

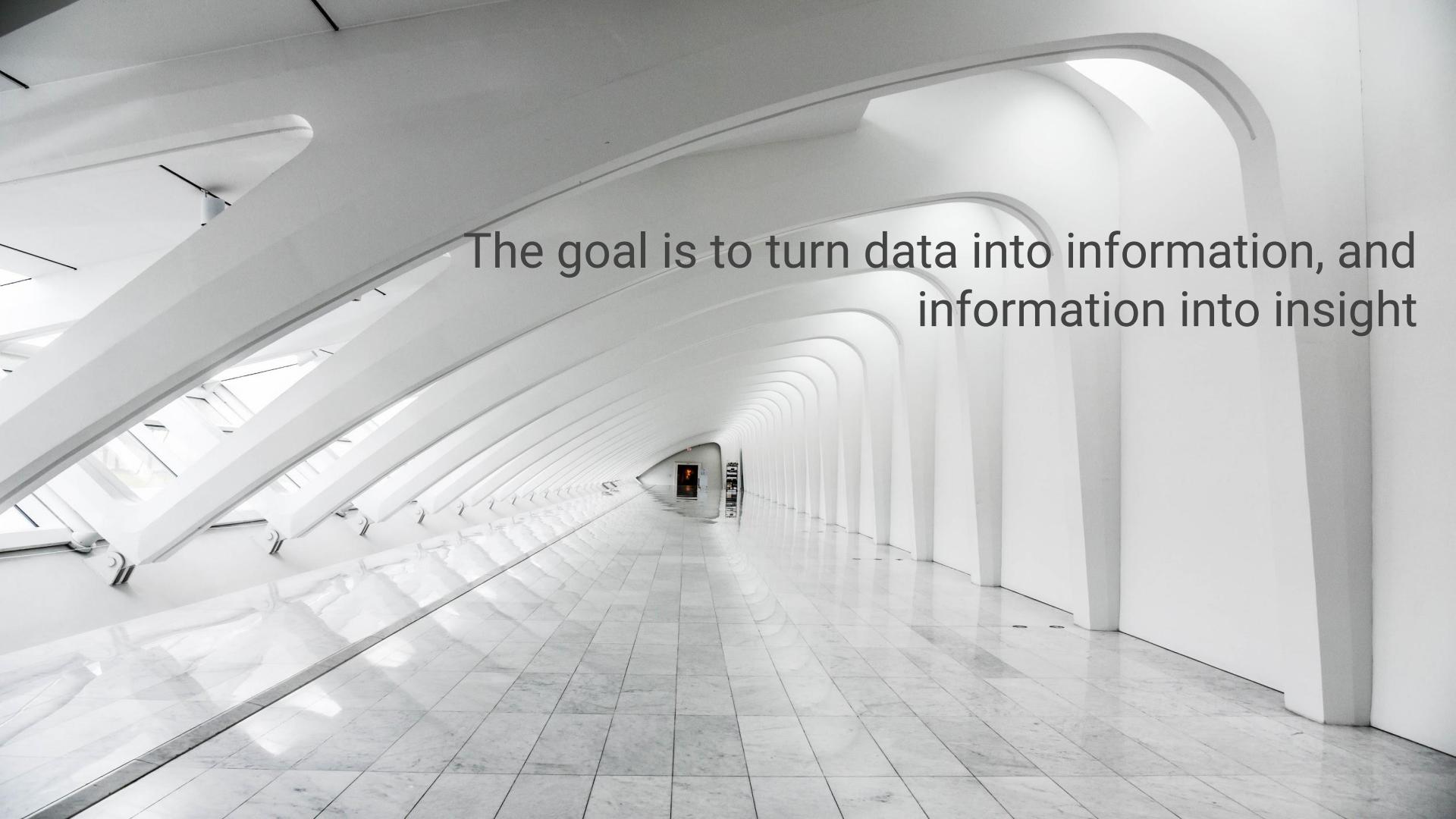
Empleos del Futuro





INTRODUCCIÓN A LA CIENCIA DE DATOS

Leonardo Ignacio Córdoba

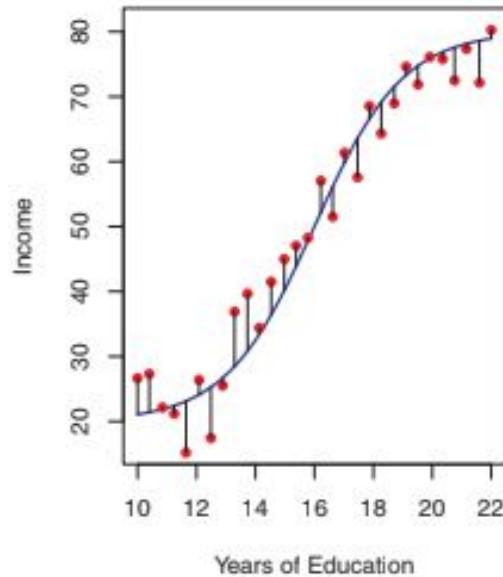
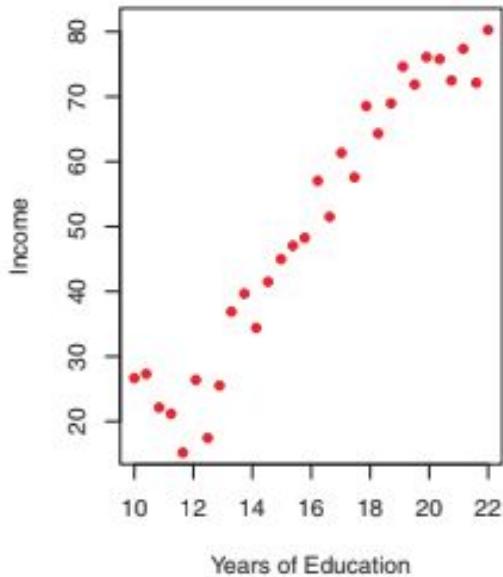
A black and white photograph of a modern architectural space. The ceiling and walls are white and curved. On the left, there is a glass partition with a metal frame. The floor is made of large, light-colored tiles in a grid pattern. The perspective leads to a small, dark opening at the end of the corridor.

The goal is to turn data into information, and
information into insight

Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- **Predecir:** en ciertos casos tenemos acceso a la información de entrada (X) pero no al target o output (Y).

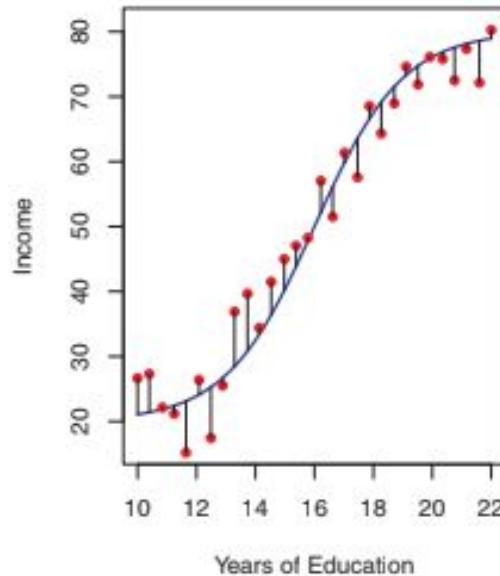
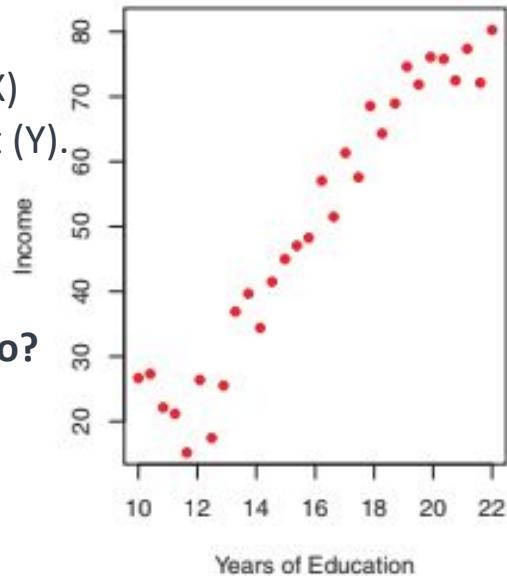


Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- **Predecir:** en ciertos casos tenemos acceso a la información de entrada (X) pero no al target o output (Y).

