Machine Learning

Big Data

Métodos de clusterización

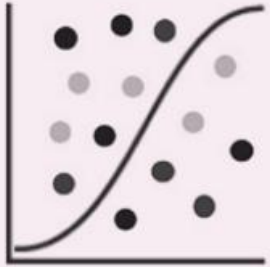
An executive’s guide to AI, McKinsey & Company, 2018

Tipos de análisis

Descriptive

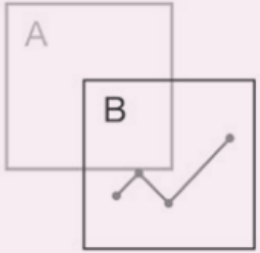
- Describe **what happened**
- Employed heavily across all industries

Predictive



- Anticipate **what will happen** (inherently probabilistic)
- Employed in data-driven organizations as a key source of insight

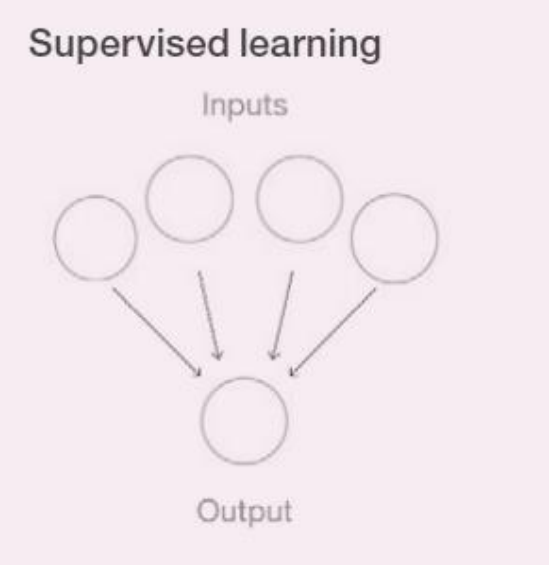
Prescriptive

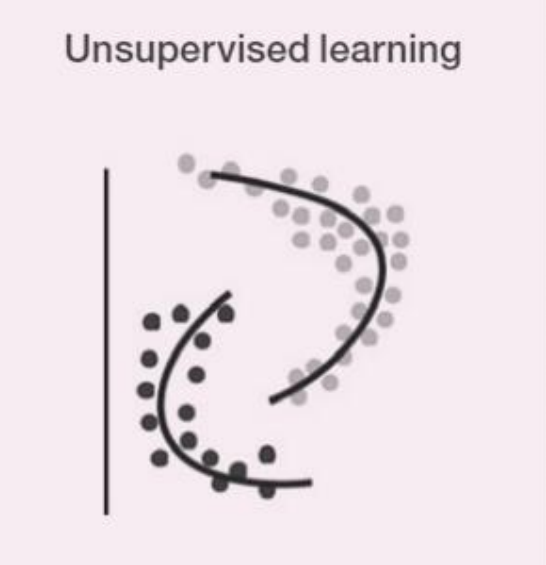


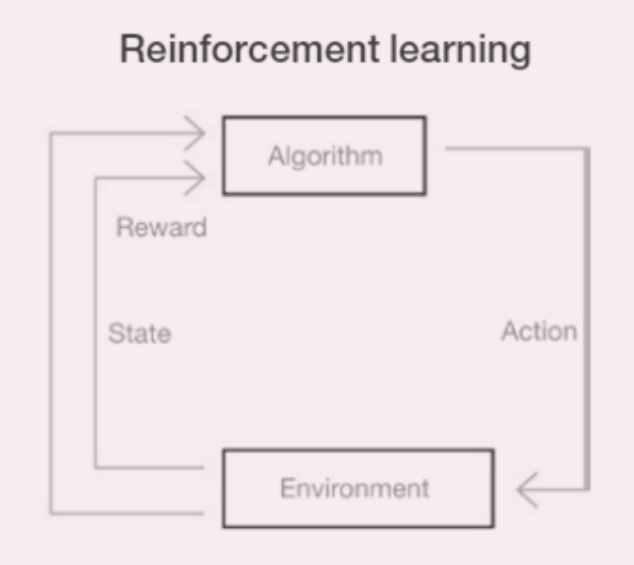
- Provide recommendations on **what to do** to achieve goals
- Employed heavily by leading data and Internet companies

