# Módulo 6 MDA - Advanced Analytics
# Machine Learning
## Sesión 1 - Clasificación
## Sesión 2 - Clusterización

José Juan Martínez Carrascosa
Jose J. Martinez
jjmcarrascosa@gmail.com

<ISDI>

## Índice (sesión teoría)

1. Sobre **las dinámicas y tiempos de clase**
2. **Inteligencia Artificial y Machine Learning**
    a. Activación de conocimientos previos
    b. Inteligencia Artificial
    c. Negocio e IA (McKinsey 2017)
    d. Data Science y Data Analytics. Estadística y ML.
    e. Machine Learning. Deep Learning
3. **ML supervisado vs no supervisado**
4. ML supervisado: **Clasificación**
    a. Por tipo de problema: binaria, multiclase, multietiqueta
    b. Por tipo de algoritmo: regresión, modelos bayesianos, descenso por gradiente, SVM, árboles de decisión y boosting
    c. Métricas
5. ML no supervisado: **Clusterización**
    a. Por tipo: de partición, de densidad, de rejilla.
    b. Métricas
6. Práctica 1: Clasificación binaria
7. Práctica 2: Clusterización

<ISDI>

# Sobre las dinámicas y tiempos de clase

# Dinámicas de clase

*FRONTAL*

*Introducción a conceptos nuevos, explicaciones magistrales*

**Teoría: 2 horas**
**Práctica: 45 minutos**

*GRUPOS*

*Anticipación, deducción, brainstorming*

**Teoría: 1 hora**
**Práctica: 2h 15 minutos**

# Inteligencia Artifical y Machine Learning

## Activación de conocimientos previos

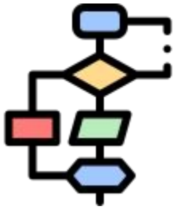
https://forms.gle/bn2KqyQRjaRTQCHF9

<ISDI>

# Inteligencia Artificial

> La **Inteligencia Artificial** (IA) es la combinación de algoritmos alimentados con datos que cuyo propósito es el crear modelos *estadísticos* que presenten las mismas capacidades de ejecución que el ser humano