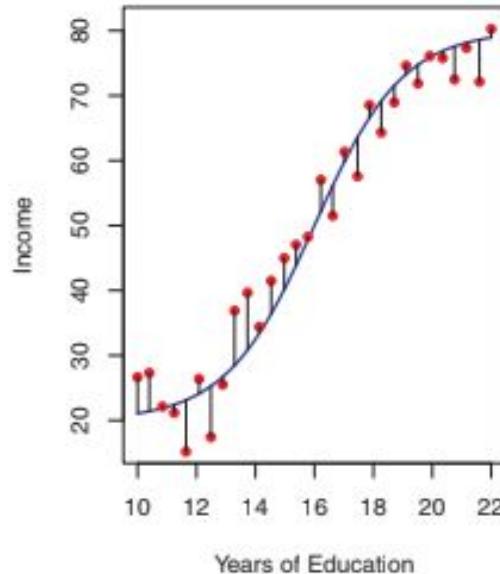
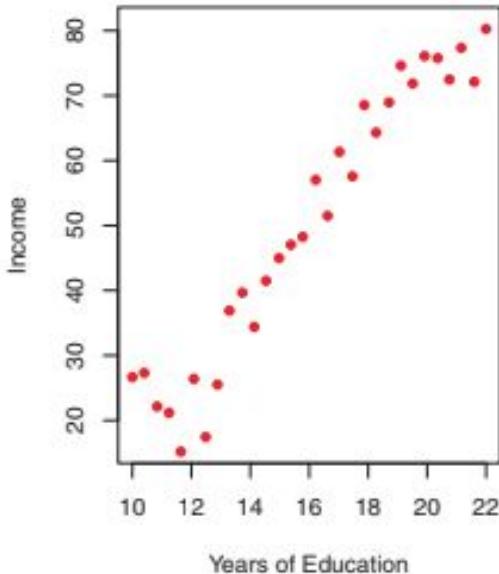
¿Pueden mencionar un ejemplo?



Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- **Entender la relación entre X e Y:** en ciertas ocasiones nuestro principal interés no es predecir sino entender cómo X afecta a Y
- **Entender si existen patrones en nuestros datos:** especialmente cuando la estadística descriptiva básica es insuficiente

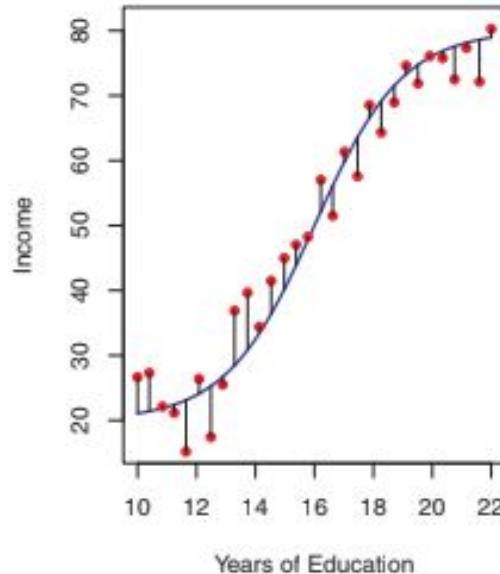
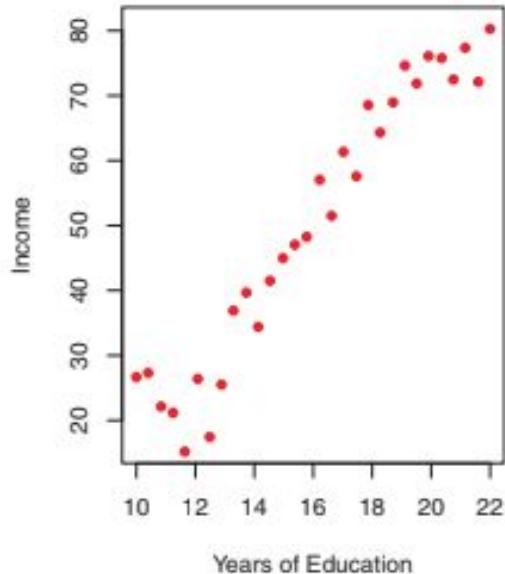


Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- Entender la relación entre X e Y: en ciertas ocasiones nuestro principal interés no es predecir sino entender cómo X afecta a Y

¿Se les ocurre cuándo nos puede interesar este caso?

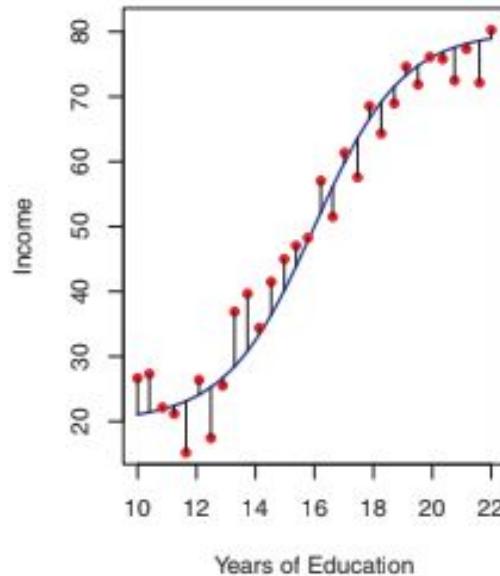
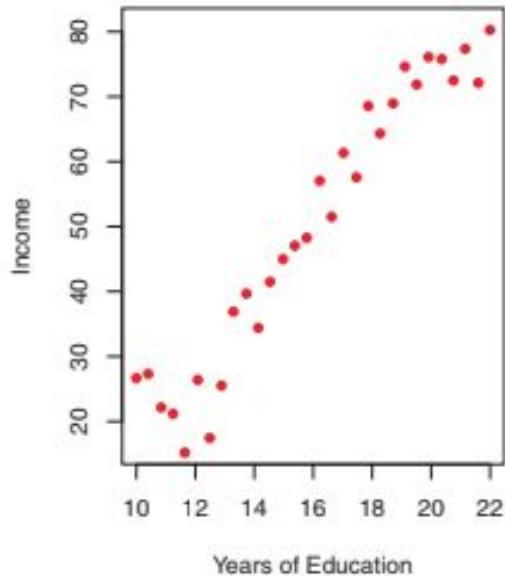


Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- Entender si existen patrones en nuestros datos: especialmente cuando la estadística descriptiva básica es insuficiente

¿Y ésto?

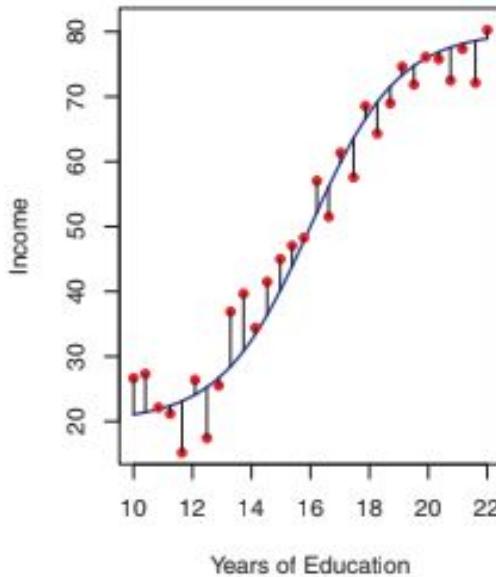
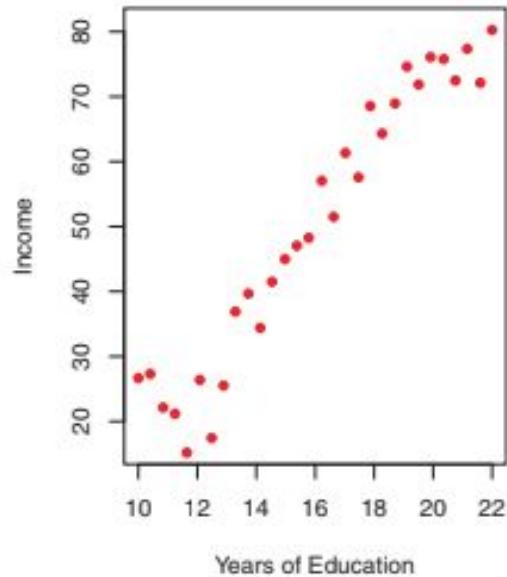


Introducción a la ciencia de datos

¿Ciencia de datos para qué?