Focus of machine learning

Tipos de Machine Learning







Machine Learning - Supervisado

- Supone la existencia de dos variables
- Una variable que queremos pronosticar (Dependiente)
- Un set de variables que queremos usar para construir ese pronostico o dependencia (Independientes)

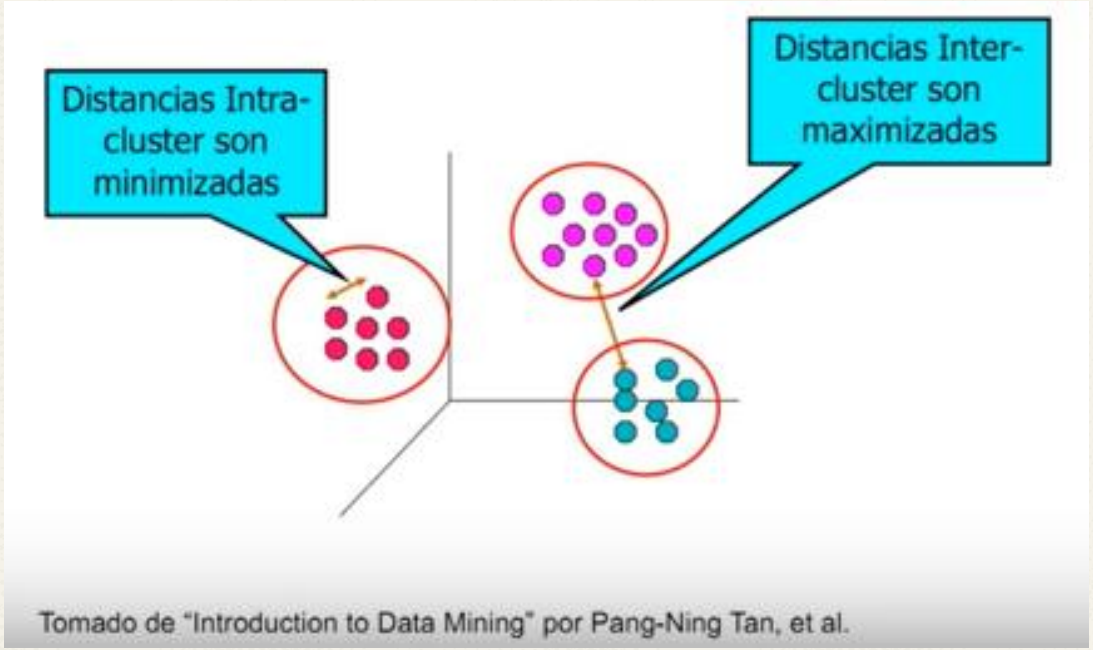
Machine Learning – No Supervisado

- Aquí no hay distinción de variables
- Todas juegan el mismo rol
- Todas tienen el mismo status o importancia
- Entre los tipos de ML No Supervisada podemos mencionar:
  - Clustering
  - Reglas de Asociación

Clusterización - K Means

- Es un método de agrupación o clustering
- Su objetivo es encontrar grupos de observaciones (clusters) con características semejantes
- Similares dentro del grupo, diferentes fuera de él
- Técnicamente diríamos: **Maximizar variación *inter-cluster* y minimizar variación *intra-cluster***

