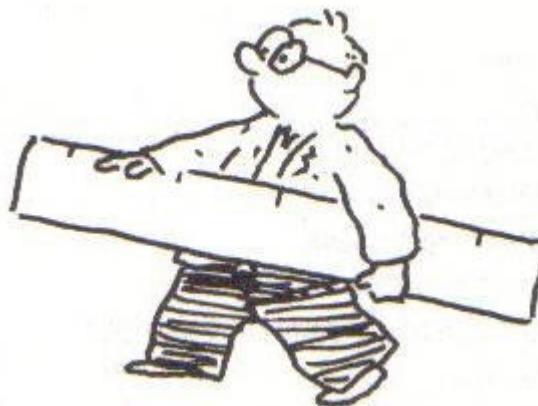
$$\frac{5 + 7}{2} = 6$$

- *La moda*

Es otra medida de tendencia central no tan usual como las anteriores; siendo esta el valor de la variable que presenta mayor repetición. Es decir la respuesta más común.



Medidas de variabilidad



La variabilidad de un conjunto de datos es la dispersión de las observaciones en el conjunto

- **Varianza**

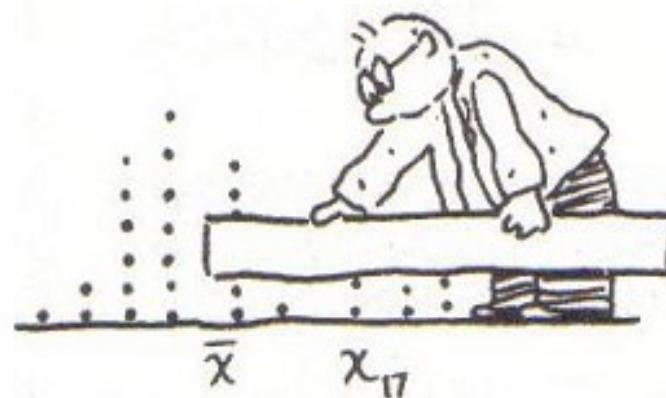
Es la distancia al cuadrado de las observaciones a la media, sobre el número de observaciones

distancia promedio al cuadrado

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

*Por razones técnicas
usamos $n-1$ en el
denominador, y se
define como varianza*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$



- *Desviación estándar*

Una medida de dispersión debería tener las mismas unidades que los datos originales, entonces lo mas indicado es sacar la raíz cuadrada de la varianza.



$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

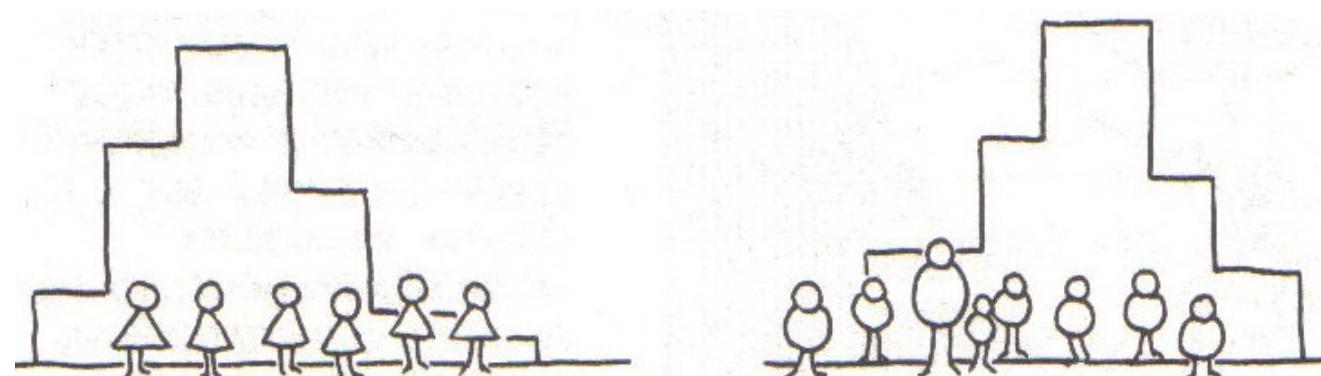
- **Coeficiente de variación**

Mide la dispersión de los datos, se calcula

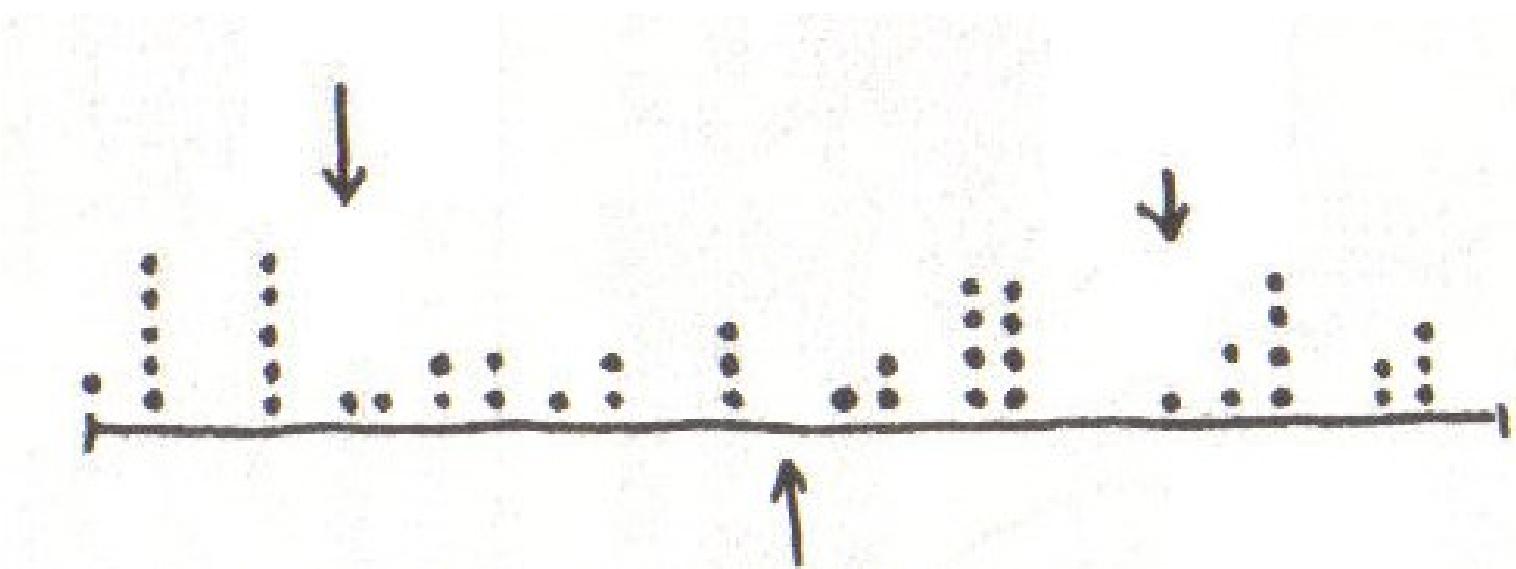
desviación estándar

$$\textbf{CV} = \frac{\text{Media aritmética}}{\text{desviación estándar}} \times 100$$

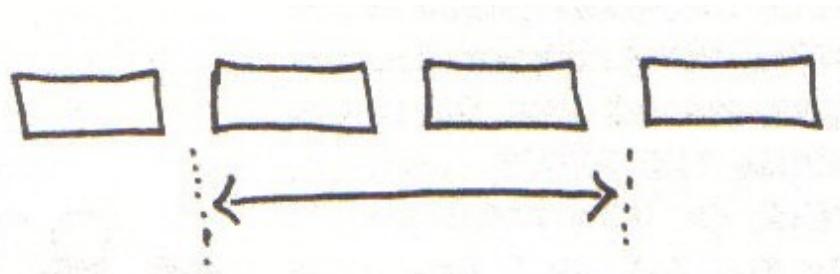
Esta nos permite comparar la variabilidad de dos o mas grupos de datos en los que inclusive la información no tienen las mismas unidades o se trata de datos diferentes.



Medidas de posición



- Los **cuartiles, deciles y percentiles** son valores que dividen los datos en 4, 10 y 100 partes iguales.



- **Frecuencias**

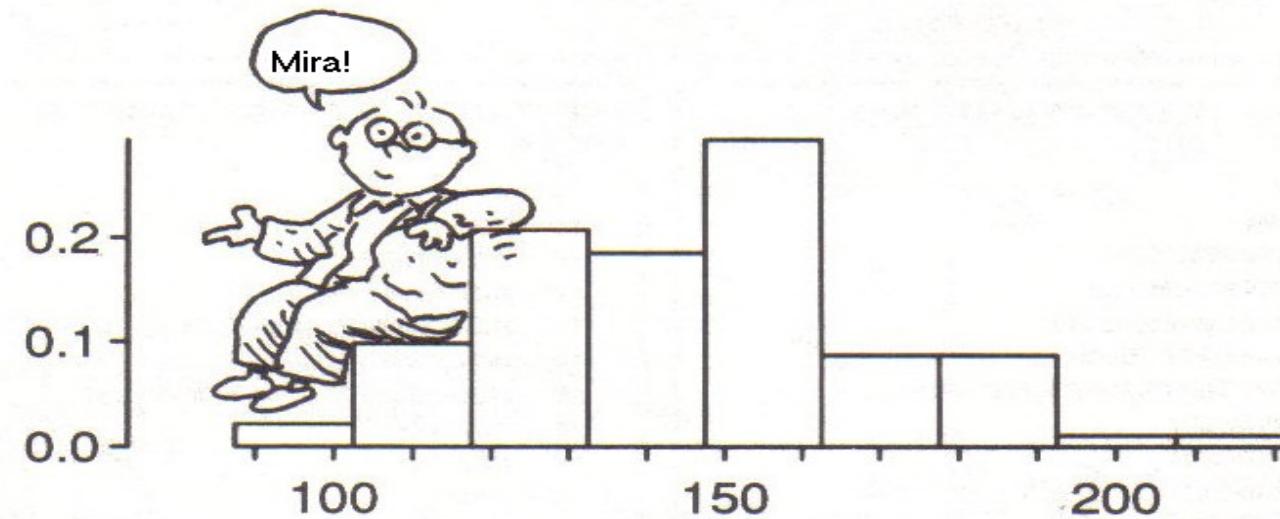
Frecuencia absoluta: Llamaremos así al número de repeticiones que presenta una observación. Se representa por n_i .

Frecuencia relativa: Es la frecuencia absoluta dividida por el número total de datos.

La suma de todas las frecuencias relativas, siempre debe ser igual a la unidad.

$$f_i = \frac{n_i}{n}$$

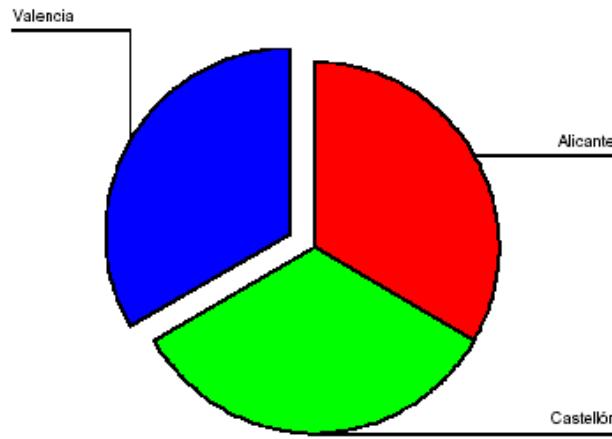
Gráficos



Los gráficos simplifican y aclaran los datos de la investigación facilitando el resumen y la comunicación del significado de los datos.

- **Diagrama de pastel**

Es un círculo dividido en sectores de tamaño proporcional a la frecuencia (sea absoluta o relativa) de cada valor de la variable.



Se utilizan cuando hay pocos valores que representar (máximo de 7).

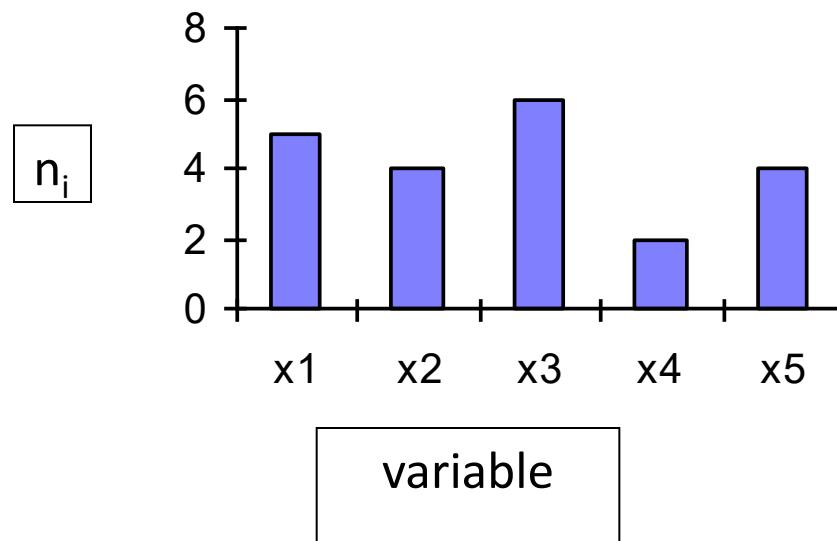
Este diagrama se utiliza para cualquier tipo de variable.

- **Diagrama de barras**

Sobre un eje horizontal se representan los distintos valores de una variable categórica.

Sobre cada valor se levanta un rectángulo cuya base está separada de las contiguas.

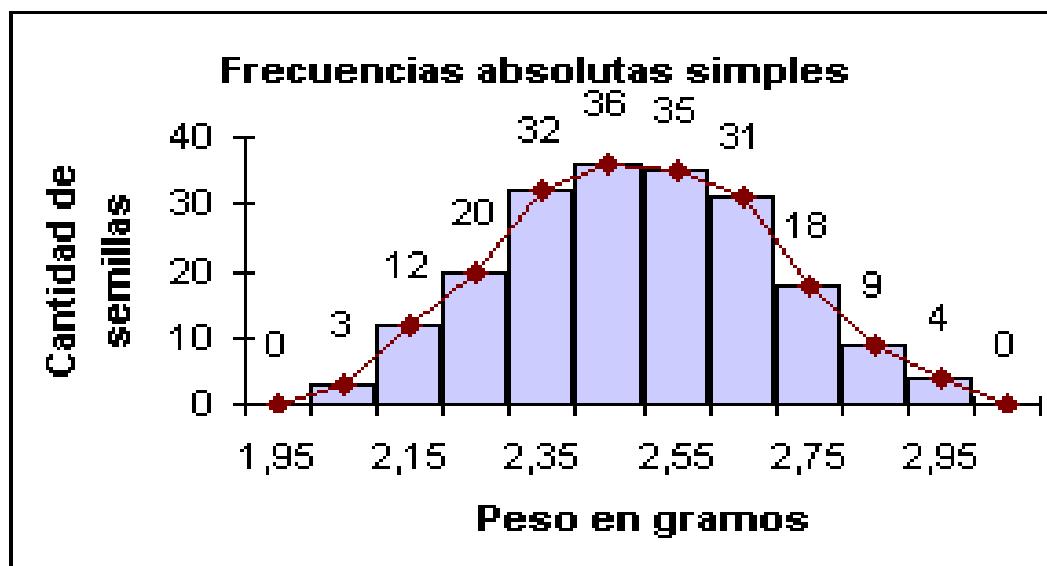
En un eje de escala vertical se representa una característica numérica de la variable por ejemplo el número de casos.