1) Predecir

2) Entender



Introducción a la ciencia de datos

¿Ciencia de datos para qué?

- 1) Predecir
- 2) Entender

NO SON LO MISMO

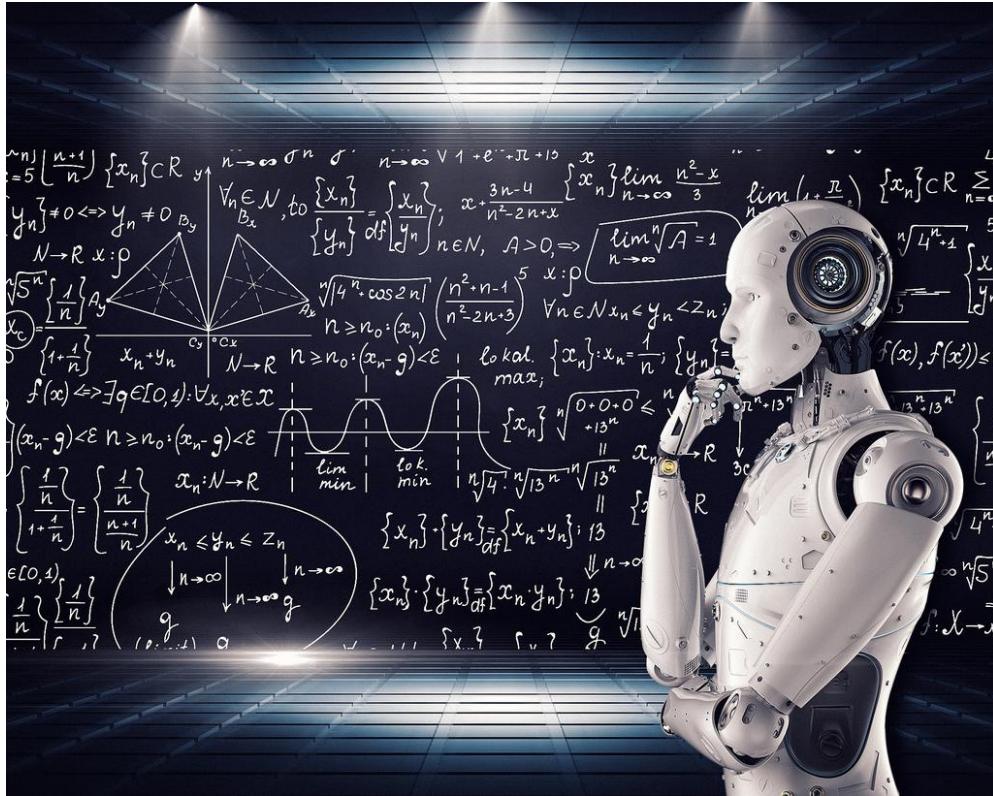


SIMILAR IS NOT THE SAME



Introducción a la ciencia de datos

¿Machine Learning?

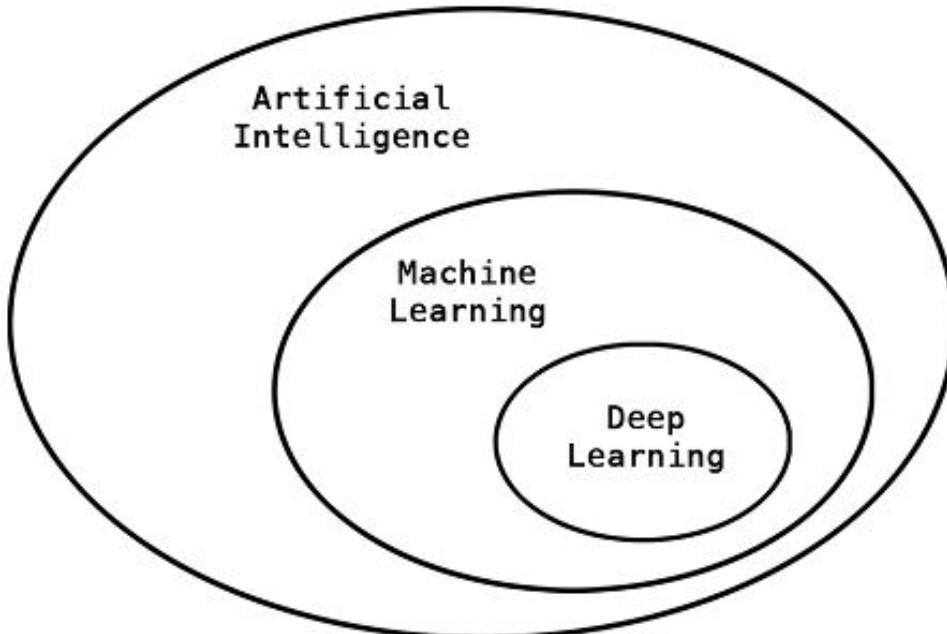


Decimos que un programa “aprende” si para cierta tarea, su performance mejora cuando crece la experiencia.

T. Mitchell

Introducción a la ciencia de datos

Machine Learning



- La “**Inteligencia Artificial Simbólica**” dominó el paradigma de IA desde 1950 hasta la parte de los ‘80. Su pico de popularidad: boom de los “**sistemas expertos**”.

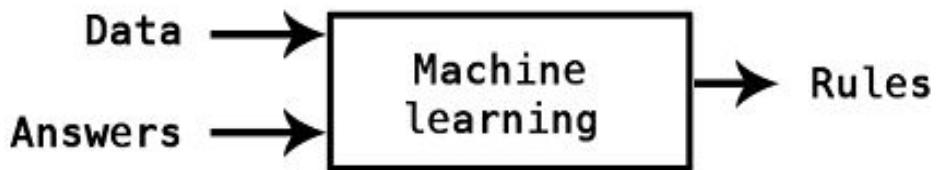
Introducción a la ciencia de datos

Machine Learning

Programación clásica (el paradigma de la IA simbólica): los humanos ingresan reglas (un programa), datos a ser procesados según esas reglas, y de este proceso resultan las respuestas esperadas.

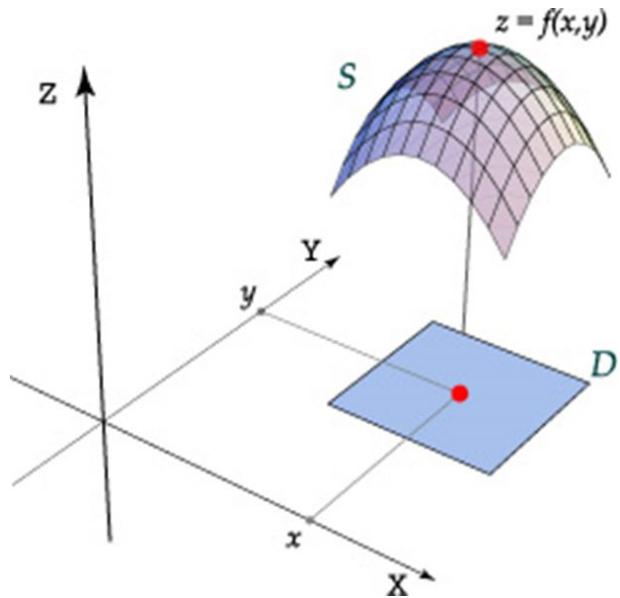


Machine Learning: los humanos ingresan datos como input además de las respuestas esperadas, y las reglas surgen como output. Estas reglas pueden luego ser aplicadas a nuevos datos para producir respuestas originales.