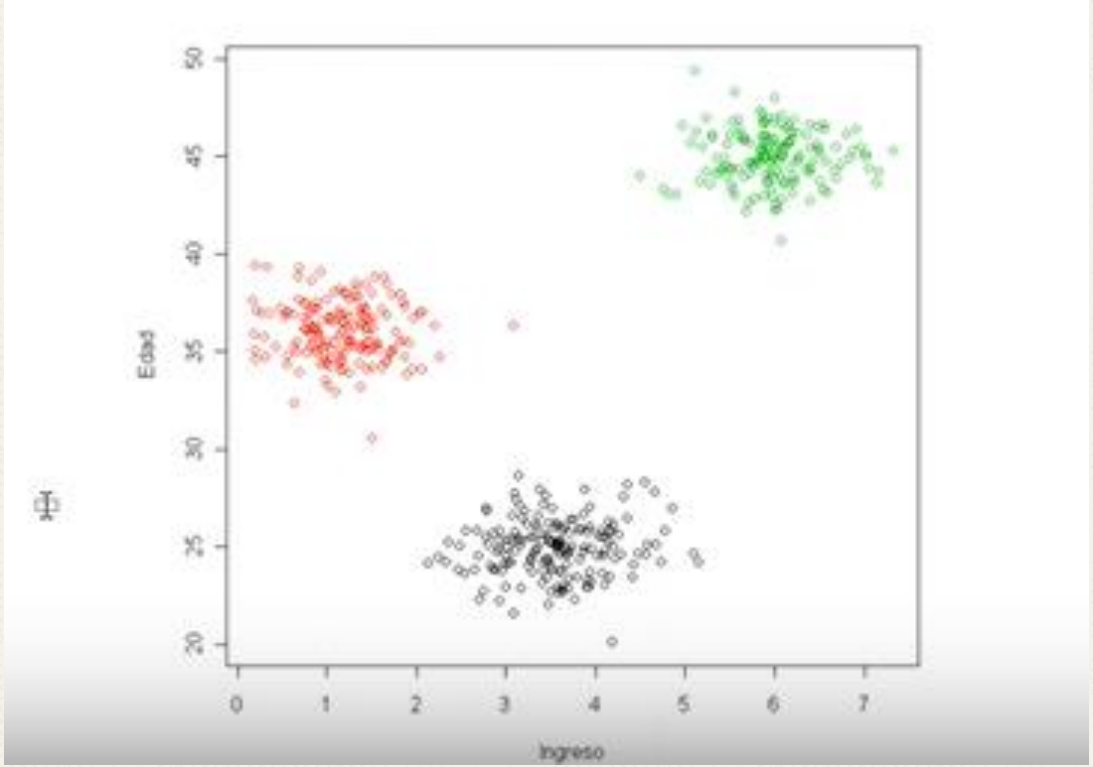
K Means


Distancias Intra-cluster son minimizadas
Distancias Inter-cluster son maximizadas
Tomado de "Introduction to Data Mining" por Pang-Ning Tan, et al.

Ejemplo

- Tenemos datos sobre edad e ingreso para un grupo de consumidores
- La pregunta que surge es: Existen grupos de consumidores con características similares?
- Con dos variables es algo sencillo de interpretar

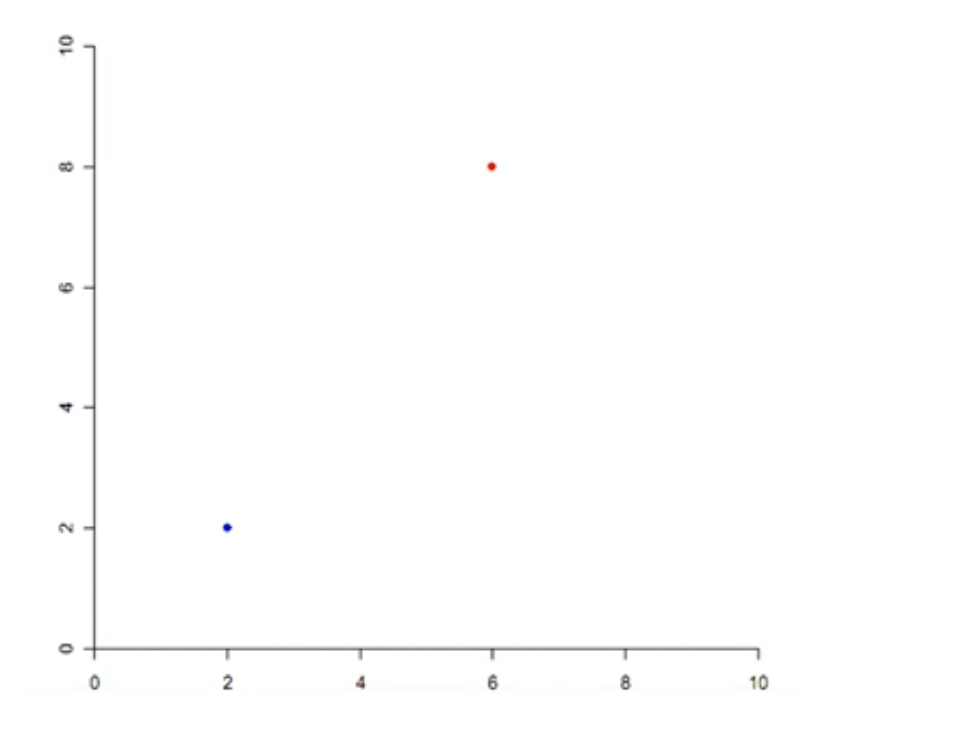
¿Cuántos grupo hay?