- **Histograma**

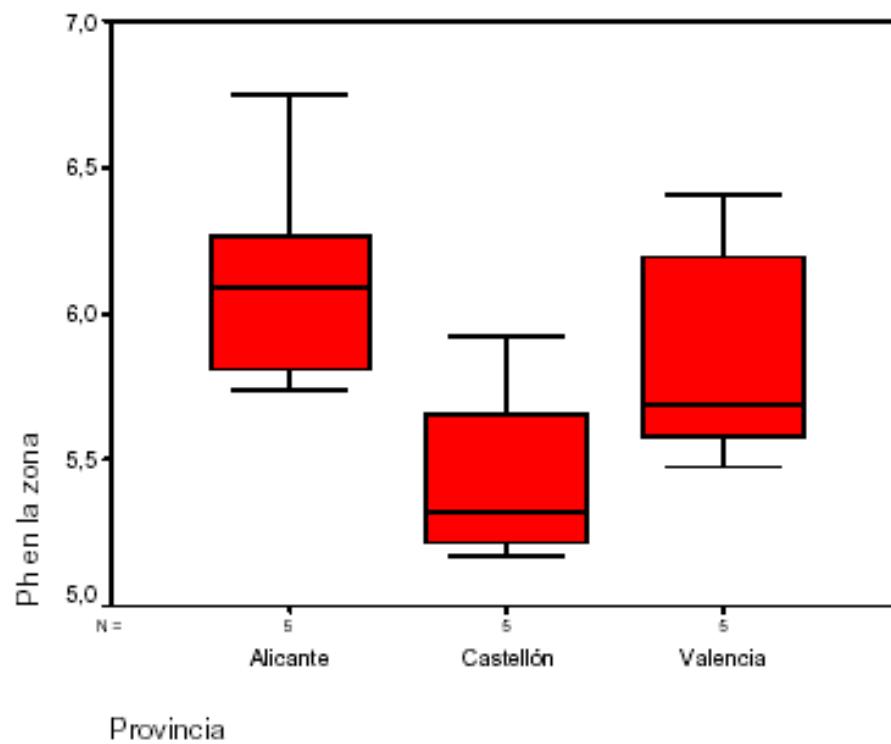
Parecidos en forma a los diagramas de barras, pero su uso se restringe a variables continuas.

Los histogramas representan frecuencias agrupadas de una variable continua sobre intervalos, dibujan rectángulos unidos entre sí, lo que significa que existe una continuidad en la variable.



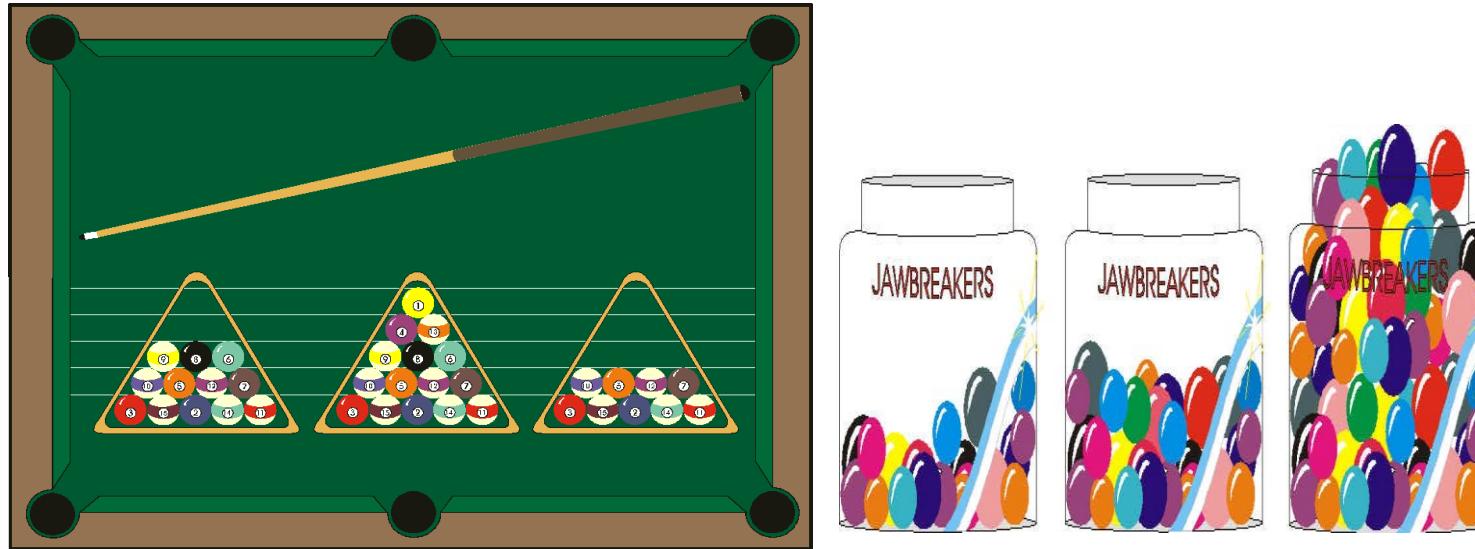
- *Diagramas de caja*

A diferencia de los otros gráficos ya vistos, los diagramas de caja hacen énfasis en las medidas de posición. Es muy útil para hacer comparaciones entre muestras de distintas poblaciones.



- **Pictograma**

Se utilizan para expresar un atributo mediante iconos que se identifiquen con la variable (ejemplo un coche) y su tamaño suele guardar relación con la frecuencia.



Conceptos teóricos



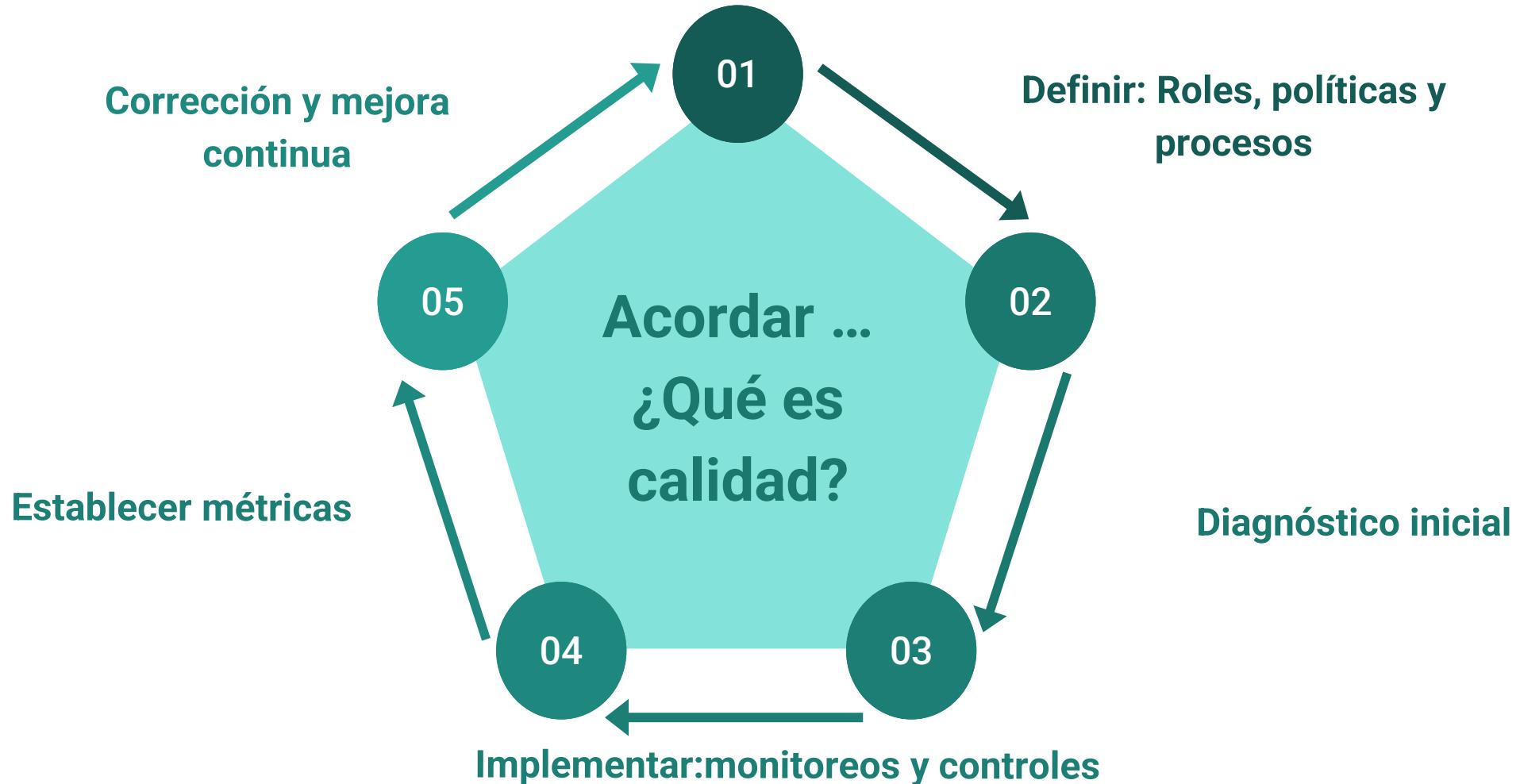
Calidad de datos:

El aseguramiento de la calidad busca que los datos:

- Cumplan su propósito
- Estén disponibles para la toma de decisiones
- Los resultados obtenidos en los análisis sean confiables
- Se de cumplimiento a requerimientos regulatorios



Calidad de datos: ¿Qué abarca este frente?



Calidad de datos: Atributos clave de la calidad de datos



Exactitud	Compleitud	Consistencia	Actualización	Validez	Unicidad
Mide qué tan bien el dato representa la realidad	Los datos tienen todos los elementos requeridos para ser usados. • Existen • % de llenado • Cobertura • Granularidad	Hay coherencia en el dato entre distintos sistemas	Los datos son vigentes y muestran un estado reciente	Los datos cumplen con los formatos y estándares definidos	No se presenta duplicidad en los datos

Calidad de datos: Ejemplos de problemas de calidad

1. Falta de homologación de contenidos:

Los contenidos de los datos de referencia no están todos regidos por un estándar que asegure su correcto uso en el nuevo sistema.

Genero:

“Masculino”,
“Femenino”, “M”, “F”,
“1”, “2”,

2. Campos sin ordenamiento.

Dentro del espacio establecido para almacenamiento no existe un orden estandarizado entre todos los registros.

Nombre:

“Jorge Arturo López”
“Rodriguez Torres Andrea”

3. Formatos diferentes en un mismo campo.

Los contenidos registran diferentes formatos para reportar el dato y al momento de migración puede no cumplir con el estándar requerido

Fecha de Nacimiento:

“2012-03-15” ,
“20101112”
“16/03/1998” ,
“13/23/98”

Calidad de datos: Ejemplos de problemas de calidad

4. Duplicidad de registros maestros.

Un duplicado es un registro Maestro con diferente código de identificación pero que se registraron como nuevos clientes o nuevos productos.

Datos Cliente:	
19257412	Jorge Arturo Lopez
1257412	Jorge Lopez

5. Datos Faltantes.

Los datos no fueron reportados y en su lugar se reportó datos no consistentes.

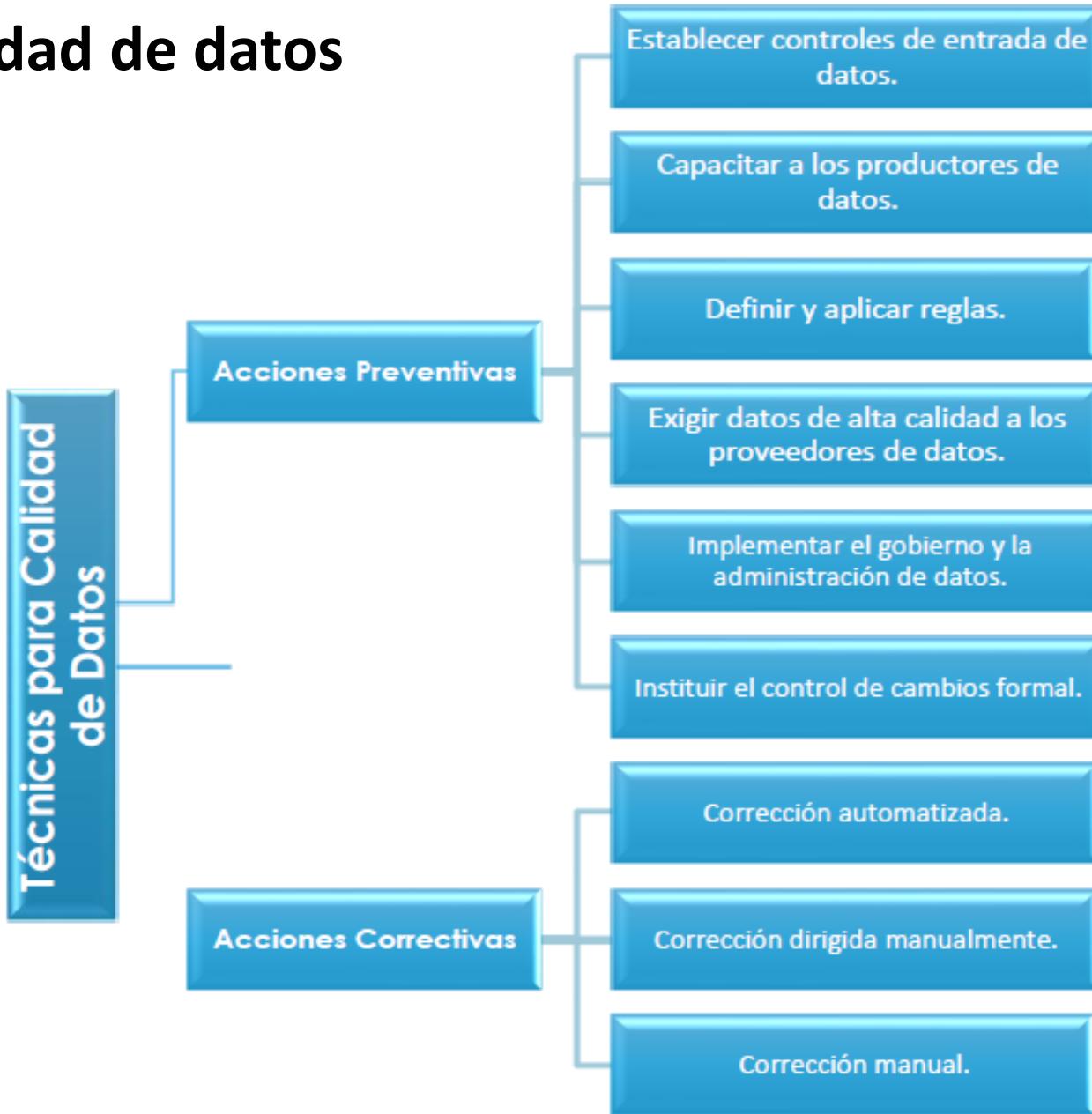
Dirección de Residencia:
“1”
“No Dijo”

6. No hay diferenciación de clientes empresa y clientes personas.

No permite saber de forma automatizada si el cliente es persona o empresa.