Introducción a la ciencia de datos

Machine Learning



- Podemos entender a este campo como un conjunto de *conocimientos* en donde lo que se busca es *automatizar* dos tipos de tareas:
 - El descubrimiento de **patrones** en la información.
 - El **mapeo** de observaciones a valores, es decir, a las observaciones asignarles un valor numérico. Típicamente, en estos casos hablamos de problemas de predicción y de clasificación.

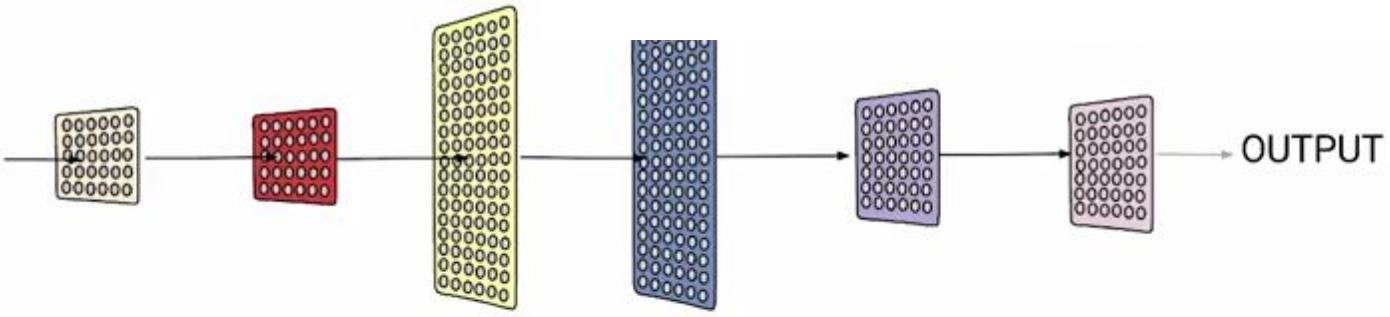
Introducción a la ciencia de datos

Machine Learning

Target



En clasificación, el proceso de ajuste consiste en hacer modificaciones a la función del modelo de forma tal de que, para cada input, el output se acerque al target correspondiente (resultado esperado).



Un modelo de ML es
una **función**
matemática

Introducción a la ciencia de datos

Machine Learning

En nuestro caso la experiencia es un conjunto de **observaciones** con **atributos**. En algunos casos, estas observaciones tienen una variable **target** que queremos predecir. El **target** es optativo.

Para que los modelos de **Machine Learning** aprendan es necesario que la información esté organizada de manera matricial como a continuación:

Diagrama que ilustra la estructura de datos para Machine Learning:

Las columnas representan los **Atributos**: user_id, mes, page_views, tiempo, compras previas.

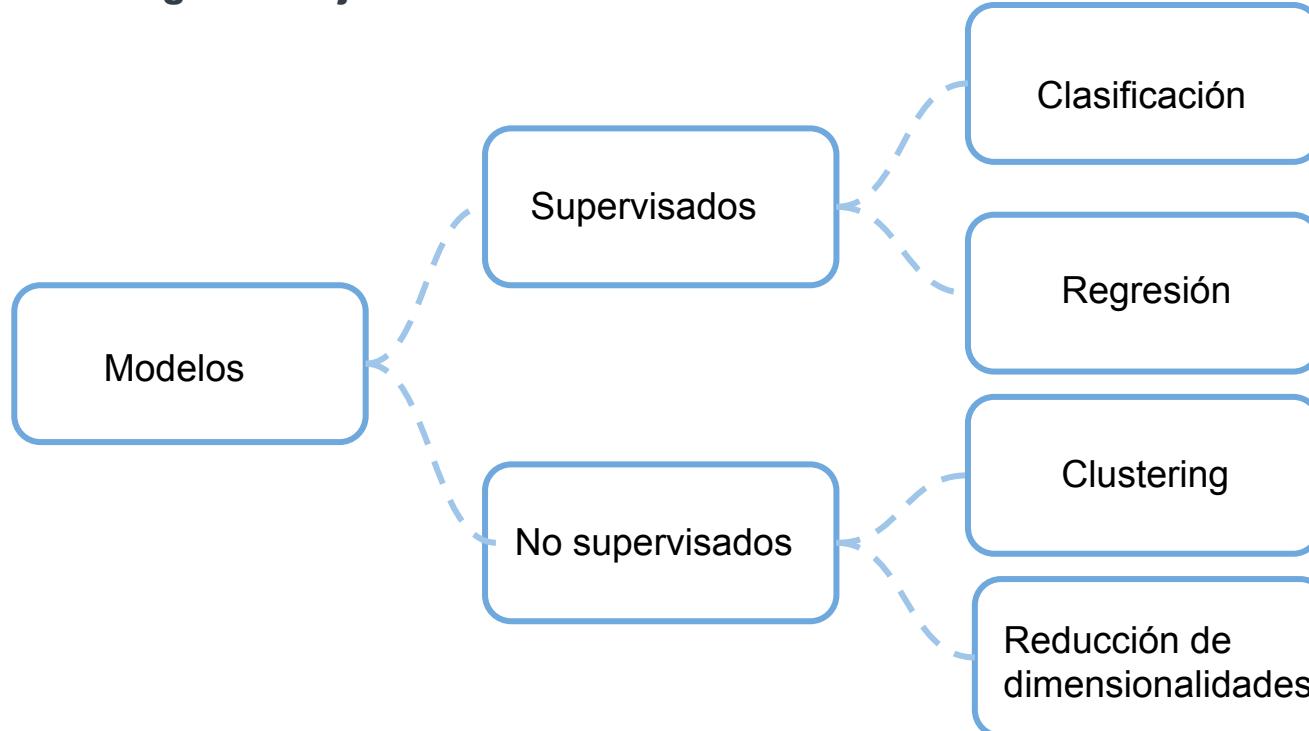
La primera columna es el **Target**, que en este caso es "upsell".

Las filas representan las **Observación**s.

upsell	user_id	mes	page_views	tiempo	compras previas
1	144332	5	23	15	3
0	634631	5	14	10	1
0	123126	5	10	8	0