Algoritmos



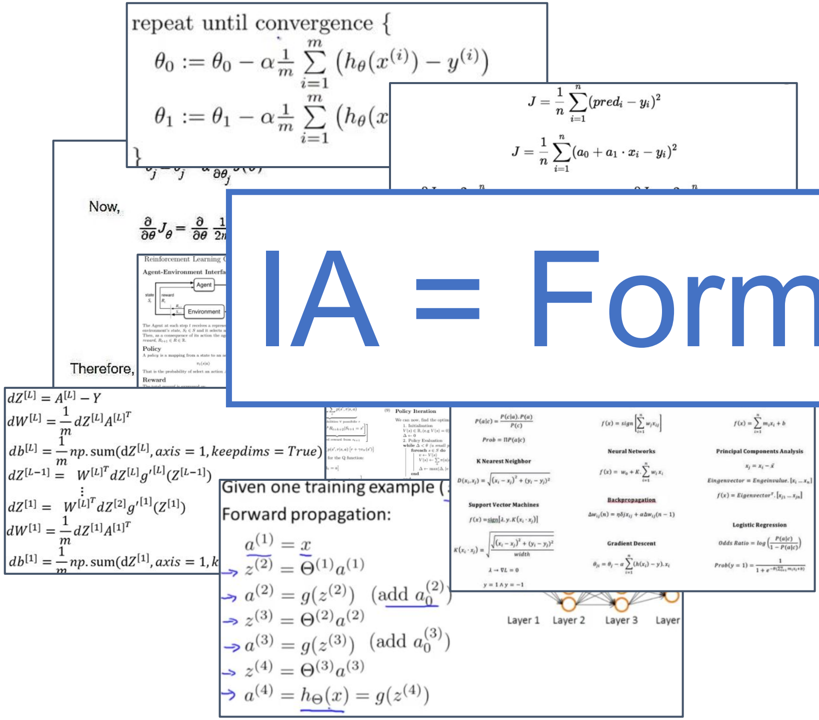
+

Datos

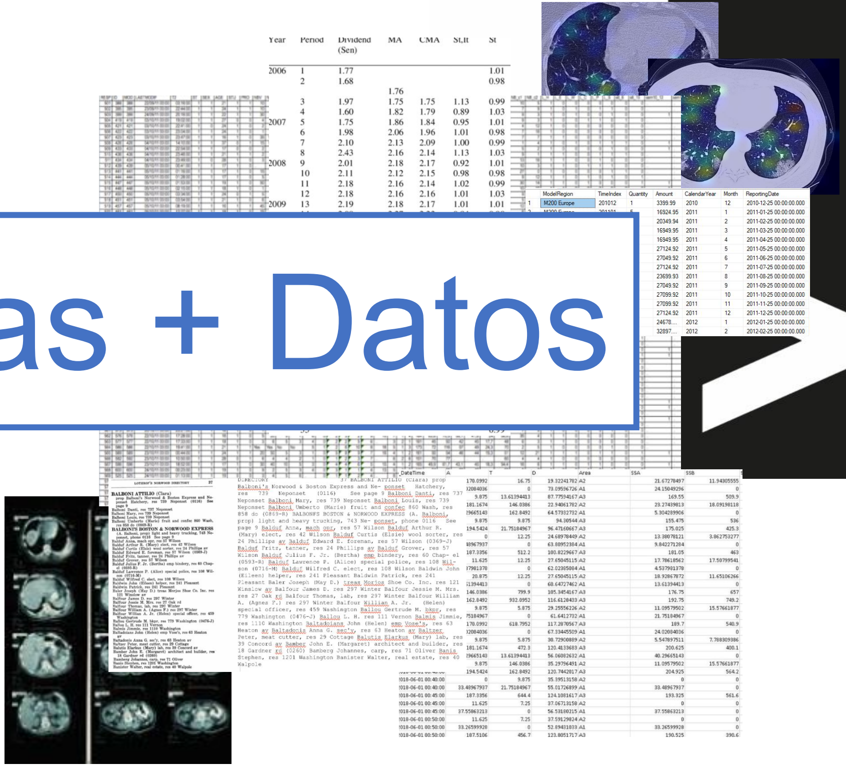
=

Modelos estadísticos