Distancia

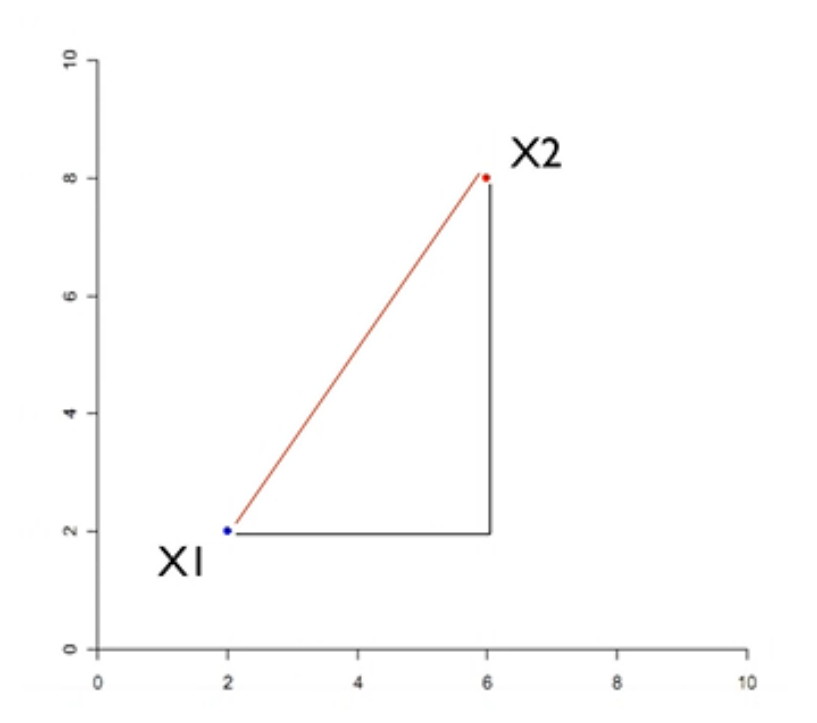
- ¿Cómo se define el concepto de parecido entre dos observaciones?
- Observaciones que están cerca
- Pero que quiere decir que esta cerca o lejos
- Necesitamos definir un concepto de distancia para poder hablar de cercanía o lejanía

Distancia



Distancia

Podemos utilizar la medida de distancia que aprendimos en el teorema de Pitágoras



Técnicamente se conoce como distancia Euclidiana

La distancia entre el punto
x1=(2,2) y el punto
x2=(6,8) es igual a:

D(x1,x2) = [(6-2)^2 + (8-2)2]^0.5
D(x1,x2) = 7.21

Hipotenusa = suma de los catetos del triangulo

Distancia

- En general, cuando tenemos más de dos variables o dimensiones, todavía podemos definir la distancia euclidiana entre dos puntos Xi y Xj:
- Estos puntos ahora son vectores de P dimensiones

$$d(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

K Means

- El método comienza asumiendo que conocemos el número de grupos o clusters
- El método entonces encuentra la “mejor” asignación de puntos a los distintos grupos o clusters
- “Mejor” en el sentido previo de maximizar distancias inter-cluster y minimizar distancia intra-cluster.

K Means

- Decidir # de clusters. Llamemos a este número K
- Un posible método de inicialización: Tomar K observaciones de la muestra al azar. Estas observaciones se convierte en los **centroides** iniciales
- Para cada una de las N-K observaciones restantes, calculamos las distancias entre la observación correspondiente y cada uno de los centroides
- Cada observaciones es entonces asignada al centroide más cercano

K Means

- Cuando terminamos de asignar observaciones tenemos K grupos de observaciones
- Para cada uno de estos grupos, calculamos nuevos centroides.
- El centroide es un vector de medidas de todas las variables utilizadas para las observaciones dentro de cada grupo
- Repetimos el proceso hasta que ya no haya reasignaciones

K Means

Cosas a tener en cuenta

- No hay ninguna garantía de que el algoritmo encuentre la solución óptima
- Una mala selección inicial de centroides puede resultar en un pobre agrupamiento
- Recomendación: Recomenzar el algoritmo varias veces desde puntos diferentes. Quedarse con la mejor solución.

K Means

- K Means y cualquier otro algoritmo que se calcule basado en distancias, puede ser afectado por las unidades en que las variables se miden.
- Variables medidas en las unidades más grandes dominaran la construcción de los clusters
- Recomendación: Estandarizar las variables antes de iniciar la búsqueda de clusters.