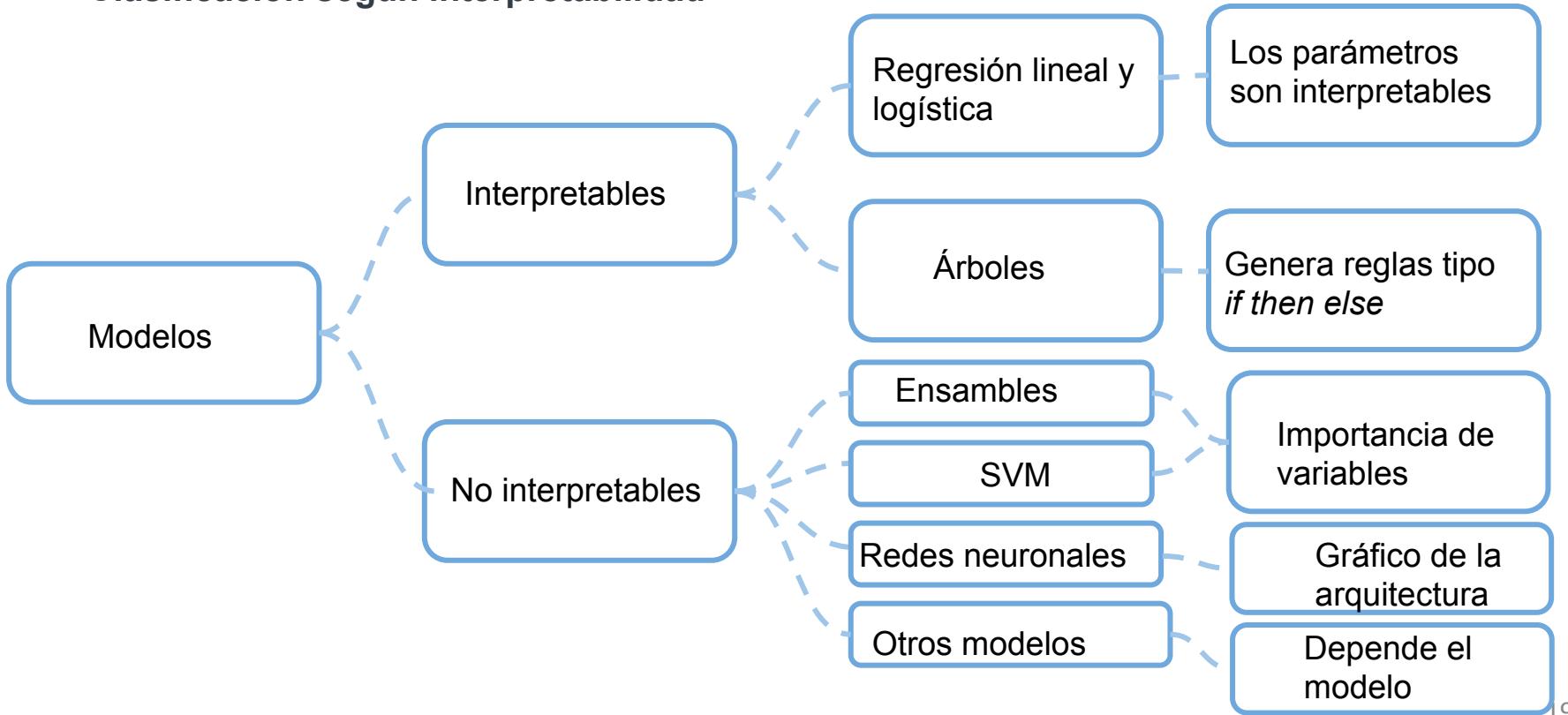
Modelos supervisados y no supervisados

Clasificación según el objetivo



Modelos supervisados

Clasificación según interpretabilidad



Casos de negocio



Casos de negocio: problemas supervisados

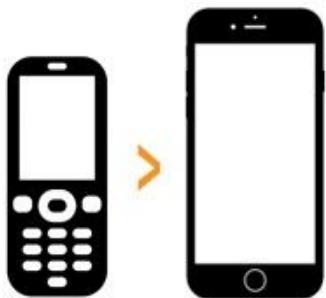
Predicción de propensión a conversión online



- **Objetivo:** predecir qué tan probable es que un visitante a una app realice una conversión. Ésta puede ser una compra online, un clic, completar un formulario, etc.
- Esto se usa para decidir cuánto pagar en una subasta, asignar un precio o producto dinámicamente, etc.

Casos de negocio: problemas supervisados

Modelos de upselling/cross-selling



Upselling

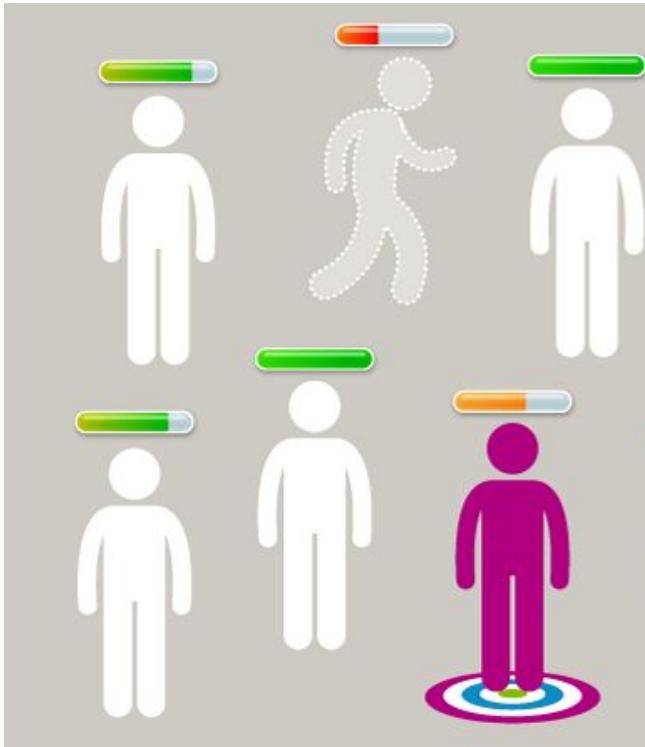


Cross-selling

- **Objetivo:** predecir la probabilidad de que un cliente compre un producto más caro (upselling) o complementario (cross-selling).
- Esto se puede integrar a un CRM o una herramienta de marketing para realizar campañas automáticamente.

Casos de negocio: problemas supervisados

Predicción de churn



- **Objetivo:** predecir la probabilidad de que un cliente se dé de baja en determinado período.
- Sabiendo quiénes son los más propensos podemos generar un incentivo para impedirlo.

Casos de negocio: problemas supervisados

Predicción de fraude



- **Objetivo:** predecir la probabilidad de que una transacción sea fraudulenta

Casos de negocio: problemas supervisados

Sistemas de recomendación



- **Objetivo:** optimizar qué productos ofrecer en una plataforma de ventas online



Casos de negocio: problemas supervisados

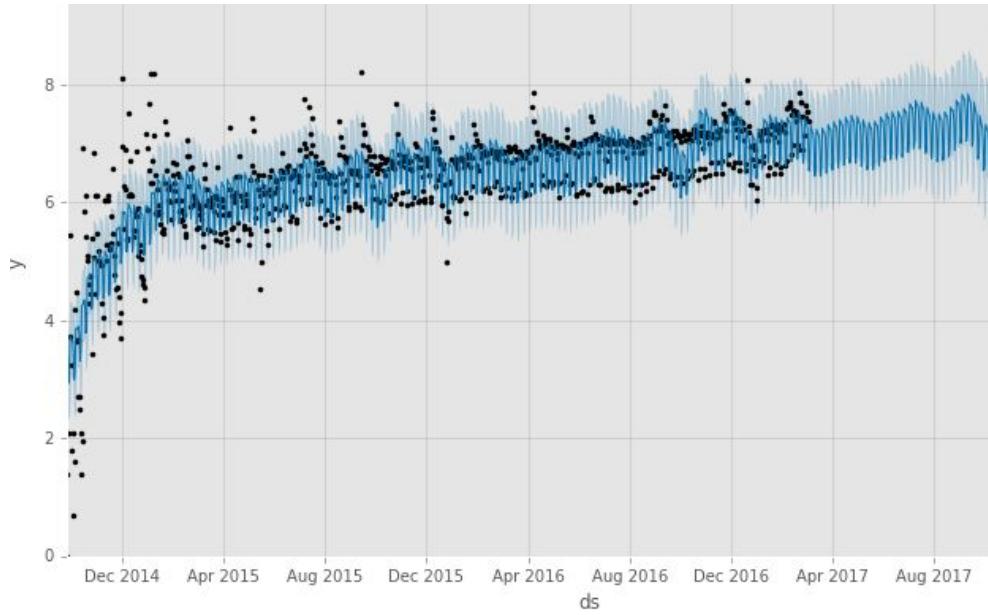
Marketing Mix Model



- **Objetivo:** optimización de la asignación de pauta entre los distintos canales publicitarios
- Conocer la importancia de los distintos canales y/o campañas
- Simular el impacto de cierto canal en las ventas o kpi

Casos de negocio: problemas supervisados

Forecasting



- **Objetivo:** conociendo una serie histórica de algunas pocas variables predecir el valor en el período siguiente.
- Ejemplos de forecasting pueden ser predicción de ventas, precio de activos financieros, despacho de cemento, etc.