# Inteligencia Artificial

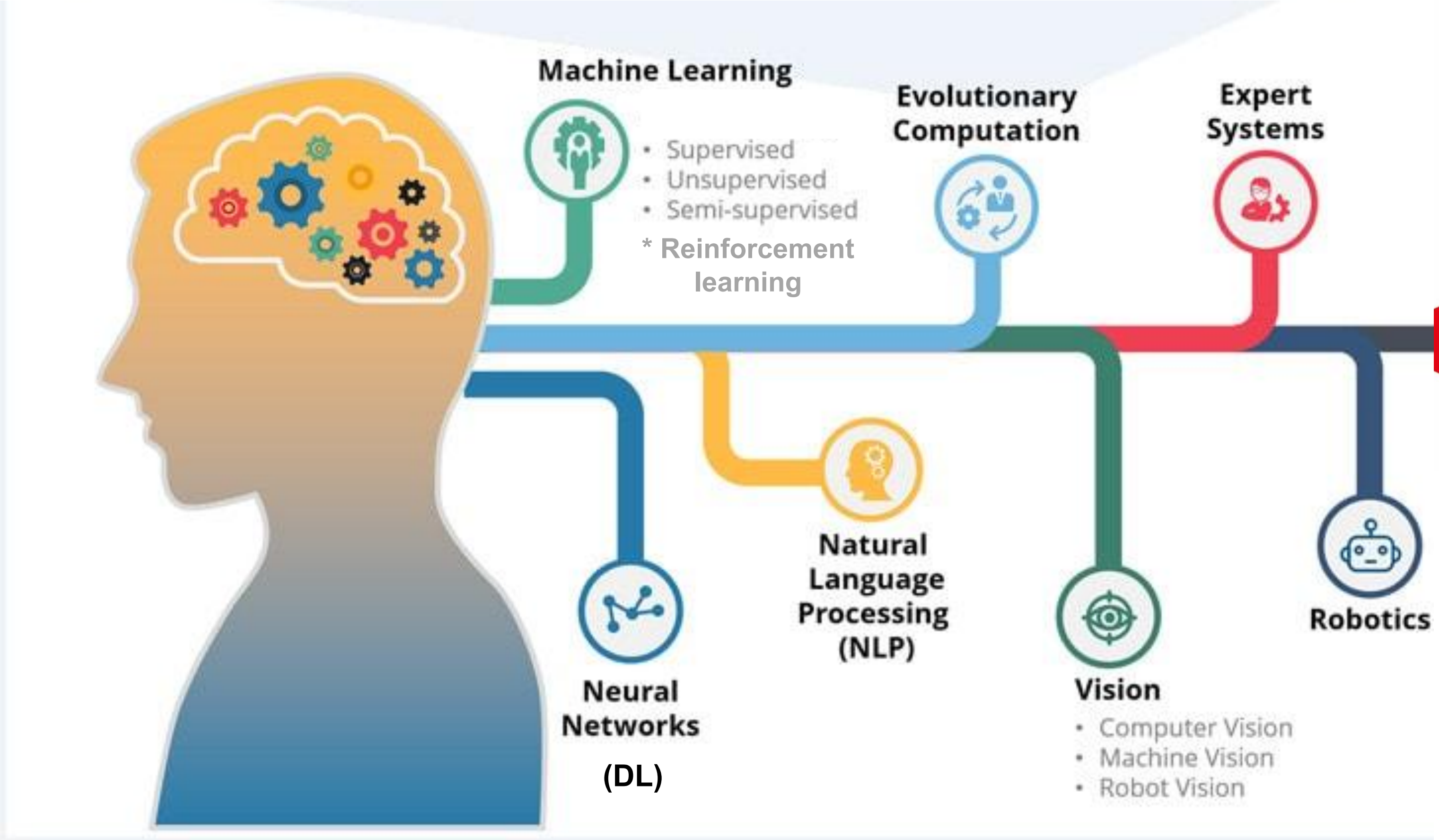



# IA = Formulas + Datos



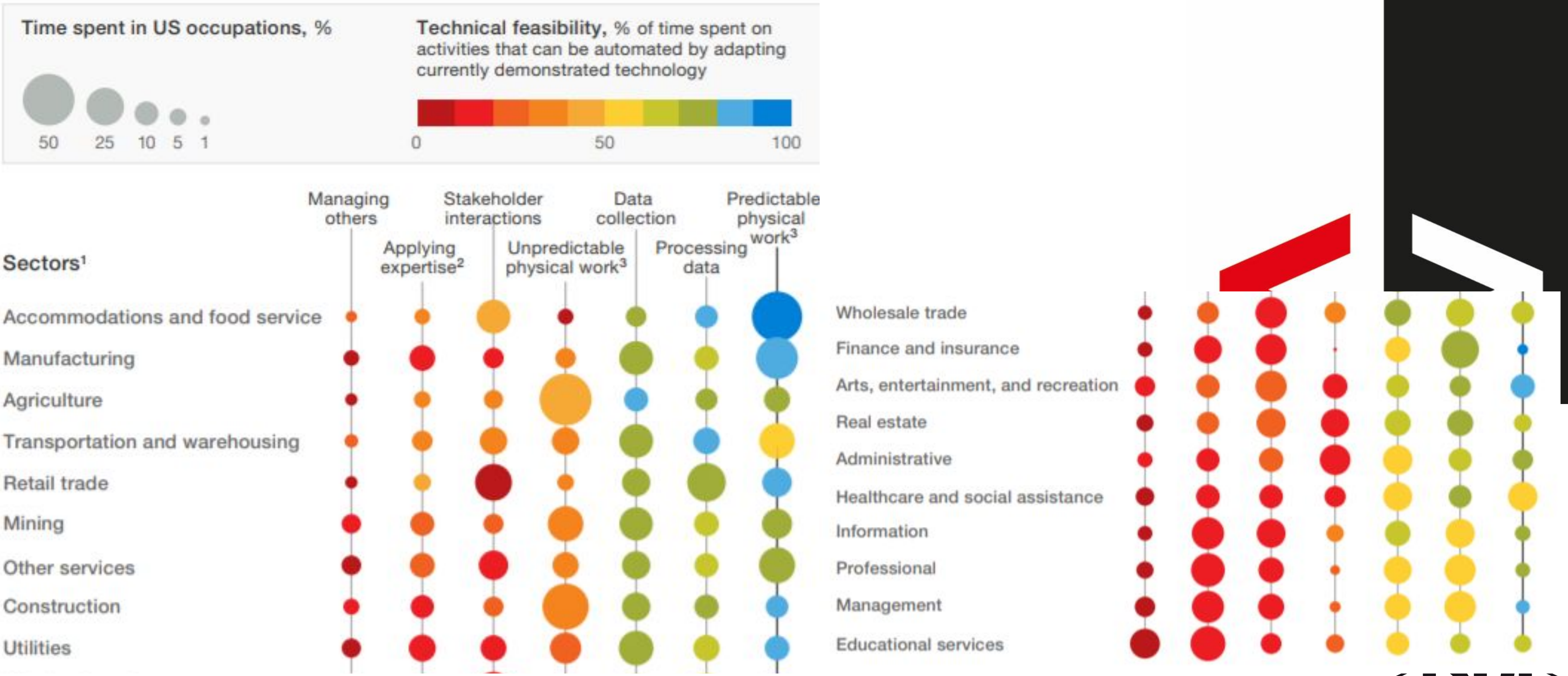