Nombre:
“Jorge Arturo López Moncada”
“Industrias Paquita”
“Almacén la Estrella”

Técnicas de Calidad de datos

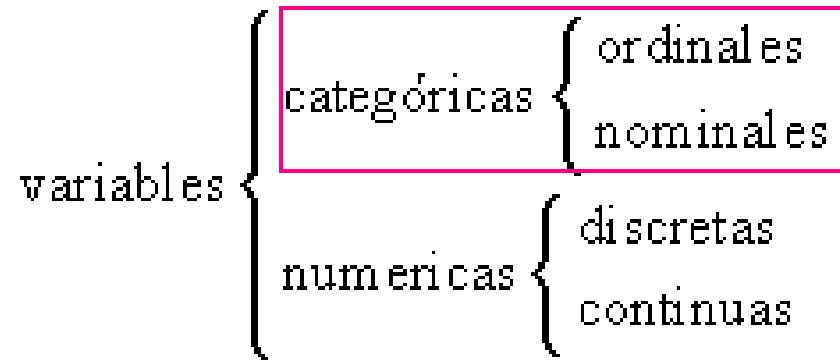


Métodos de asociación

TEMAS DE LA SESIÓN

- Tablas de frecuencias
- Análisis bivariado Variables categóricas
- Tablas de contingencia y prueba Chi-cuadrado de independencia
- Correlación variables continuas

VARIABLES CATEGORICAS



Las **variables categóricas** son aquellas cuyos valores indican categorías o son etiquetas alfanuméricas o "nombres". A su vez se clasifican en:

- **nominales:** Sus posibles valores son mutuamente excluyentes entre sí y no tienen alguna forma "natural" de ordenación
- **ordinales:** tienen algún orden. Por ejemplo: "nunca sucede", "la mitad de las veces" y "siempre sucede".

Tablas de frecuencia

- Una **tabla de frecuencias** organiza los datos en **clases**, es decir, en grupos de valores que describen una característica de los datos y muestra el número de observaciones del conjunto de datos que caen en cada una de las clases definidas.

Tablas de frecuencias

FRECUENCIA ABSOLUTA: Se calcula para cada categoría contando cuantos datos pertenecen a esta.

FRECUENCIA RELATIVA: Se calcula como la frecuencia en porcentaje sobre el total de datos e indica el porcentaje de datos que esta en determinada categoría.

Tablas de frecuencias

Algunas veces, las clases de las tablas están dadas por alguna variable de agrupación o categórica tales como el género o el estado civil, sin embargo, hay variables continuas como el peso o el tiempo hasta algún evento que no tienen las clases definidas, para estos efectos entonces se recomienda:

Tablas de frecuencias

1. Determinar a su criterio una cantidad de clases no muy alta, pues si son demasiadas, distorsionará la distribución de los datos dejando clases de baja frecuencia.
2. Para ayudarse para el número de clases correctas mediante una formula matemática se tiene la formula de Sturges:

$$\text{Nº de clases} = 1 + 3.332 \log N$$

Tablas de frecuencias

3. La longitud de cada clase también puede determinarse mediante la fórmula:

$$\text{Long de clase} = (\text{Max.-min.})/k$$

Después de realizar esta clasificación se procede a graficar los datos para ver su forma.

EJEMPLO

Considere los siguientes 50 datos:

12	11	4	6	6	11	3	10	12	4
10	1	1	2	4	5	2	4	4	8
8	7	8	4	10	4	2	6	2	9
5	6	6	4	12	8	1	12	1	7
7	6	8	4	6	9	3	7	7	5

- N° de clases = $1 + 3.332\log (50) = 6$
- Tamaño de clase = $11/6 = 2$

EJEMPLO

Clase	Intervalo		Frec. Absoluta	Frec. Relativa	Frec. Porcentual
	LI	LS			
1	1	2.9	8	0.16	16 %
2	3	4.9	11	0.22	22 %
3	5	6.9	10	0.20	20 %
4	7	8.9	10	0.20	20 %
5	9	10.9	5	0.10	10 %
6	11	12.9	6	0.12	12 %
Total			50	1	100 %

TABLAS DE CONTINGENCIA

Son un recurso estadístico para que registrar y analizar la relación entre **dos o más variables**, de naturaleza cualitativa.

Var 1 / Var 2	Categoría a	Categoría b	total
Categoría 1	N _{1a}	N _{1b}	N _{1total}
Categoría 2	N _{2a}	N _{2b}	N _{2total}
total	N _{a total}	N _{b total}	N

N_{1a}: Número de elementos, personas, que asumen la categoría a en la variable uno y la categoría 1 en la variable dos.

N_{1total}: Cantidad total de elementos, personas que asumen la categoría 1 de la variable uno.

N: Total de elementos, personas, registradas en la tabla.

EJEMPLO

Tenemos dos variables categóricas para una muestra de 100 personas económicamente activas:

- ▶ Sexo
- ▶ Ocupación: Empleado, desempleado

Sexo / Ocup	Empleado	desempleado	total
Hombre	43	9	52
Mujer	44	4	48
total	87	13	100

Frecuencias marginales y relativas

Las frecuencias absolutas pueden distorsionar la realidad de la relación entre las dos variables, para esto se utilizan las frecuencias relativas.

- Relativas a la fila
- Relativas a la columna
- Relativas al total

Frecuencias relativas a la fila

Continuando con el ejemplo de las variables sexo y ocupación, las frecuencias relativas a la fila son:

Sexo/Ocup	Empleado	Desempleado	Total
Hombre	0,83	0,17	1
Mujer	0,92	0,08	1
Total	0,87	0,13	1

Distribución de la ocupación según el sexo

Frecuencias relativas a la columna

Continuando con el ejemplo de las variables sexo y ocupación, las frecuencias relativas a la columna son:

Sexo/ocup	Empleado	desempleado	Total
Hombre	0,49	0,69	0,52
Mujer	0,51	0,31	0,48
Total	1	1	1

Distribución de sexos según ocupación

Frecuencias relativas al total

Para el ejemplo de las variables sexo y ocupación, las frecuencias relativas a la columna son:

Sexo/ocup	Empleado	desempleado	Total
Hombre	0,43	0,09	0,52
Mujer	0,44	0,04	0,48
Total	0,87	0,13	1

Distribución general del total de la población

Prueba Chi-cuadrado

Determina si el hecho de que un individuo tome una categoría X de una de las variables esta relacionada con el hecho que también tome la categoría z de la otra variable, es decir si son o no independientes.



Prueba Chi-cuadrado

Supongamos una tabla de contingencia así:

X \ Y	y_1	y_2	y_i	y_J	$F_i = \sum_j O_{ij}$
x_1	O_{11}	O_{12}	O_{1j}	O_{1J}	F_1
x_2	O_{21}	O_{22}	O_{2j}	O_{2J}	F_2
....
x_i	O_{i1}	O_{i2}	O_{ij}	O_{iJ}	F_i
...
x_J	O_{J1}	O_{J2}	O_{Jj}	O_{JJ}	F_J
$C_j = \sum_i O_{ij}$	C_1	C_2	C_i	C_J	T

Con todas las frecuencias mayores a cinco

Prueba Chi-cuadrado

H_0 = Las variables son independientes

H_1 = Las variables no son independientes

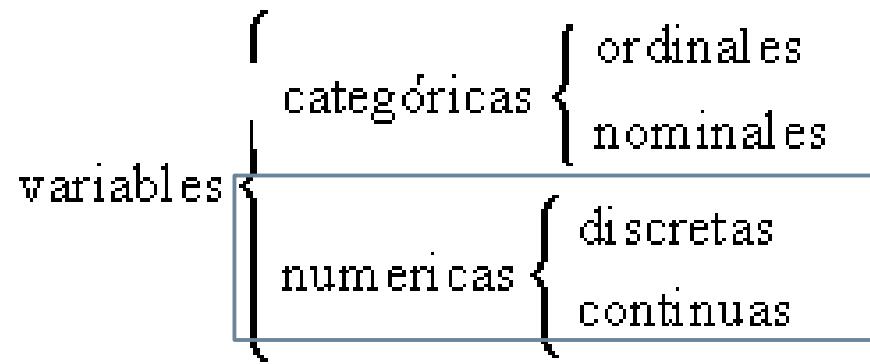
El estadístico de prueba para este sistema es:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

$$E_{ij} = \frac{F_i C_j}{T}$$

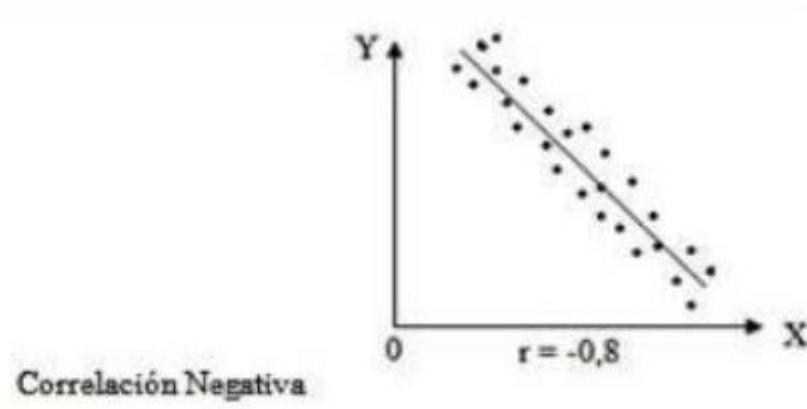
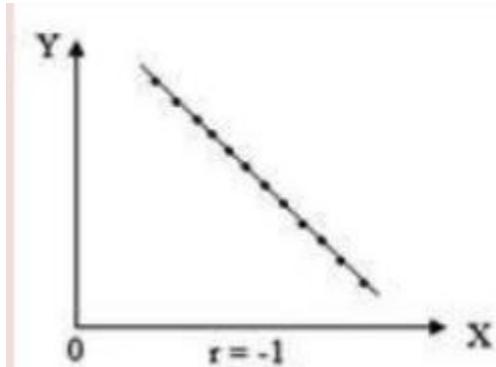
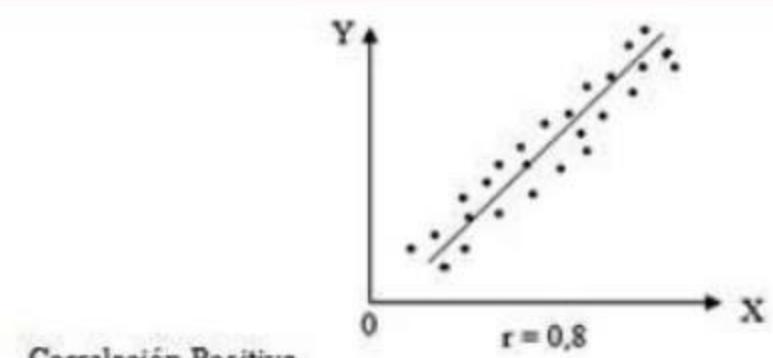
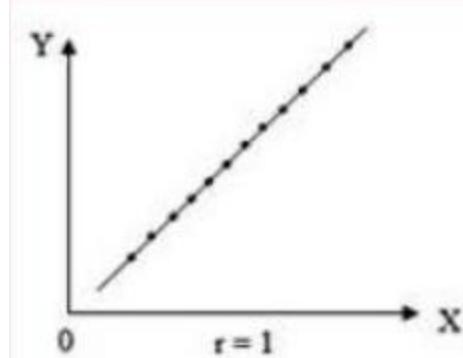
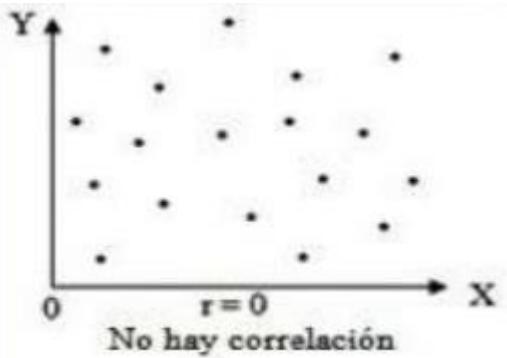
VARIABLES CONTINUAS



Las **variables categóricas** son aquellas cuyos valores indican categorías o son etiquetas alfanuméricas o "nombres". A su vez se clasifican en:

- **nominales**: Sus posibles valores son mutuamente excluyentes entre sí y no tienen alguna forma "natural" de ordenación
- **ordinales**: tienen algún orden. Por ejemplo: "nunca sucede", "la mitad de las veces" y "siempre sucede".

Correlación lineal



Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (r) es un índice que mide la magnitud de la relación lineal entre 2 variables cuantitativas, así como el sentido, positivo o negativo, de dicha relación.

Indica en qué grado 2 variables X e Y fluctúan simultáneamente, es decir cuánto aumenta X al aumentar Y (correlación positiva), o cuánto aumenta X al disminuir Y (correlación negativa).

Coeficiente de correlación de Pearson

Se define la covarianza como:

$$\text{Covarianza} = \frac{\sum (\bar{X} - X) * (\bar{Y} - Y)}{n - 1}$$

Entonces el coeficiente de correlación entre X y Y:

$$r = \frac{\text{covarianza}}{S_x * S_y}$$

Ejemplo

Hallando el coeficiente de correlación entre las variables X y Y

Y	X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X}) * (Y - \bar{Y})$
9	72	5.65	1.4	7.91
10	76	9.65	2.4	23.16
6	59	-7.35	-1.6	11.76
8	68	1.65	0.4	0.66
10	60	-6.35	2.4	-15.24
5	58	-8.35	-2.6	21.71
8	70	3.65	0.4	1.46
7	65	-1.35	-0.6	0.81
4	54	-12.35	-3.6	44.46
11	83	16.65	3.4	56.61
7	64	-2.35	-0.6	1.41
7	66	-0.35	-0.6	0.21
6	61	-5.35	-1.6	8.56
8	66	-0.35	0.4	-0.14
5	57	-9.35	-2.6	24.31
11	81	14.65	3.4	49.81
5	59	-7.35	-2.6	19.11
9	71	4.65	1.4	6.51
6	62	-4.35	-1.6	6.96
10	75	8.65	2.4	20.76
				$\sum 290.8$

Ejemplo

$$\bar{X} = 66.35$$

$$\bar{Y} = 7.6$$

$$Covarianza = \frac{\sum (\bar{X} - X) * (\bar{Y} - Y)}{n - 1} = \frac{290.8}{19} = 15.30$$

$$r = \frac{covarianza}{S_x * S_y} = \frac{15.30}{8.087 * 2.137} = 0.885$$

Correlación de rango de Spearman

- **Planteamiento de Hipótesis:**

$\rho=0$ No hay correlación

$\rho \neq 0$ Hay correlación

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

<https://es.slideshare.net/byrong/correlacion-rango-spearman>

Correlación de rango de Spearman

Vendedor	Lugar en potencial	Ventas (u) en dos años	Lugar según ventas en 2 años	di	di2
A	2	400	1	1	1
B	4	360	3	1	1
C	7	300	5	2	4
D	1	295	6	-5	25
E	6	280	7	-1	1
F	3	350	4	-1	1
G	10	200	10	0	0
H	9	260	8	1	1
I	8	220	9	-1	1
J	5	385	2	3	9

Sumatoria 44

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$
$$r_s = 1 - \frac{6 * 44}{10(100 - 1)} = 0.73$$

IA : Programas con la capacidad de aprender y razonar como humanos

AM : Algoritmos con la habilidad de aprender sin ser explícitamente programados

AP : Subconjunto de **AM** que mediante redes neurales adaptativas aprenden de grandes cantidades de datos