Casos de negocio: problemas no supervisados

Segmentación automática

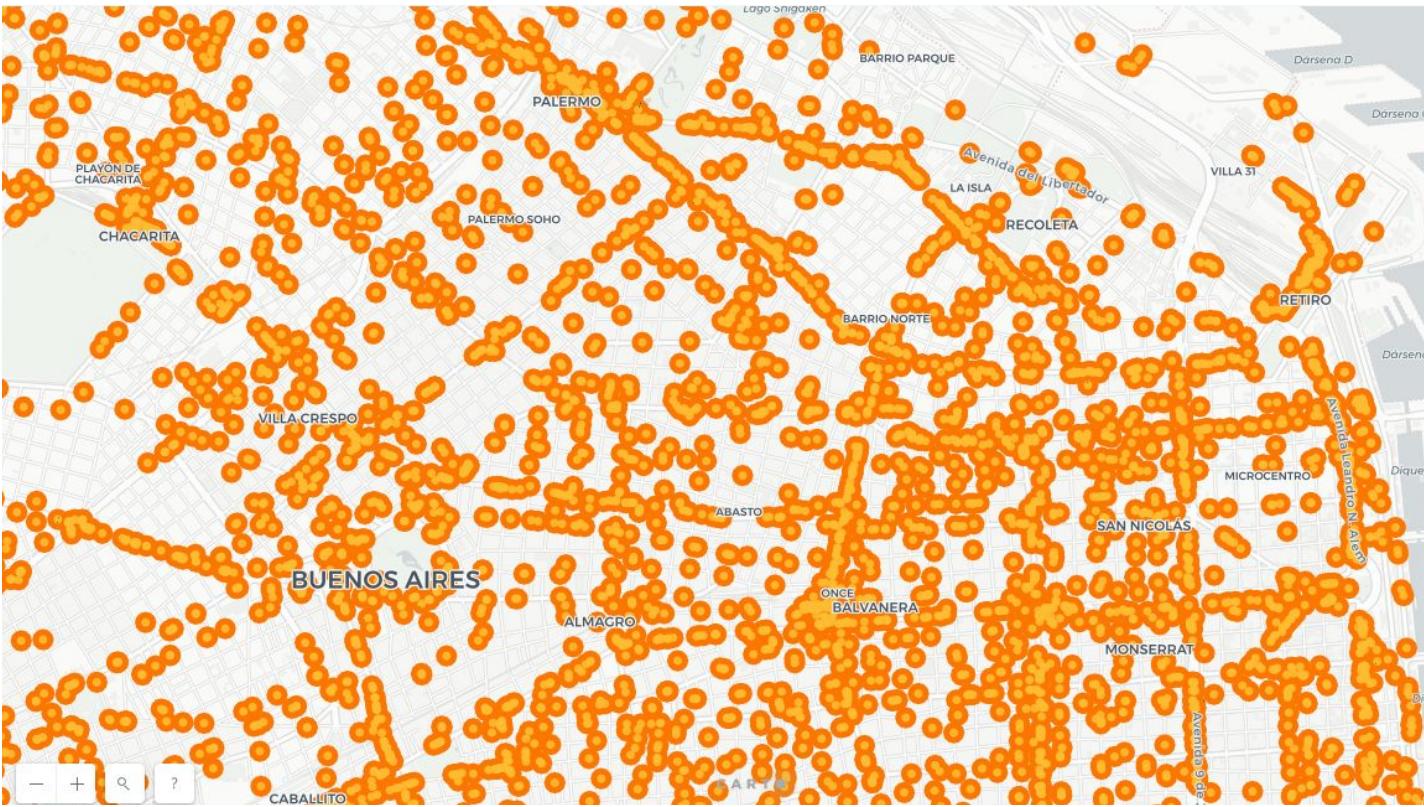
- Generación de segmentos a partir de información demográfica y de comportamiento.
- Con esta técnica se pueden generar K segmentos automáticamente, de modo de accionar de manera distinta sobre cada uno de ellos.



Casos de negocio: problemas no supervisados

Clustering geográfico I

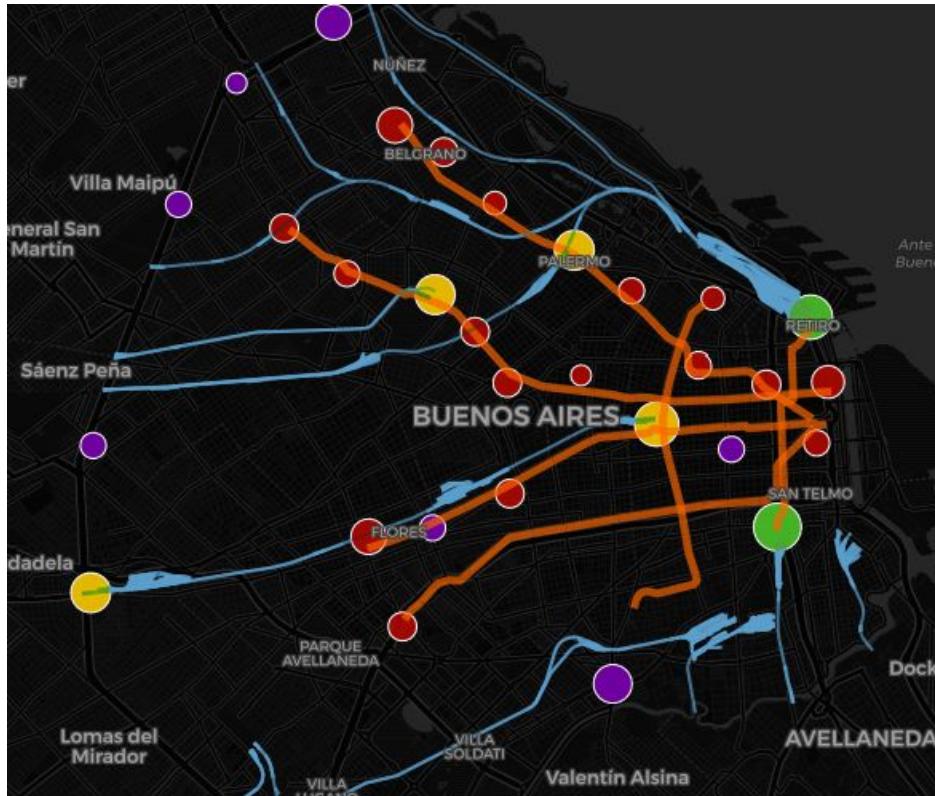
- Encontrar patrones geográficos automáticamente.
- Por ejemplo: paradas de colectivos, transacciones de clientes,etc.



Casos de negocio: problemas no supervisados

Clustering geográfico II

- En este caso, en vez de clusterizar puntos GPS, clusterizamos centros de transbordos para generar categorías



A black and white photograph of a modern building's facade. The facade features a complex grid of intersecting white lines, creating a sense of depth and perspective. The lines form a series of triangles and rectangles, some of which appear to be recessed panels or louvers. The overall effect is a high-contrast, geometric pattern against a dark background.

Regresión lineal

Regresión lineal

Marketing Mix Model

Imaginemos la siguiente situación:

Una empresa posee una presupuesto X para administrar su pauta presupuestaria entre distintos medios (TV, radio, diarios). Además, tiene los registros históricos de sus ventas y de sus inversiones.