<ISDI>

# Inteligencia Artificial



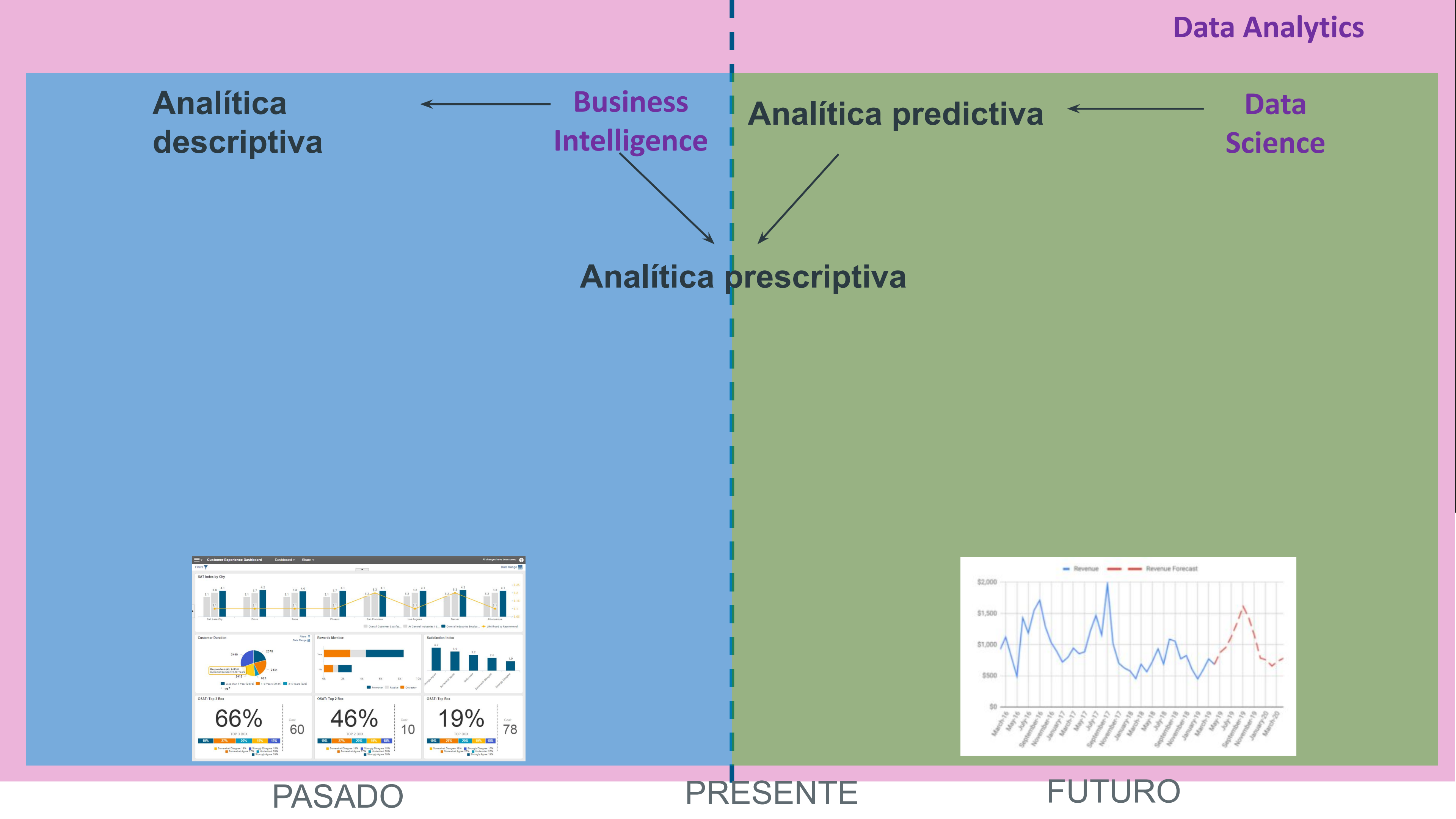
# Potencial de automatización por area de negocio IA