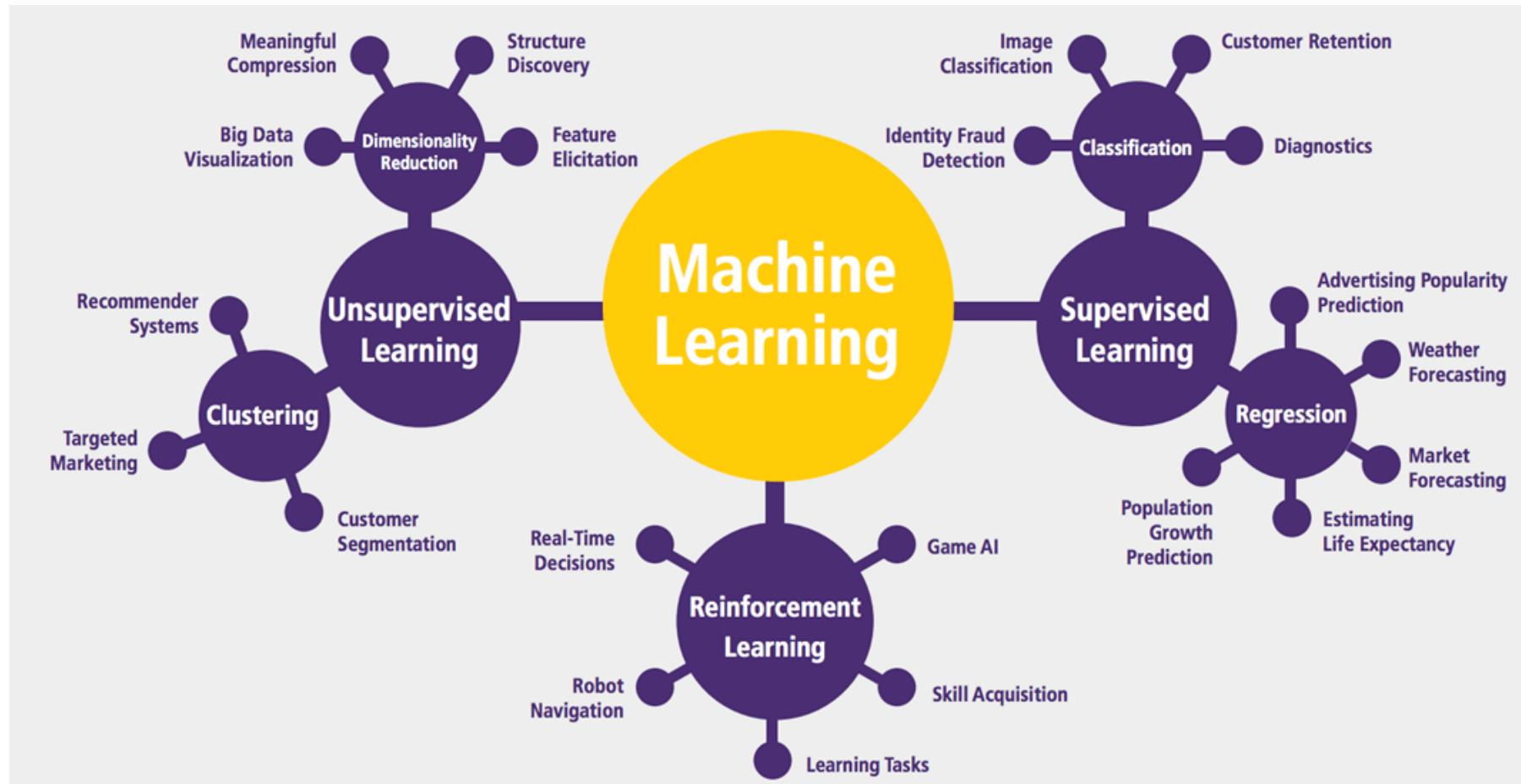
IA GEN : Crear nuevos contenidos de texto, imagen y voz mediante órdenes o datos