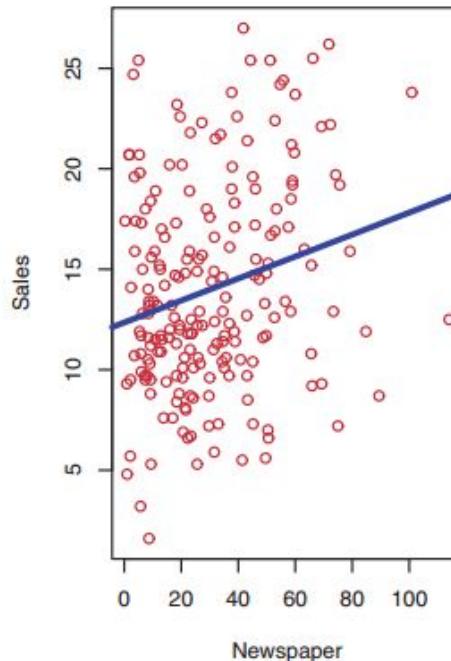
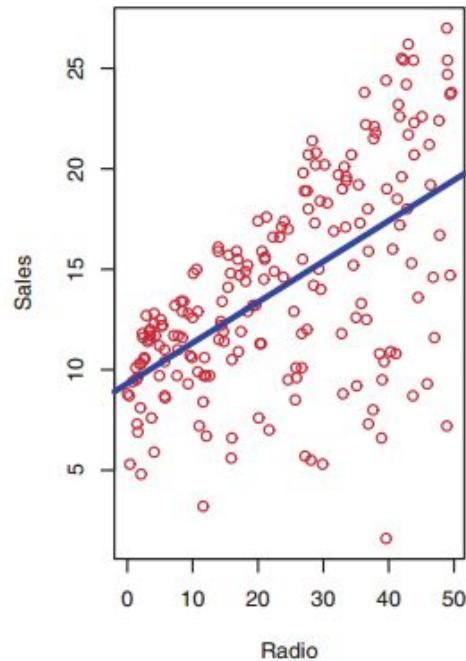
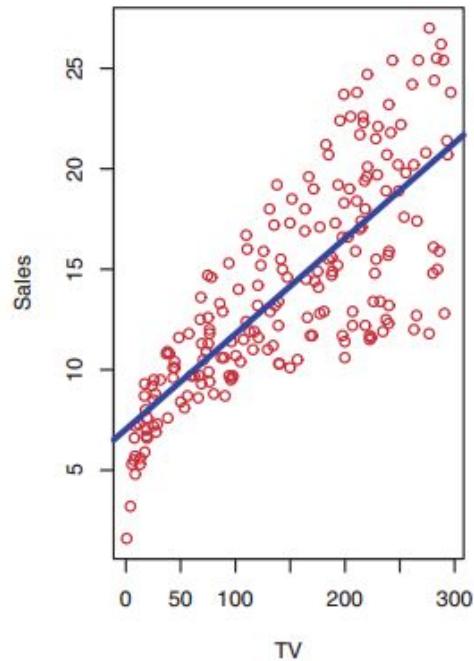
Con esta información... ¿podemos mejorar la distribución del presupuesto entre cada uno de los medios?

Con esta idea surge un modelo empleado por agencias publicitarias conocido como Marketing Mix Model.

“Mix” se refiere a que nos permite determinar la mejor combinación de medios.

Regresión lineal

Marketing Mix Model



Regresión lineal

Marketing Mix Model

¿Hay alguna **relación** entre el presupuesto en publicidad y las ventas?

¿Qué tan fuerte es esa relación?

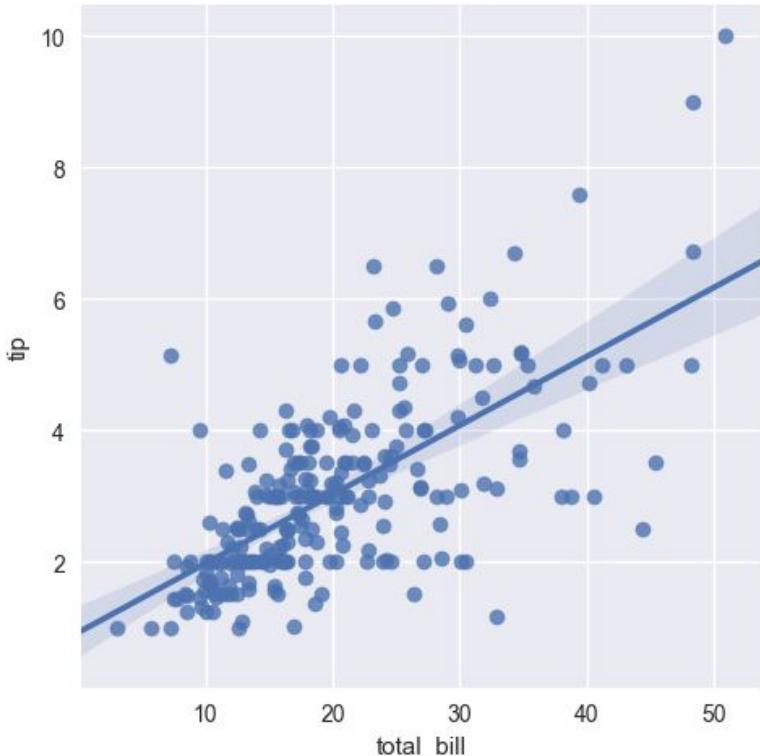
¿Cuáles de los medios mencionados contribuyen a las ventas?

¿Con cuánta **precisión** podemos predecir las ventas futuras?

¿Es esta **relación lineal**?

Modelo supervisado

Regresión lineal



$$y_i = \beta_0 + \beta_1 x_1 + \epsilon$$

La regresión lineal es uno de los modelos más sencillos y, por este motivo, ampliamente estudiado.

Si bien no tiene una gran capacidad predictiva los parámetros pueden ser interpretados.

¿Qué nos dicen los “betas”?

Modelo supervisado

Regresión lineal

β_0 y β_1 son dos constantes que representan la ordenada al origen y la pendiente en el modelo lineal. β_0 se conoce además como intercepto.

β_0 y β_1 son los parámetros del modelo.

Los parámetros deben ser calculados de modo que el modelo tenga la mayor calidad predictiva posible. Una vez hecho ésto podemos predecir futuras ventas en base a un valor particular de X1.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \text{ donde } \hat{\cdot} \text{ indica una predicción de Y en base a X = x.}$$

Modelo supervisado

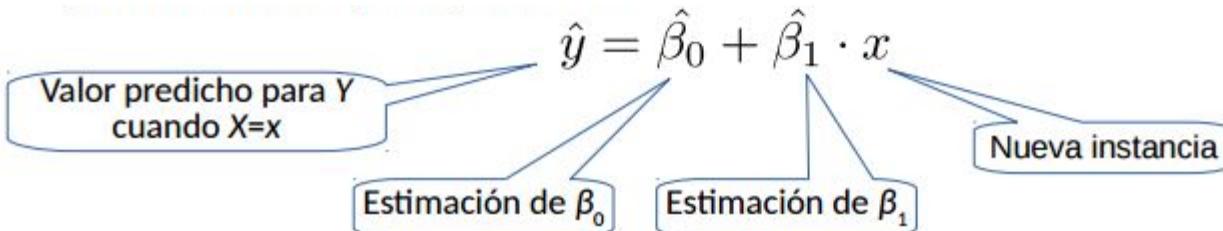
Regresión lineal

Entonces:

- Consiste en predecir una respuesta cuantitativa Y en base a una única variable predictora X.

$$Y \approx \beta_0 + \beta_1 \cdot X$$

- β_0 y β_1 son coeficientes desconocidos que vamos a estimar, o ajustar en base a los datos de entrenamiento. Una vez estimados, los podemos usar para predecir:



Modelo supervisado

Regresión lineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

En una regresión lineal generalmente uno tiende a incluir más de una variable independiente, ¿se imaginan por qué?