# Potencial de automatización por áreas




McKinsey, 2017

# Data Science y Data Analytics


Data Analytics
Analítica descriptiva
Business Intelligence
Analítica predictiva
Data Science
Analítica prescriptiva
Customer Experience Dashboard
66%
60
46%
10
19%
78
PASADO
PRESENTE
FUTURO

Data Analytics

**Analítica descriptiva** ← **Business Intelligence**

**Analítica predictiva** ← **Data Science**

**Analítica prescriptiva**

Paneles de control y reportes

KPI

Clusterización de clientes

Monitorización de consumos

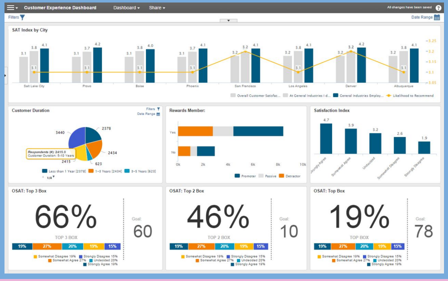
Cruce de datos, agregaciones, proyecciones

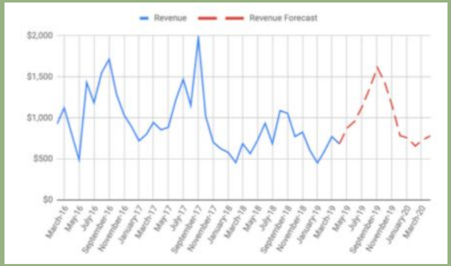
Forecasting de consumos, ventas, ambulatorización, contagios...