Inteligencia Artificial

Aprendizaje de Maquina

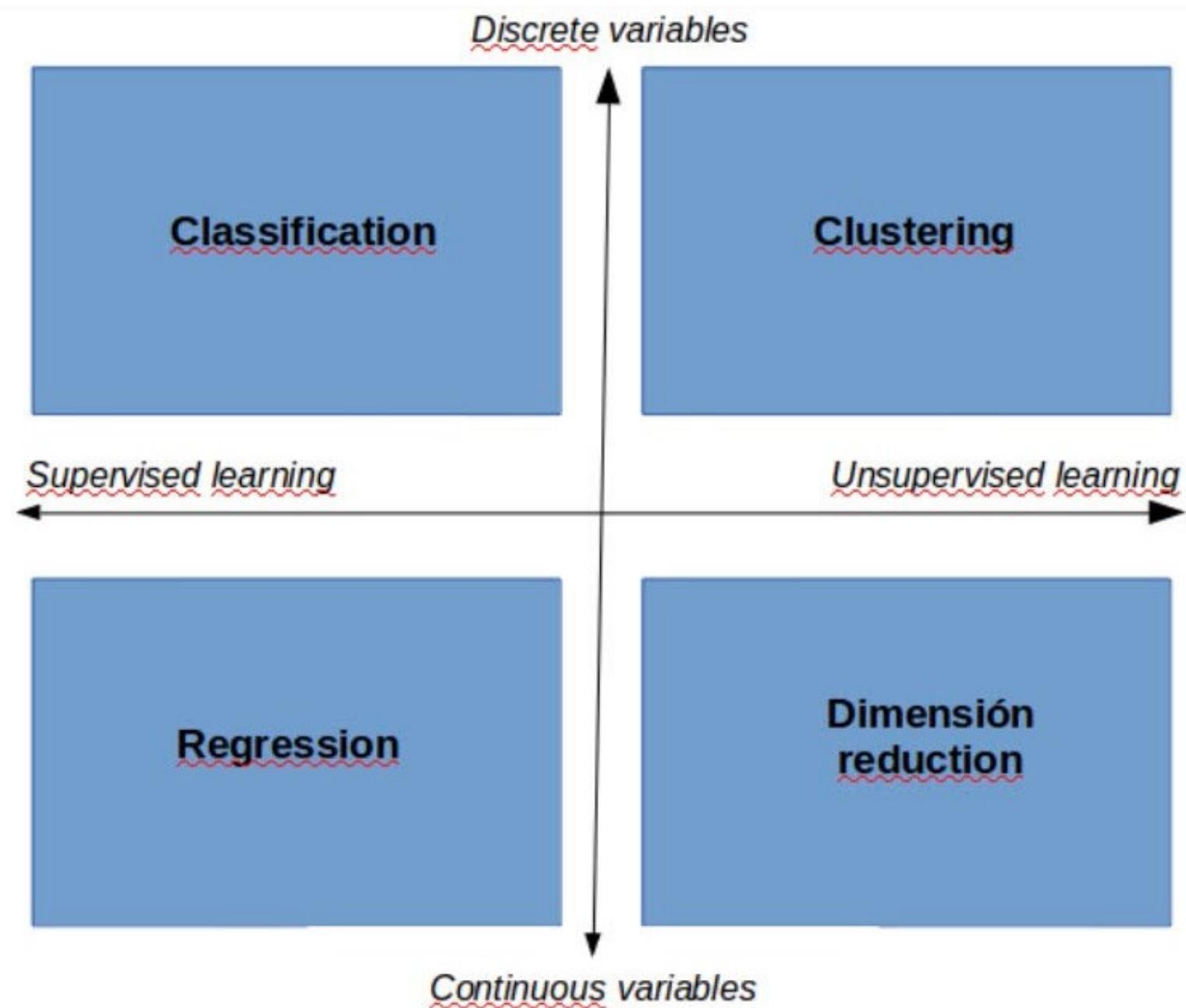
Aprendizaje Profundo

IA Generativa

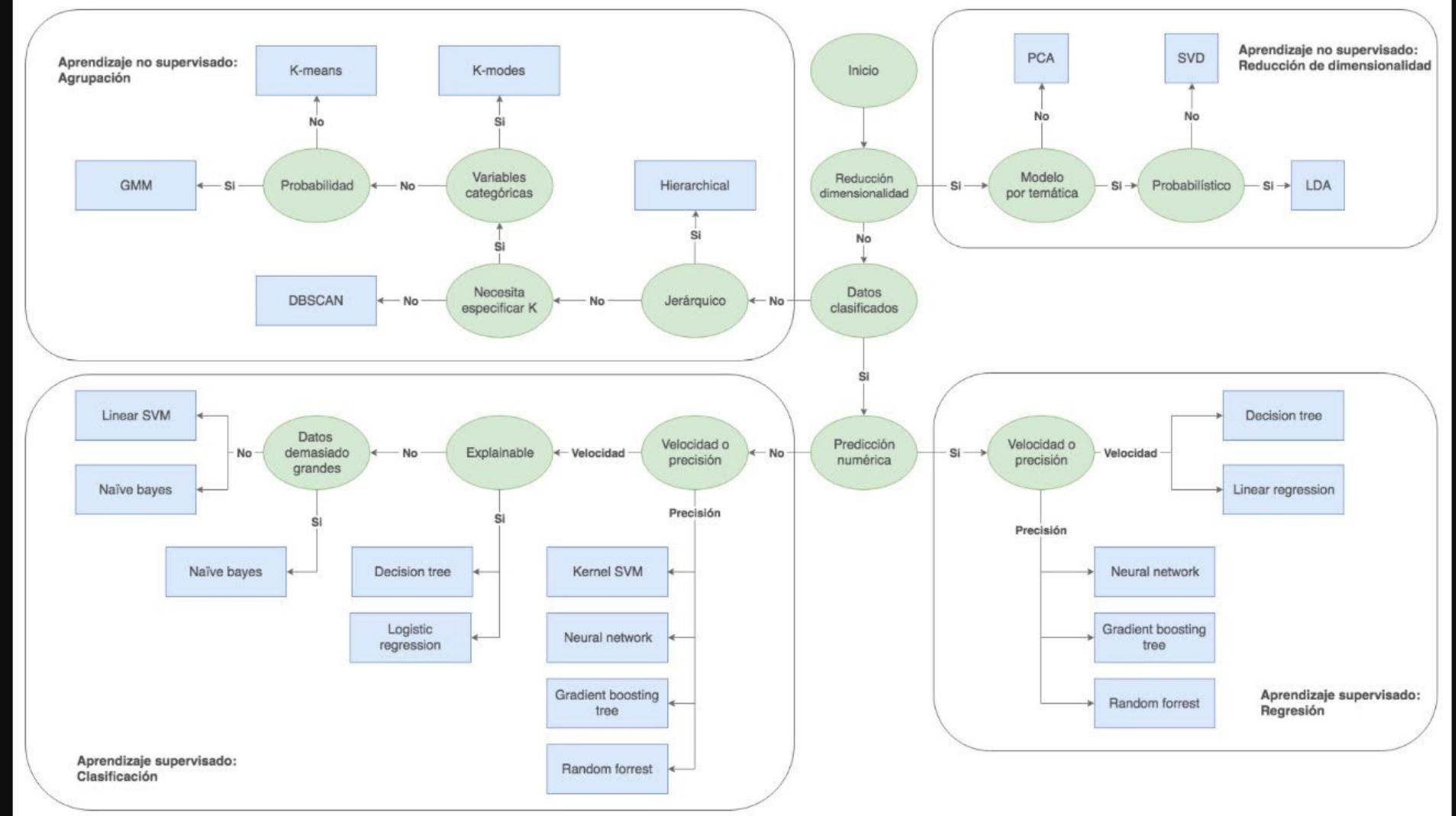


Tipos de Algoritmos

- Clasificación
- Detección de anomalías
- Regresión
- Clusterización

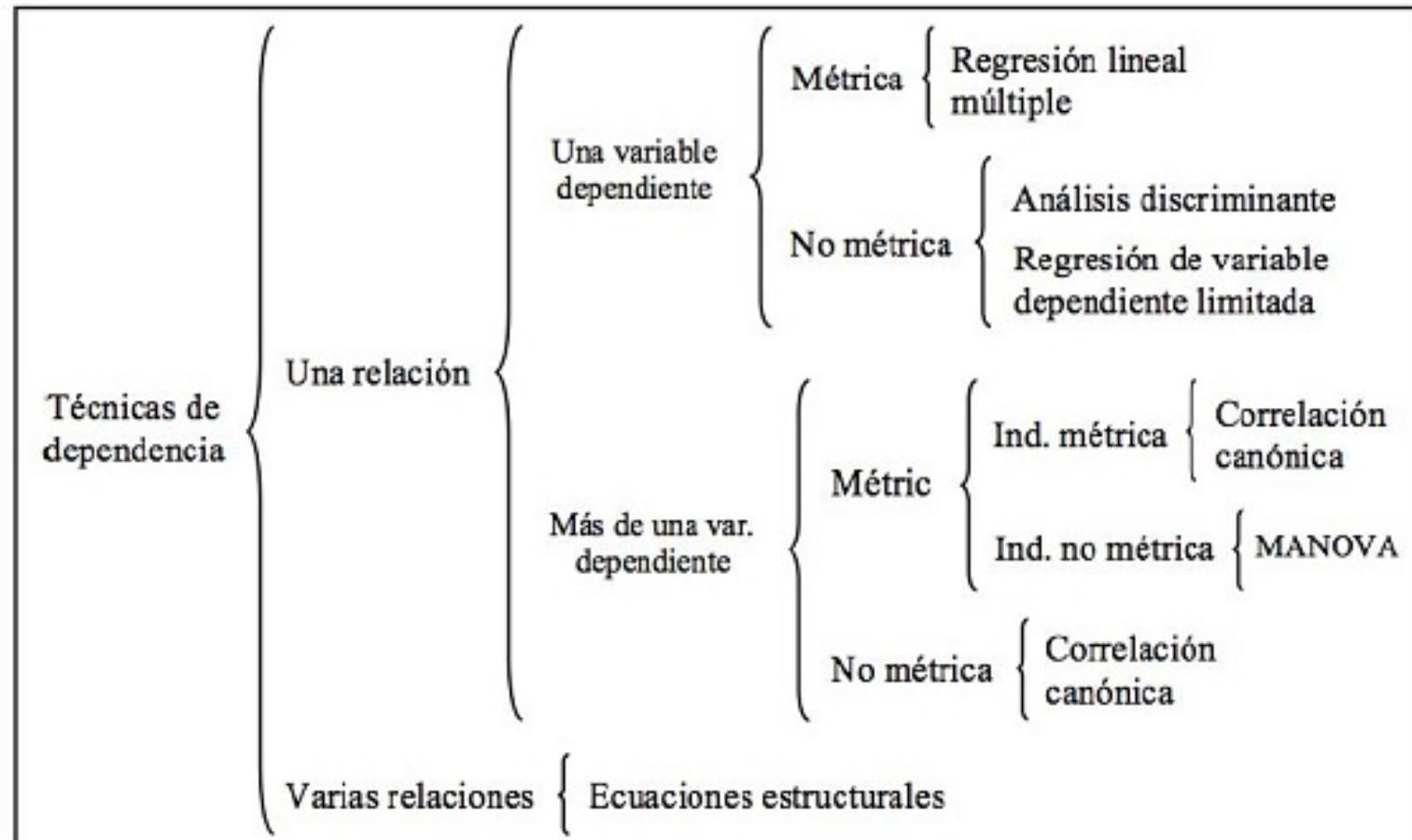




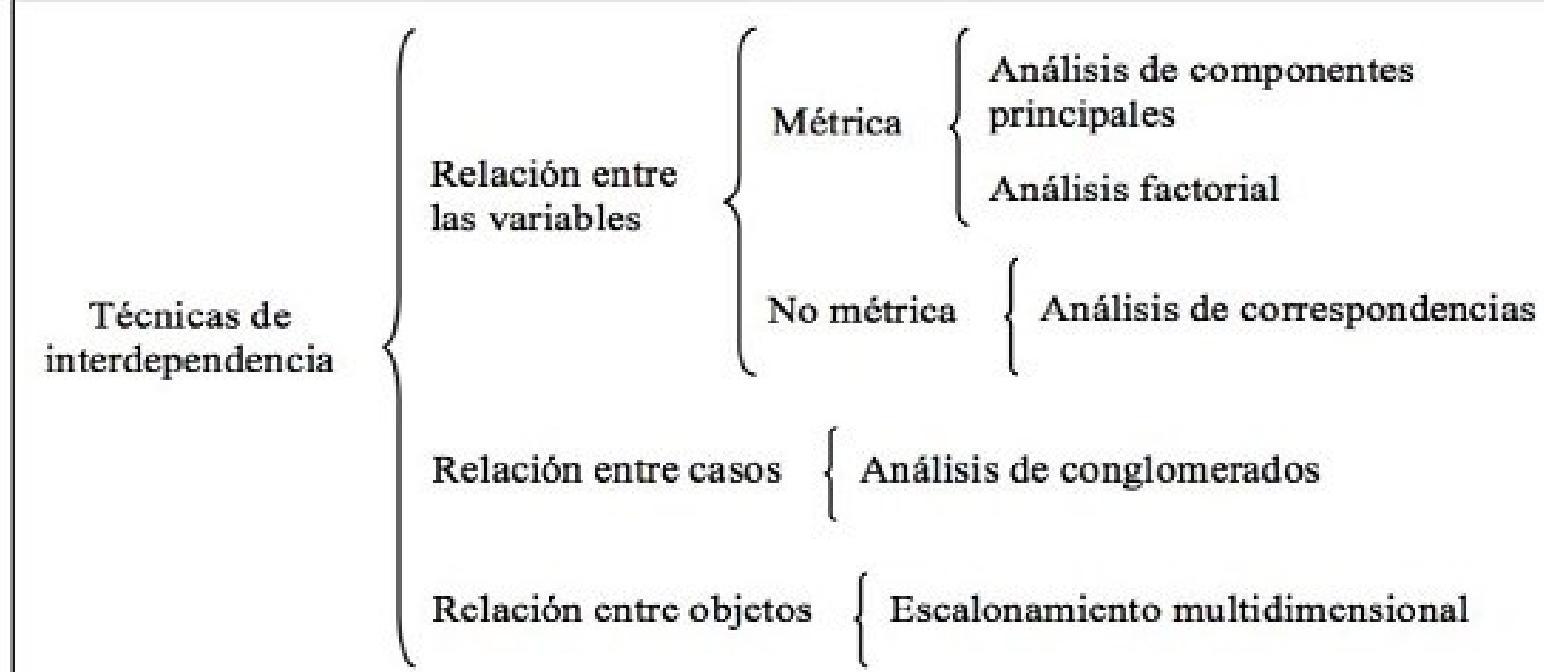


Type	Name	Description	Advantages	Disadvantages
Linear	Linear regression	The "best fit" line through all data points. Predictions are numerical.	Easy to understand – you clearly see what the biggest drivers of the model are.	<ul style="list-style-type: none"> ✗ Sometimes too simple to capture complex relationships between variables. ✗ Tendency for the model to "overfit".
	Logistic regression	The adaptation of linear regression to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	<ul style="list-style-type: none"> ✗ Sometimes too simple to capture complex relationships between variables. ✗ Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a branching method to match all possible outcomes of a decision.	Easy to understand and implement.	<ul style="list-style-type: none"> ✗ Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but by combining them we get better overall performance.	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	<ul style="list-style-type: none"> ✗ Can be slow to output predictions relative to other algorithms. ✗ Not easy to understand predictions.
	Gradient Boosting	Uses even weaker decision trees, that are increasingly focused on "hard" examples.	High-performing.	<ul style="list-style-type: none"> ✗ A small change in the feature set or training set can create radical changes in the model. ✗ Not easy to understand predictions.
Neural networks	Neural networks	Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several layers of neural networks put one after the other.	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	<ul style="list-style-type: none"> ✗ Very, very slow to train, because they have so many layers. Require a lot of power. ✗ Almost impossible to understand predictions.





Fuente: Adaptado de Uriel y Aldás (2005).



Fuente: Adaptado de Uriel y Aldás (2005).

Variables endógenas exógenas y Regresión lineal

TEMAS DE LA SESIÓN

- Definición de variables endógenas y exógenas.
- Definición de regresión lineal simple, utilidad y marco teórico.
- Supuestos de validación, medidas de bondad de ajuste y aplicación.

SE ESPERA QUE AL FINAL USTED...

- Logre identificar en un conjunto de variables cuales pueden ser consideradas como endógenas o exógenas.
- Ejecute una regresión lineal simple y verifique que los supuestos para hacer posible su interpretación se cumplan.

Variable endógena, dependiente o a predecir

Como su palabra lo dice, son características de la realidad que se ven determinadas o que dependen del valor que asuman otros fenómenos o variables independientes.

Ejemplos:

Crecimiento de una planta en cm.

Satisfacción de un cliente con un producto.

Precio de una vivienda.

Variable exógena, independiente o predictora

Son aquellas en las cuales los cambios en los valores de ellas determinan cambios en los valores de otra variable (variable dependiente).

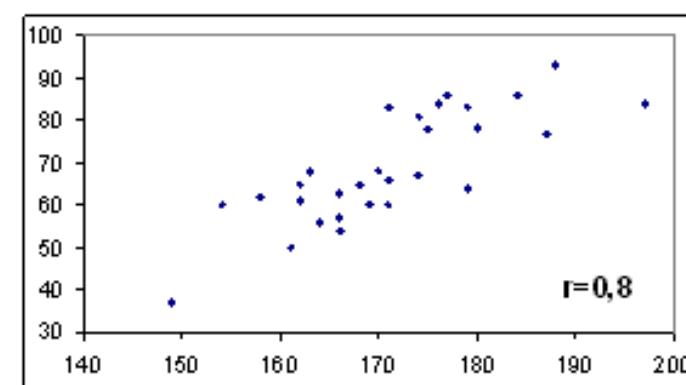
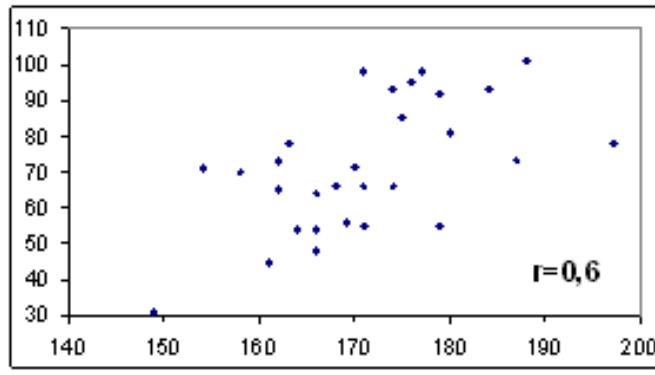
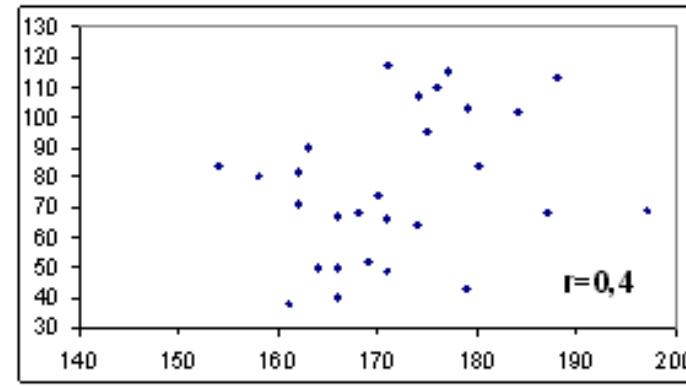
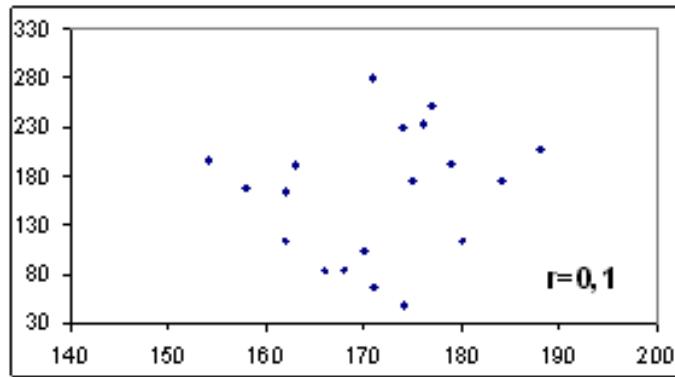
Ejemplos:

Cantidad de agua que se le pone a la planta.

Atención al cliente en puntos de venta y pago.

Material del cual se construyo la casa.

Tipos comunes de relación



Coeficiente de correlación de Pearson

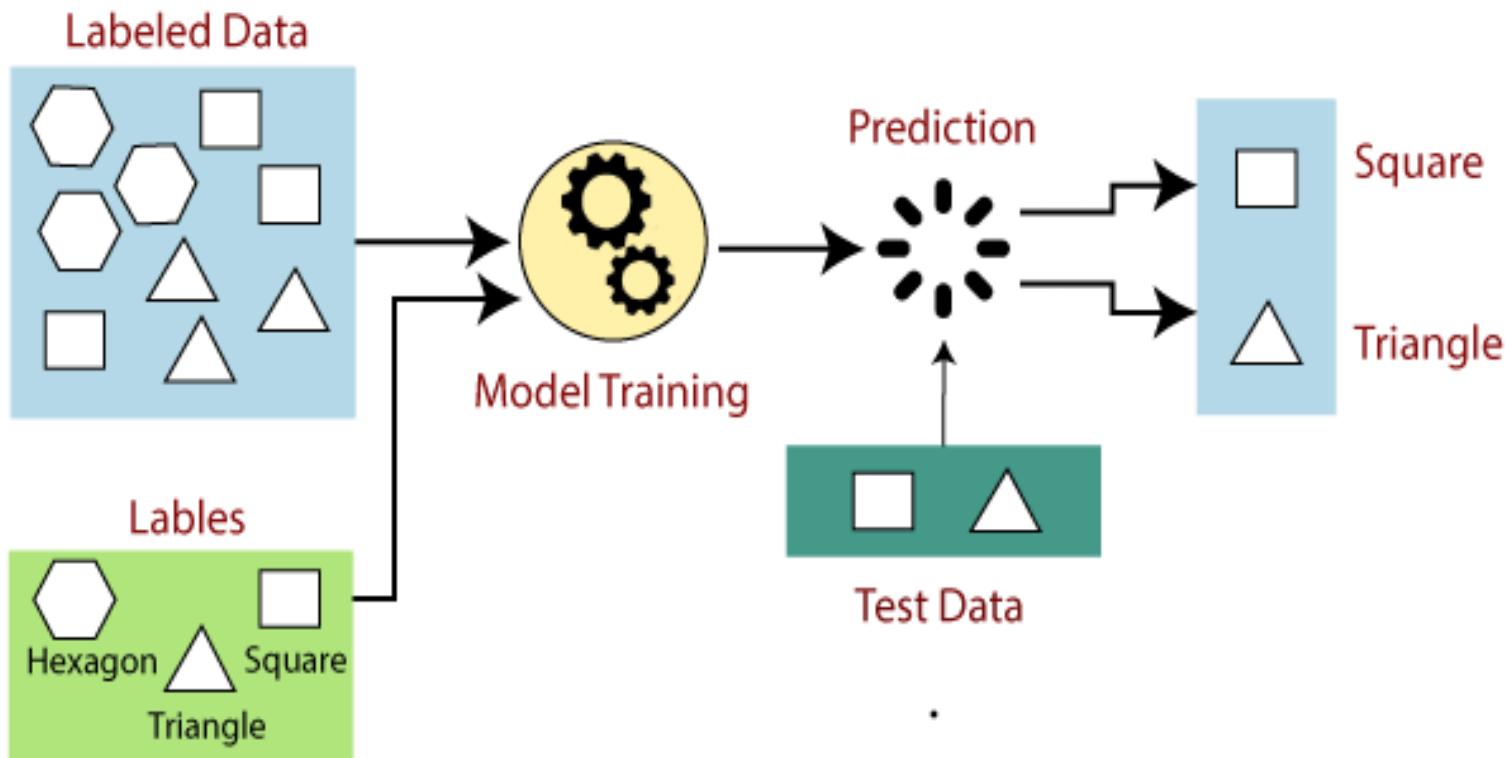
Las relaciones anteriores se miden con el coeficiente r que es útil para determinar si hay relación **lineal** entre dos variables, pero **no servirá** para otro tipo de relaciones (cuadrática, logarítmica,...)

$$r_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{\sum (X - \bar{X})^2}{N}} \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}}$$

Propiedades de r

- Es adimensional.
- Sólo toma valores en $[-1,1]$.
- Las variables son incorreladas $\Leftrightarrow r = 0$.
- Relación lineal perfecta entre dos variables $\Leftrightarrow r = +1$ o $r = -1$.
- Cuanto más cerca esté r de $+1$ o -1 mejor será el grado de relación lineal.
 - Siempre que no existan observaciones anómalas.

Modelos supervisados



Características

- Existe un label o target.
- El objetivo es pronosticar o estimar por cada individuo o registro su variable label o target.
- El label o target puede ser categórico (clasificación)

Ejemplo

- El cliente va pagar o no.
- El cliente va adquirir o no un producto
- El cliente va cancelar o no sus productos (estrategia de retención).

Sin importar la naturaleza de los datos, siempre se deben llevar una tabla estructurada

Ejemplo 1: Se desea construir un modelo de aprendizaje supervisado para identificar los clientes más propensos a cancelar sus productos en los próximos **dos** meses.

Cliente	Saldo TC	Saldo cta ahorros	...	Churn TC
1	1000000	5000000		0
2	0	100000		1

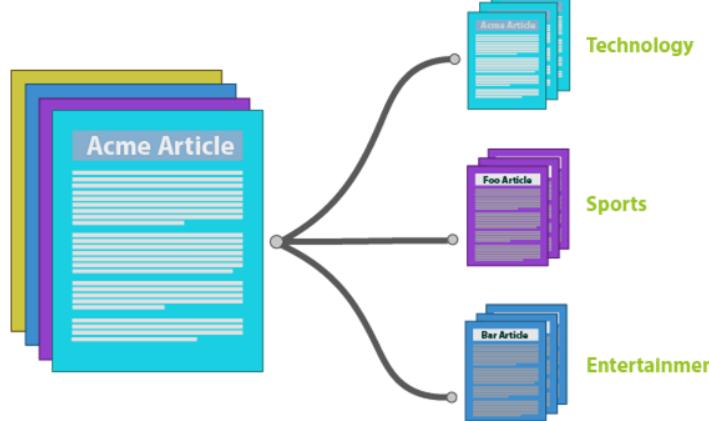
La información de los clientes debe corresponder a un mes determinado (ejemplo, Febrero 2019)

En los datos, siempre se debe tener identificado la variable objetivo o target

Se observa **dos** meses después (Abril 2019) y se revisa si el cliente ha cancelado o no sus productos.

Sin importar la naturaleza de los datos, siempre se deben llevar una tabla estructurada

Ejemplo 2 clasificación de textos: Se desea construir un modelo de aprendizaje supervisado para clasificar documentos como tecnología, deportes o entretenimiento.