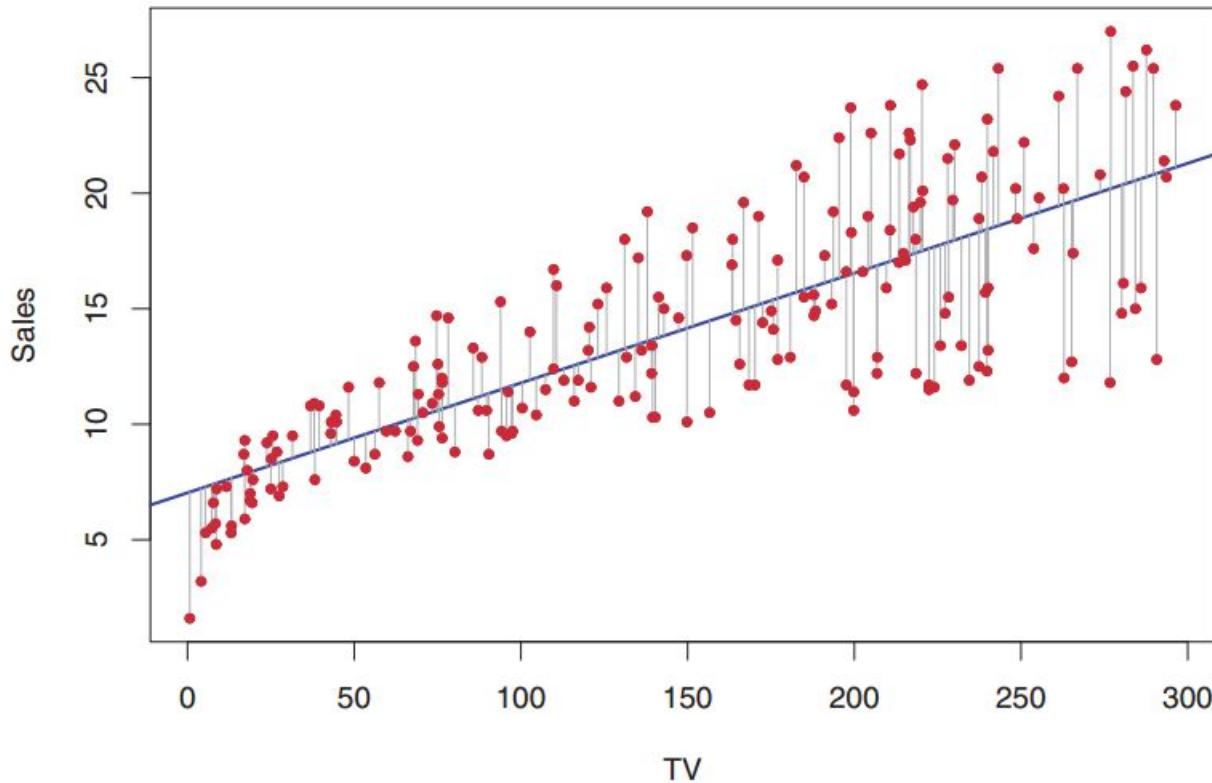
De hecho, en un Marketing Mix Model es común encontrar muchas otras variables... ¿cómo cuáles?

Modelo supervisado

Regresión lineal

En este gráfico podemos apreciar la distancia entre la predicción y el valor observado.

A la diferencia entre el valor observado y el valor predicho la conocemos como **residuo**.



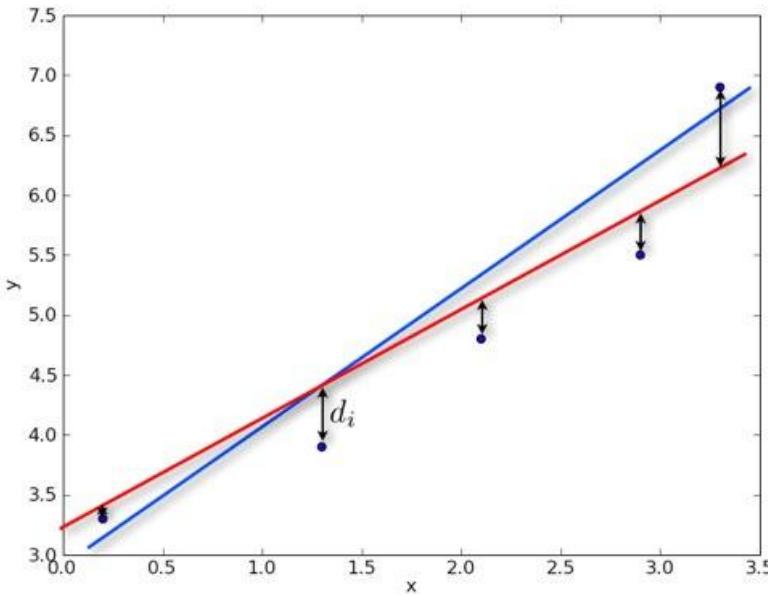
Modelo supervisado

Regresión lineal

Aquí tenemos representado el residuo de la observación i-ésima: $e_i = y_i - \hat{y}_i$

Ahora bien, dado un conjunto de puntos existen infinitas posibles rectas, ¿entonces cómo sabemos cuál elegir?

Si quisiéramos tener un criterio de la calidad del modelo podríamos considerar que un modelo con menos residuos es, a priori, mejor que otro modelo, ¿no?



Modelo supervisado

Regresión lineal

Con esta idea surge el criterio de búsqueda para los estimadores de los parámetros del modelo. Se considera que la mejor regresión lineal es aquella que minimiza la suma de errores cuadráticos.

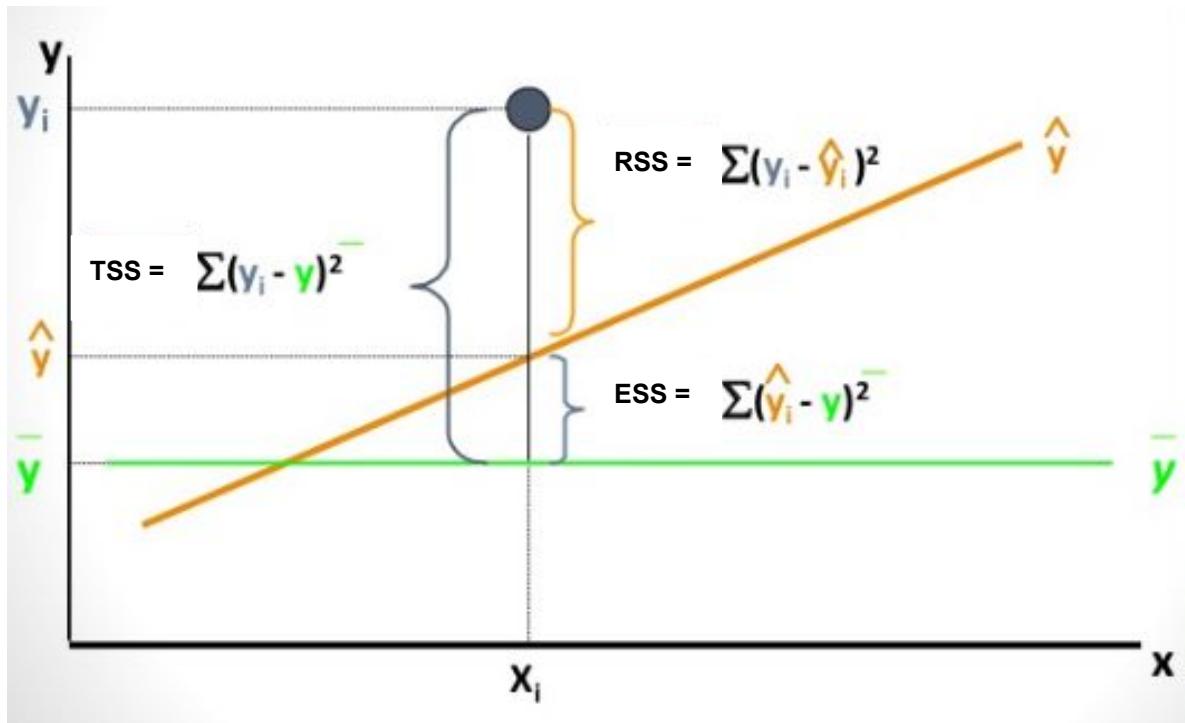
$$SS_{(residuals)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Nótese que al elevar la diferencia al cuadrado se genera el efecto de pesar más que proporcionalmente los errores más graves. Es decir, cuanto peor sea nuestra predicción para una determinada observación más va a pesar eso en el error total.

Modelo supervisado

Coeficiente de determinación

A la hora de evaluar la calidad del modelo se suele usar una medida conocida como R cuadrado o coeficiente de determinación. Esta medida compara la capacidad predictiva de nuestro modelo contra la de usar la media



Modelo supervisado

Coeficiente de determinación

RSS: Variabilidad no explicada por el modelo

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

TSS (Total Sum of Squares): Variabilidad total de los datos

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$$

R^2 : Proporción de la variabilidad explicada por el modelo

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

GRACIAS

Contacto

Docente: Leonardo Ignacio Córdoba

E-mail: cordoba.leonardoignacio@gmail.com

LinkedIn Leonardo Ignacio Córdoba



Buenos Aires Ciudad