1. Convertir a minúsculas.
2. Eliminar caracteres especiales.
3. Eliminar stopwords
4. Lematizar

1. Proceso de vectorización, TF-IDF es el método más común.
2. Cada palabra pasa a ser una variable y su valor es la multiplicación de TF e IDF

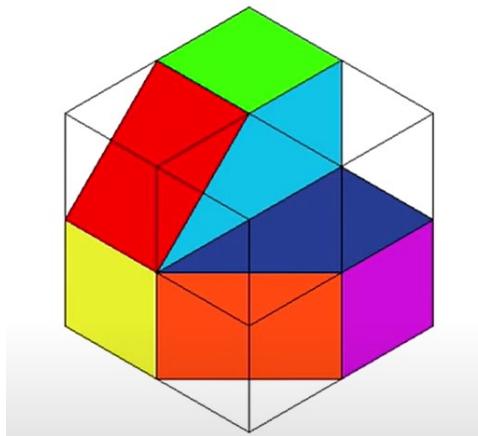
Document	FIFA	go	rock	...	Tipo
1	0	0	0.2		0
2	0.4	0	0		1

$$IDF = \log[(\# \text{ Number of documents}) / (\text{Number of documents containing the word})]$$

$$TF = (\text{Number of repetitions of word in a document}) / (\# \text{ of words in a document})$$

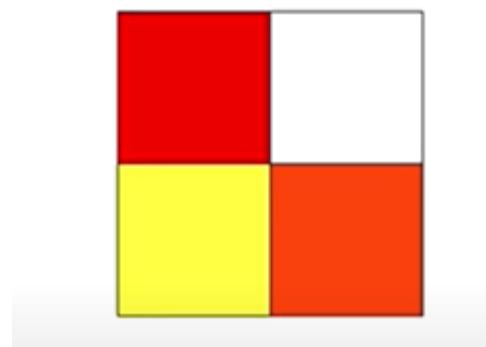
Por Qué machine learning generalmente tiene más poder predictivo?

Problema real a predecir



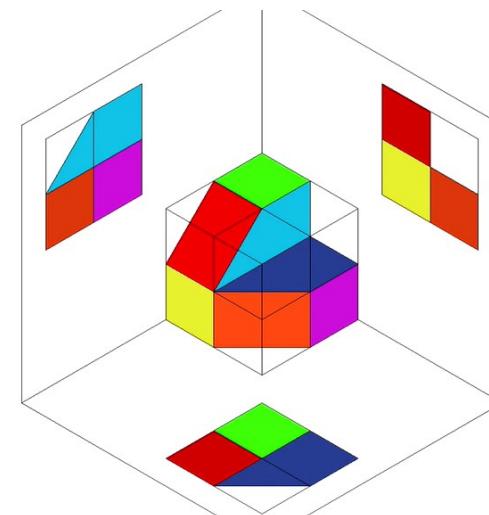
Los problemas a resolver son multidimensionales y altamente complejos.

Algoritmos tradicionales



Podemos tener una visual del problema que soluciona nos da una representación de una parte del problema

Machine learning

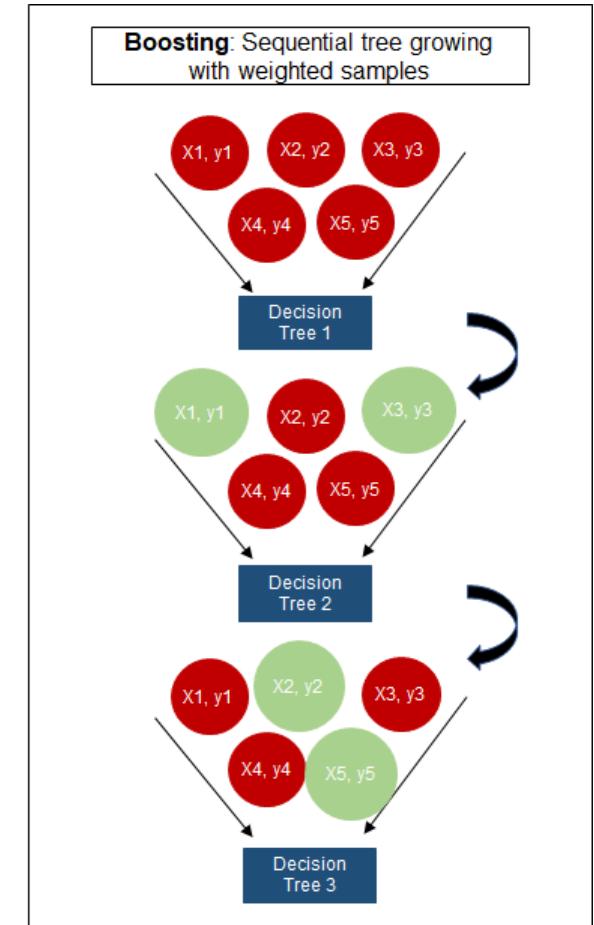
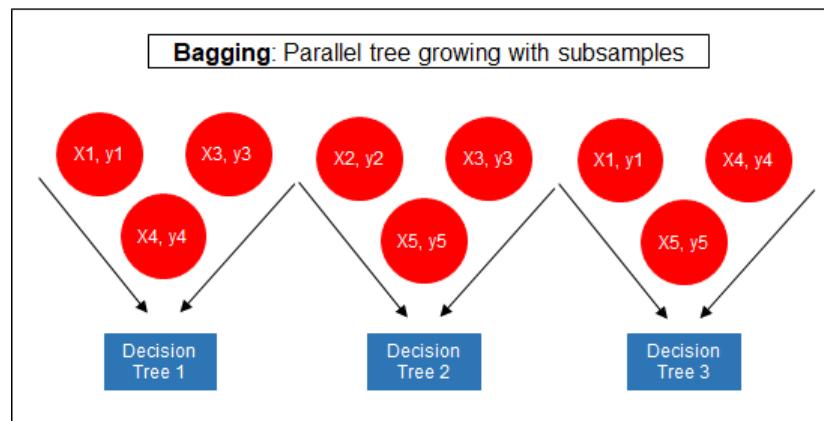
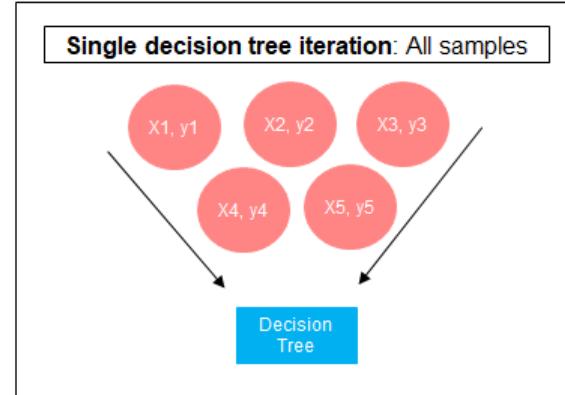


Machine learning aproxima el problema de una manera multidimensional y aunque no nos da toda las vistas nos dá más herramientas

Algunos algoritmos de aprendizaje supervisados: Xgboost

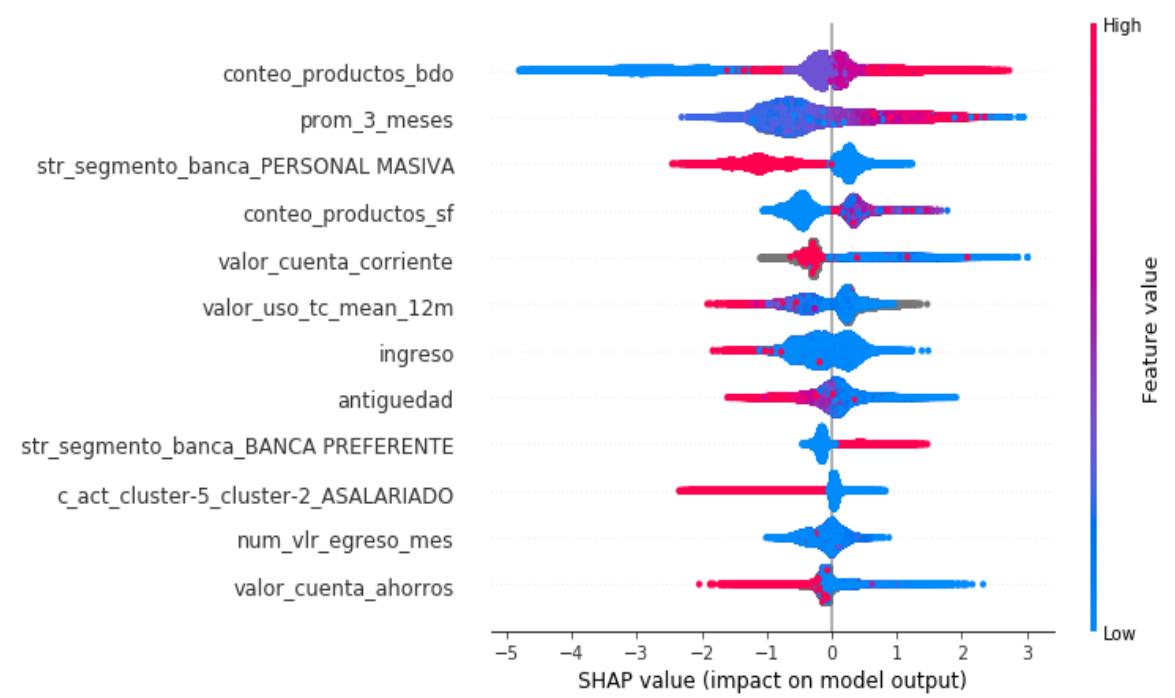
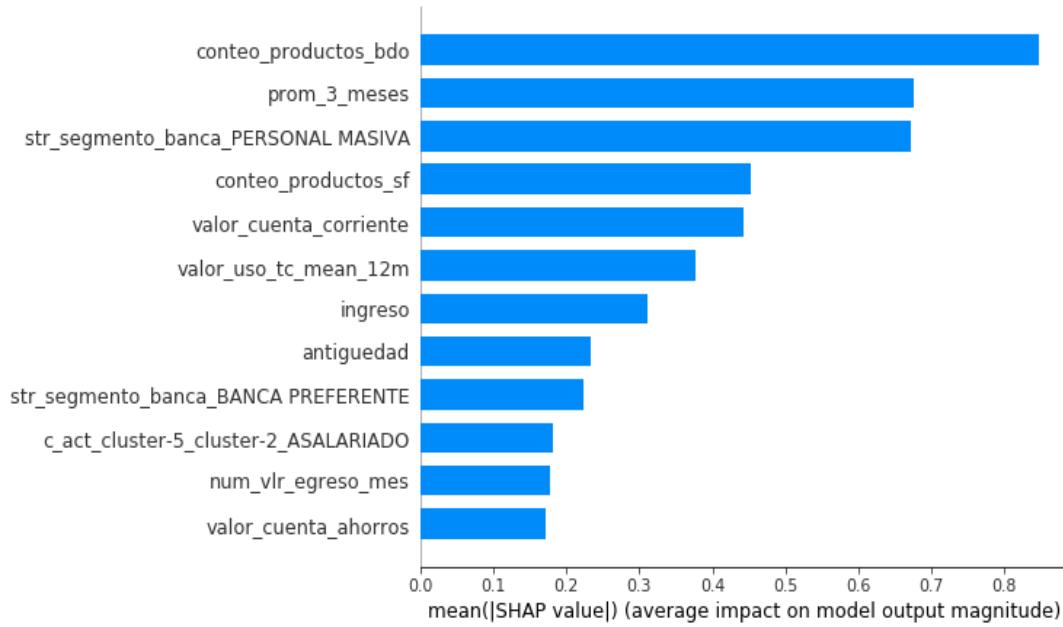
- SVM
- Árboles de decisión
- Random forest
- Regresión logística
- Regresión lineal
- Xgboost *

- Algoritmo estrella
- Tolerante a datos faltantes
- Velocidad tiempo de estimación óptima
- Presenta muy buen desempeño para data estructurada y desbalanceada



Interpretabilidad

SHAP (SHapley Additive exPlanations). SHAP asigna a cada variable un valor de importancia para una predicción particular.



Cómo medir un modelo de machine learning: clasificación

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Matriz de confusión:

Determinar el número total de individuos clasificados correcta e incorrectamente.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Accuracy:

Porcentaje de clientes que el modelo ha logrado clasificar correctamente

$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{True Negative} + \text{False Positive} + \text{False Negative} + \text{True Positive}}$$

Cómo medir un modelo de machine learning: clasificación

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision:

De todos los clientes que el modelo pronosticó que iban a cancelar sus productos, cuál es el porcentaje de estos que el modelo pronosticó correctamente.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Cómo medir un modelo de machine learning: clasificación

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

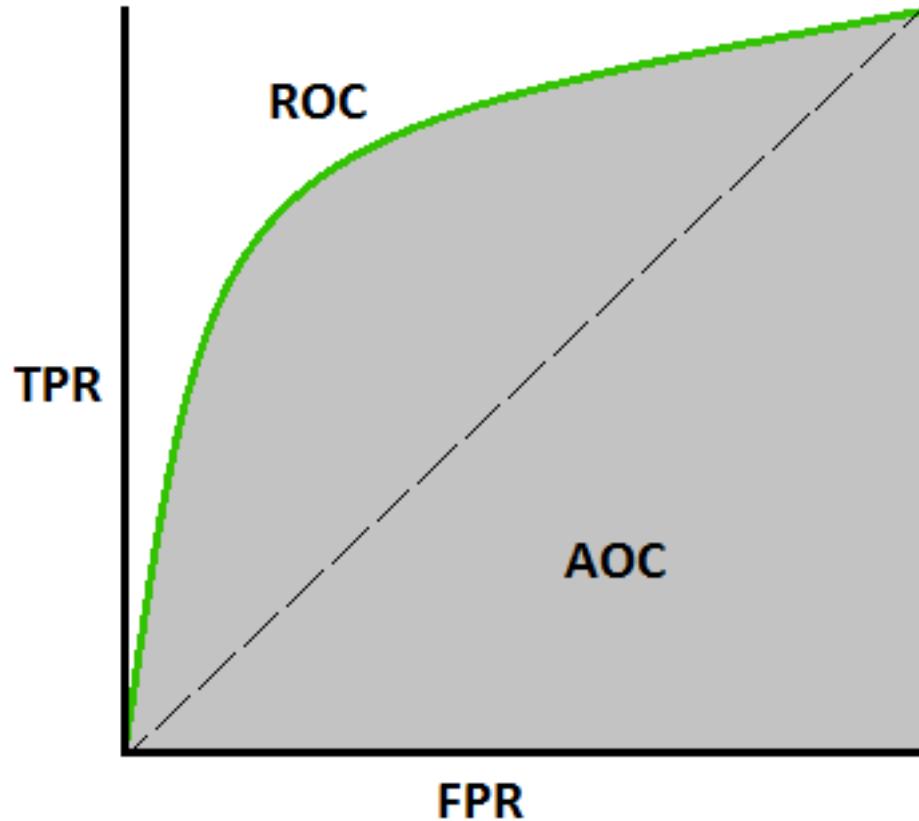
		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Recall:

Del total de clientes que cancelaron sus productos, cuál es el porcentaje de estos que el modelo ha encontrado.

$$\text{Recall} = \frac{\text{True Positive}}{\text{False Negativa} + \text{True Positive}}$$

Curvas ROC y AUC



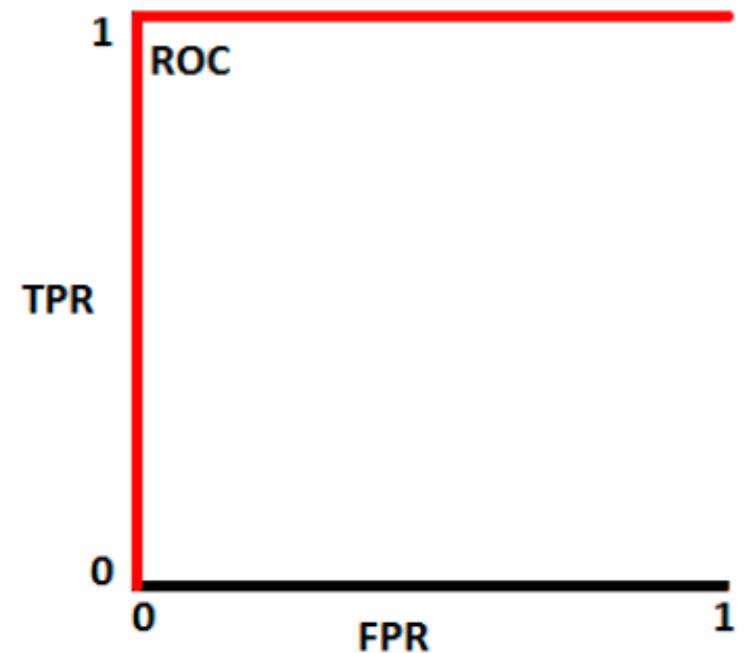
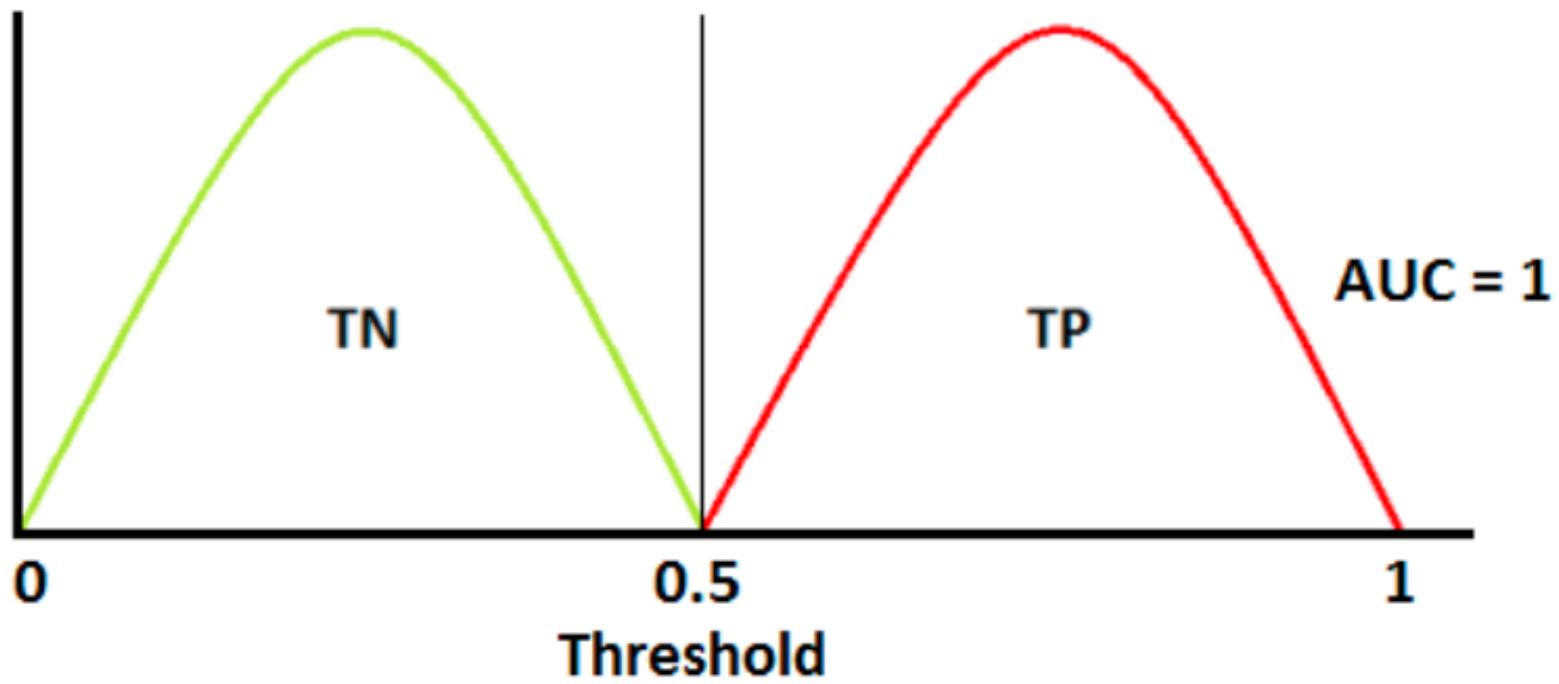
La curva ROC representa una relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) a diferentes puntos de corte.

$$TPR = Recall = \frac{\text{True Positive}}{\text{False Negativa} + \text{True Positive}}$$

$$FPR = 1 - \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}}$$

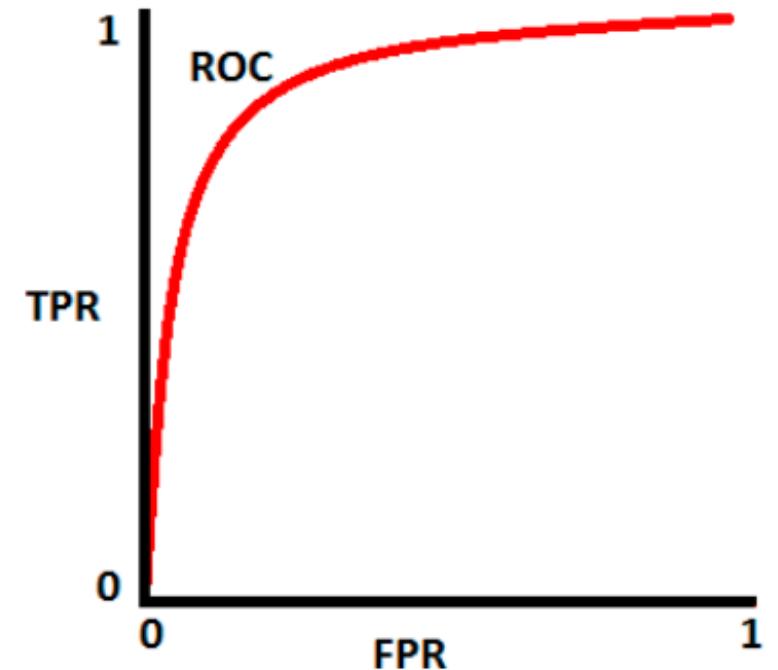
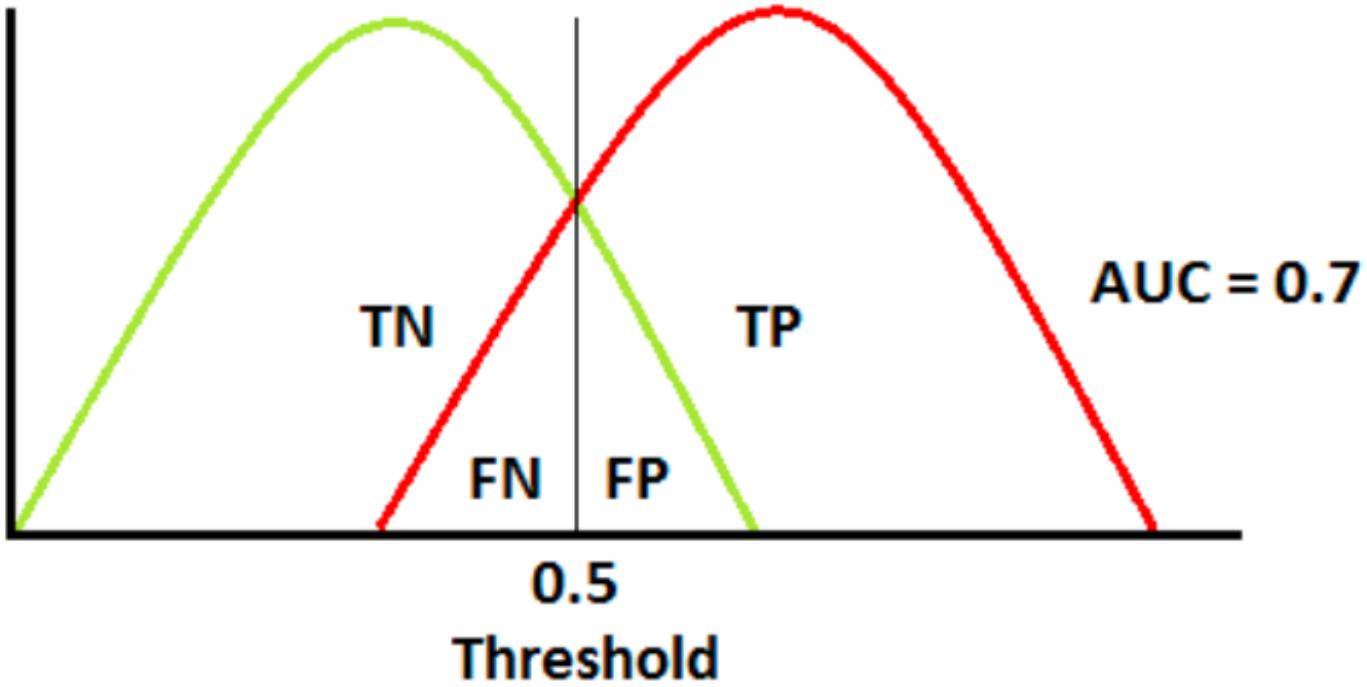
Note que el FPR es $1 - \text{Recall}$ (clase negativa)

¿Cuándo un modelo es perfecto?



El modelo es perfecto cuando AUC es igual a 1, si esto sucede, **sospeche!**

¿Cuál es el ideal de un modelo?



[Image 8 and 9] (Image courtesy: [My Photoshopped Collection](#))

Un buen modelo presenta un AUC mayor o igual a 0.7.



¿Dónde está Javier?

Esta es la capacidad que hoy tenemos para identificar las fugas de gas donde van a suceder o quien no va a pagar las cuotas de su cupo de crédito





Esta es la capacidad que tenemos con Analítica Avanzada con un modelo supervisado

Los dos conceptos **más importantes** cuando se desarrolla un modelo analítico

- En la estimación de un modelo el aspecto más importante a tener en cuenta es aprender las características esenciales de los datos y **no** una representación exacta de los datos.
- Lo más importante en un modelo es garantizar la capacidad de generalización.

Consecuencias de aprender una representación exacta de los datos



Esta casa se adapta perfectamente al tamaño del perro, sin embargo, esta no funciona con otros perros. Lo mismo sucede con los datos!

¿Como controlar la complejidad de los modelos y garantizar la capacidad de generalización?

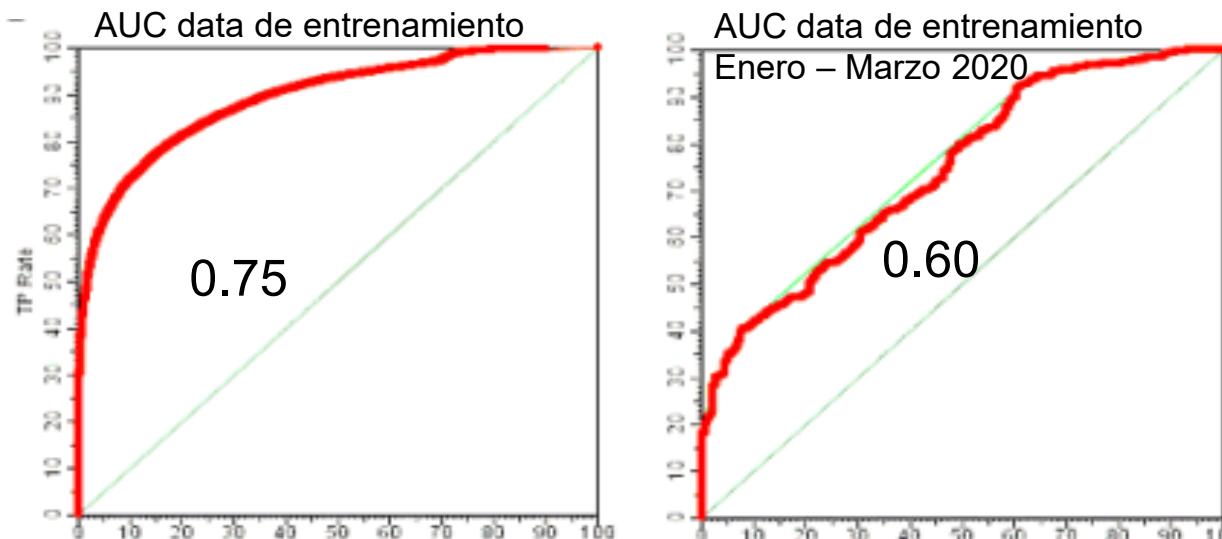
1. A partir de un conjunto de datos observados (donde se conoce su variable objetivo), dividir los datos en tres subconjuntos, denominados entrenamiento, validación y test.



¿En qué momento se debe cambiar el modelo?

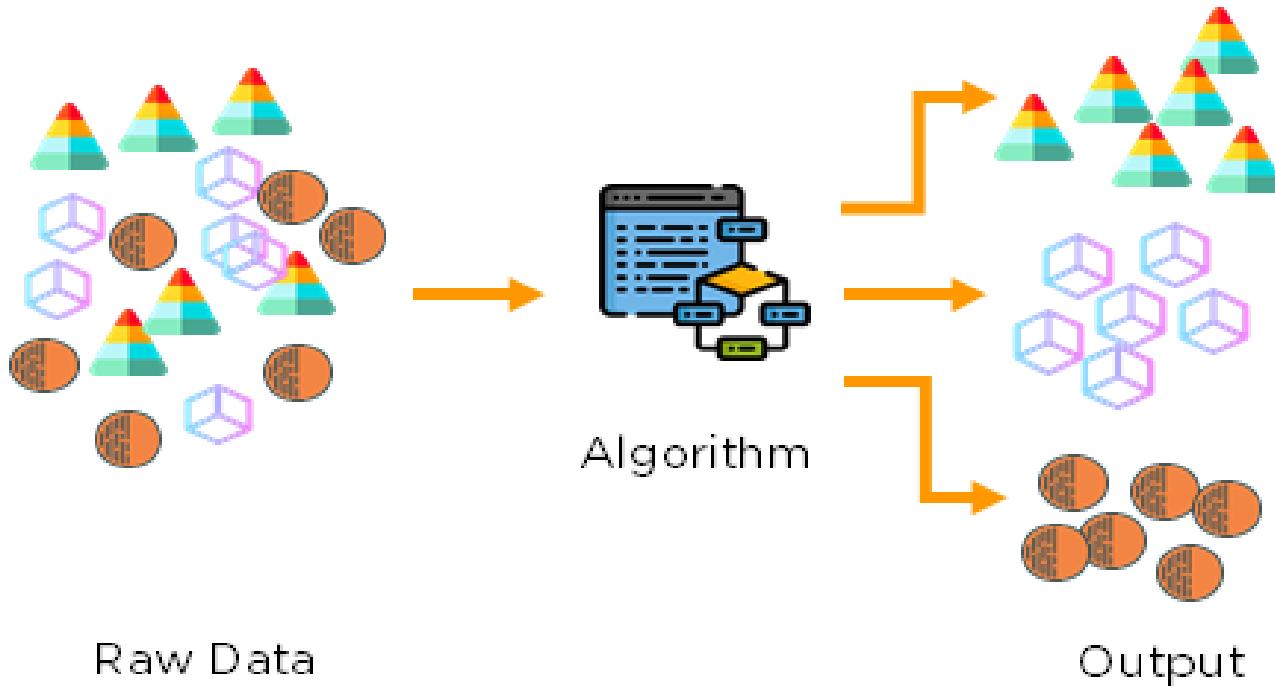
Backtesting

- Suponga que se ha construido un modelo para identificar los clientes más propensos a cancelar sus productos en los próximos **dos** meses.
- El modelo se construyó con información de 2019 y se tiene un AUC determinado.
- Se aplica el modelo a Enero de 2020.
- Con información de Marzo de 2020 se determina qué clientes(activos en Enero de 2020) cancelaron sus productos
- Se compara los resultados del modelo aplicado a los datos de Enero de 2020 con las cancelaciones reales observadas en Marzo de 2020 a través del AUC



Si el AUC ha disminuido bastante, entonces, puede ser una alerta para reentrenar o cambiar el modelo.

Aprendizaje no supervisado: clases de algoritmos



- No tenemos un label o variable objetivo que busquemos predecir o clasificar.
- El **objetivo** es **encontrar patrones** o descubrir la **estructura subyacente** de un conjunto de datos.

Clases de algoritmos:

- **Clustering:** encuentra grupos de individuos con características similares.
- **Detección de anomalías:** identifica aquellos individuos con un comportamiento atípico.
- **Reducción de dimensionalidad:** permite explicar y correlacionar la información en un conjunto de datos con la menor cantidad de variables posible.

Aprendizaje no supervisado: clases de algoritmos

Clustering: encuentra grupos de individuos con características similares.

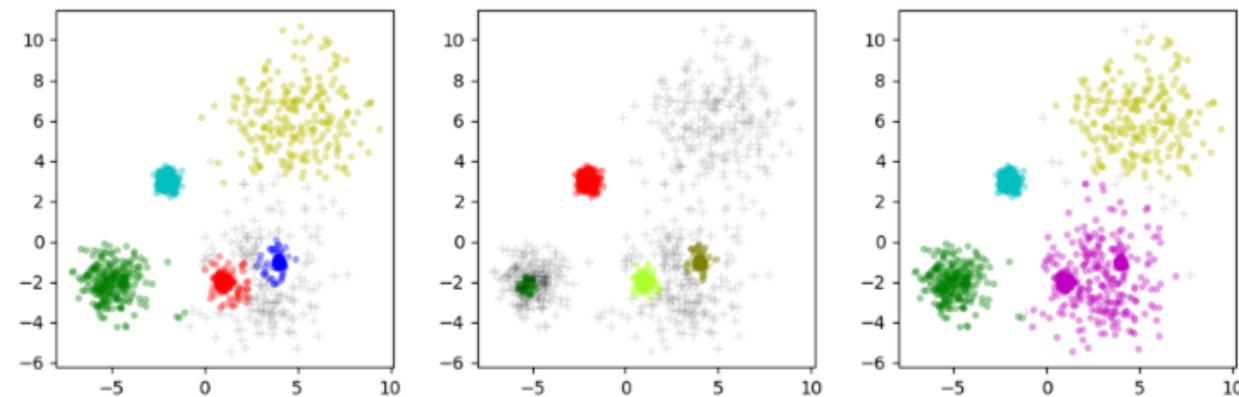
Los grupos o clusters deben ser homogéneos dentro y heterogéneos entre ellos

Algunos algoritmos para realizar *clustering* son:

- K-means
- K-modes
- Gaussian Mixture Models
- *Agrupación jerárquica*
- Affinity propagation
- DBSCAN
- Spectral Cluster

Características y aplicaciones:

- Todas las variables deben ser numéricas.
- Seleccionar el número de clusters respecto al **sentido de negocio** de los mismos y su significancia estadística.
- aplicaciones en valoración de activos, segmentación de clientes, clasificación de especies.



Aprendizaje no supervisado: clases de algoritmos

Detección de anomalías:

identifica aquellos individuos con un comportamiento atípico.

Se debe especificar la cantidad de datos *atípicos* que queremos identificar.

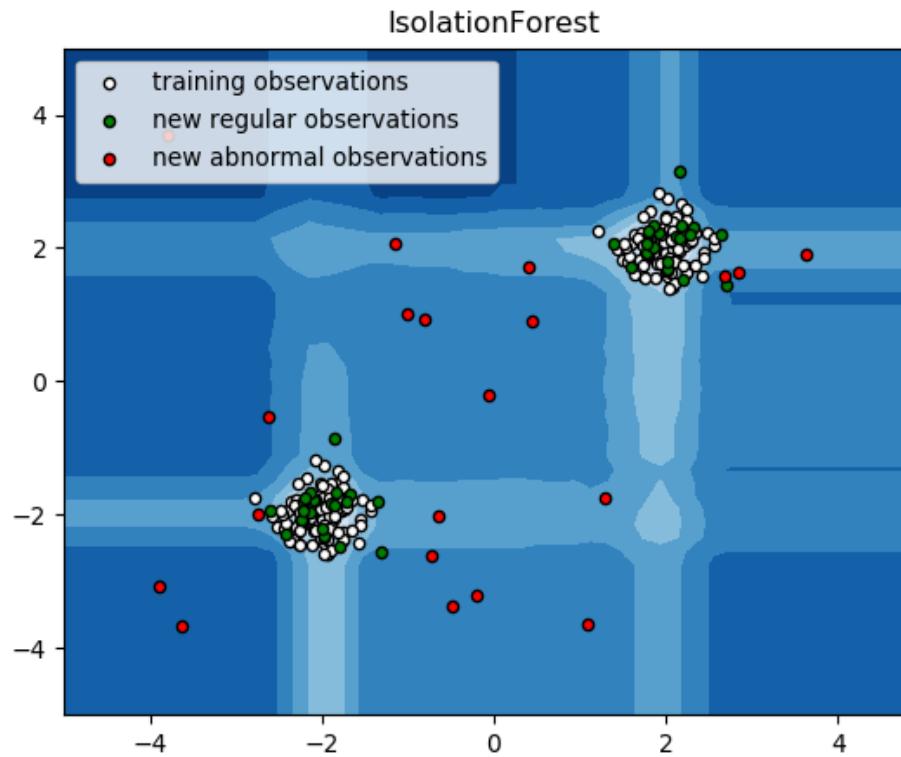
Algunos algoritmos para realizar *clustering* son:

- Isolation Forest.
- Local Outlier Factor.

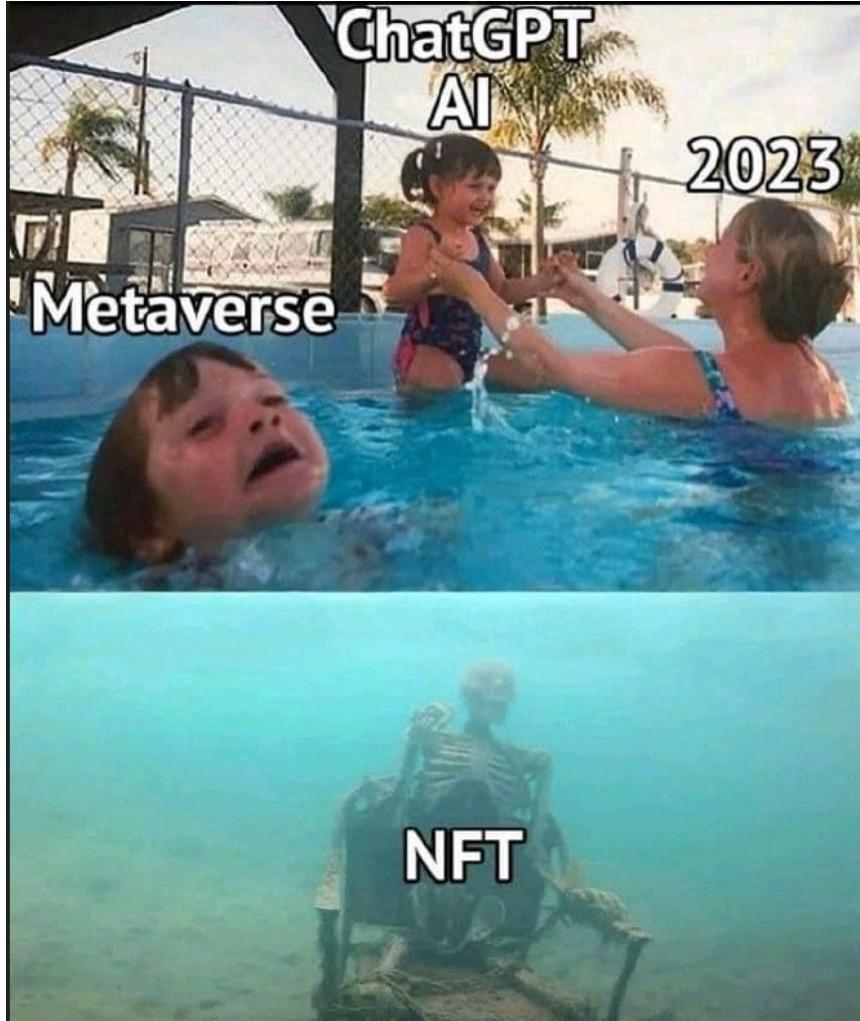
Características y aplicaciones:

SVM de una clase.

- Todas las variables deben ser numéricas.
- algunas aplicaciones se pueden encontrar en: seguros, detección de fraude, entre otros.



¿Fue Chatgpt el primer LLM en el mundo?



```
Welcome to
EEEEE LL IIII ZZZZZZ AAAAA
EE LL II ZZ AA AA
EEEEE LL II ZZZ AAAAAAA
EE LL II ZZ AA AA
EEEEE LLLLLL IIII ZZZZZZ AA AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

1964 – 1964 – LLM entrenado en MIT

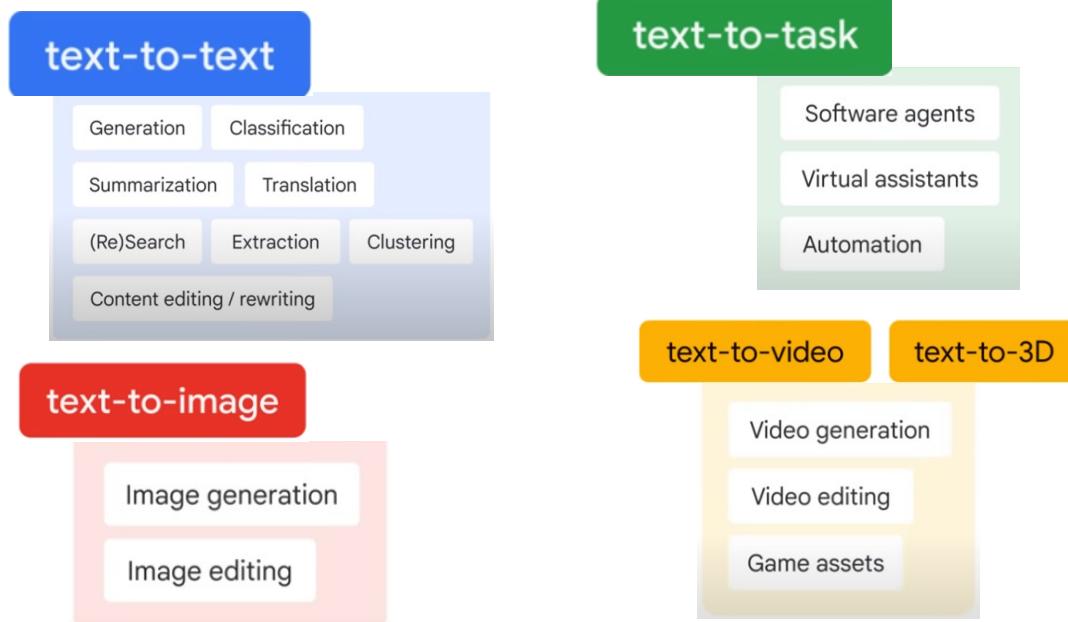
Definiciones y Conceptos Relevantes

IA Generativa es un tipo de tecnología de la inteligencia artificial que puede producir varios tipos de contenido, incluyendo texto, imágenes, audio y datos sintéticos

Gen AI → FMs → LLM

- **Gen AI** es alimentada por modelos pre.entrenados con enormes cantidades de datos (Foundation Models)
- **Large Languages Models** son un subconjunto de los modelos fundacionales, son entrenados con trillones de palabras

Tipos de modelos



Deconstrucción del nombre LLM

- **Large**: Significa que son entrenados con enormes sets de datos y una gran cantidad de parámetros. (Amplia aplicabilidad)
- **Language**: Operan bajo lenguaje humano
- **Model**: Usados para encontrar patrones o hacer predicciones dentro de los datos

Los LLM modelos descifran las relaciones entre palabras, es por esto que entienden el lenguaje tan bien. Esto se ha dado gracias a desarrollos como:

- **Transformers**: Registra las relaciones dentro de secuencias de datos, de esa manera entiende el significado de los datos y puede transformarlos.
- **Bidirectional Encoding**: Una arquitectura de ML que usa los transformers y mira en ambas direcciones para entender el significado de lenguaje ambiguo
- **Autoregressive Models**: Usan las palabras previas para predecir la siguiente palabra en una secuencia.



<https://www.scalecapital.com/stories/generative-ai-landscape-q3-2024-insights>

The Generative AI Market Map v3



A work in progress



Source: Sequoia

Herramientas para estructurar un proyecto de Analítica

Nine boxes

Artefacto que a través de 9 temas a resolver con el equipo interdisciplinario

- | | | |
|--|--|---|
| <p>1. Problema empresarial
¿Cuál es la métrica de rendimiento empresarial que se va a cambiar? (por ejemplo, tiempo, costo, calidad). ¿Qué aspecto tiene el éxito?</p> | <p>2. Valor en juego
¿Qué controladores e hipótesis están disponibles para cambiar este rendimiento? ¿Cuánto creemos que puede cambiar (% de mejora)?</p> | <p>3. Enfoque analítico propuesto
¿Puede traducirse el problema de la empresa en un problema analítico? ¿Explicativa o predictiva? ¿Potenciales variables de objetivo y métodos analíticos a utilizar? ¿Métodos de I&D?</p> |
| <p>4. Paisaje de datos y viabilidad de datos
¿Qué recursos están disponibles? Estos datos son necesarios para respaldar el enfoque analítico? ¿Estos datos conllevan algún riesgo o tienen problemas de calidad conocidos?</p> | <p>5. Intervenciones anticipadas
¿Cómo se usará la salida de análisis para impulsar el cambio? ¿Cuáles son las intervenciones probables? ¿Cambios en el proceso de negocio o herramientas interactivas? ¿Quiénes son los usuarios?</p> | <p>6. Riesgo
Posibles cosas que podrían afectar el proyecto</p> |
| <p>7. Abordaje técnico
Describe la solución técnica preferible de arquitectura Ambiente de alojamiento, infraestructura, herramientas</p> | <p>8. Alcance, restricciones e hitos
Áreas prioritarias (por ejemplo, geografía, unidad de negocio); fechas y dependencias clave a tener en cuenta.</p> | <p>9. Viabilidad de enfoque
¿Se ha hecho esto antes? Casos de uso previos y líderes del proyecto. Evidencia de éxito y oportunidad potencial</p> |

5w2h

Las 5W representan lo que sería en inglés: «Qué», «Por qué», «Quién», «Cuándo» y «Dónde», mientras que las 2H representan «Cómo» y «Cuánto».

- What: ¿Qué se hará?
- Why: ¿Por qué se hará?
- Who: ¿Quienes lo harán?
- Where: ¿Dónde se hará?
- When: ¿Cuándo se hará?
- How: ¿Cómo se hará?
- How much: ¿cuánto costará?

Gracias por su asistencia!