Retención de clientes – CHURN

Automatización de toma de decisiones

Alertas sobre tendencias futuras




PASADO

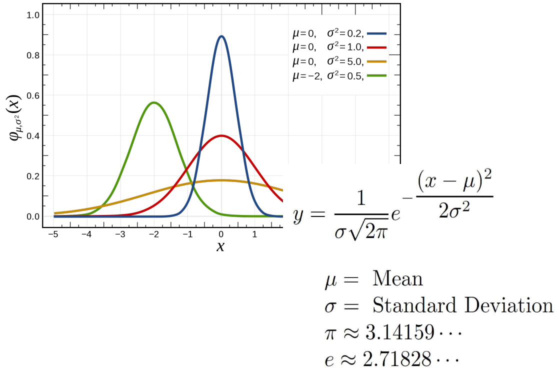
PRESENTE

FUTURO

# Estadística y ML

# Estamos rodeados de sucesos complejos y no vemos correlaciones estadísticas que ML si ve


**ESTADISTICA**
La **estadística** consiste en **conocer** y **aplicar** las mejores **formulas** que modelen los datos

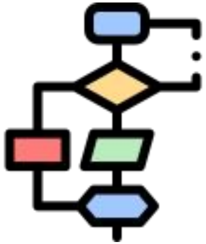



**MACHINE LEARNING**
**ML** se basa en los **datos** y **automatiza** ese modelado y aplicación de fórmulas

# Machine Learning

# ML, regresión (entrenamiento)

**Algoritmo:**
Regresión lineal

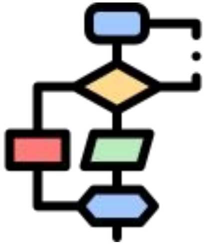


**Datos anotados (var. dep e indep)**

| MES | DIA | OUTPUT |
|---|---|---|
| Junio | 16 | 0.7 |
| Diciembre | 21 | 21 |
| ... | ... | ... |

# ML, regresión (inferencia)

**Algoritmo:**
Regresión lineal



**Datos anotados (var. dep e indep)**

| MES | DIA |
|---|---|
| Enero | 6 |
| Marzo | 11 |
| ... | ... |

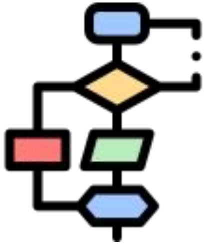
| OUTPUT |
|---|
| 19 |
| 24 |
| ... |

<ISDI>

# Deep Learning

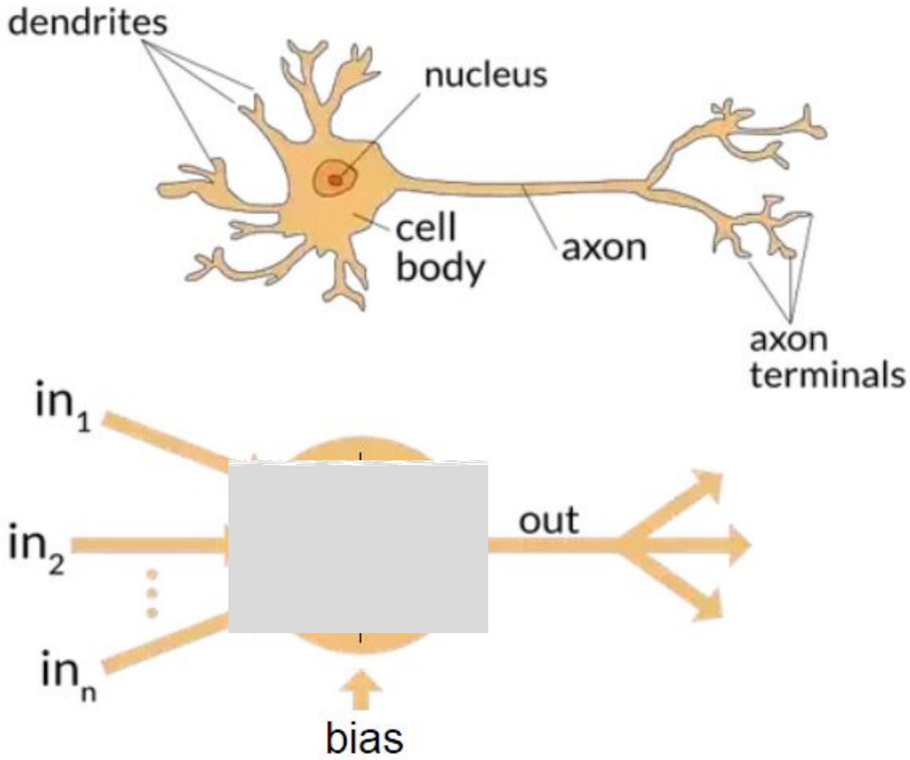
# ML, regresión (entrenamiento)

**Algoritmo:**
Regresión lineal

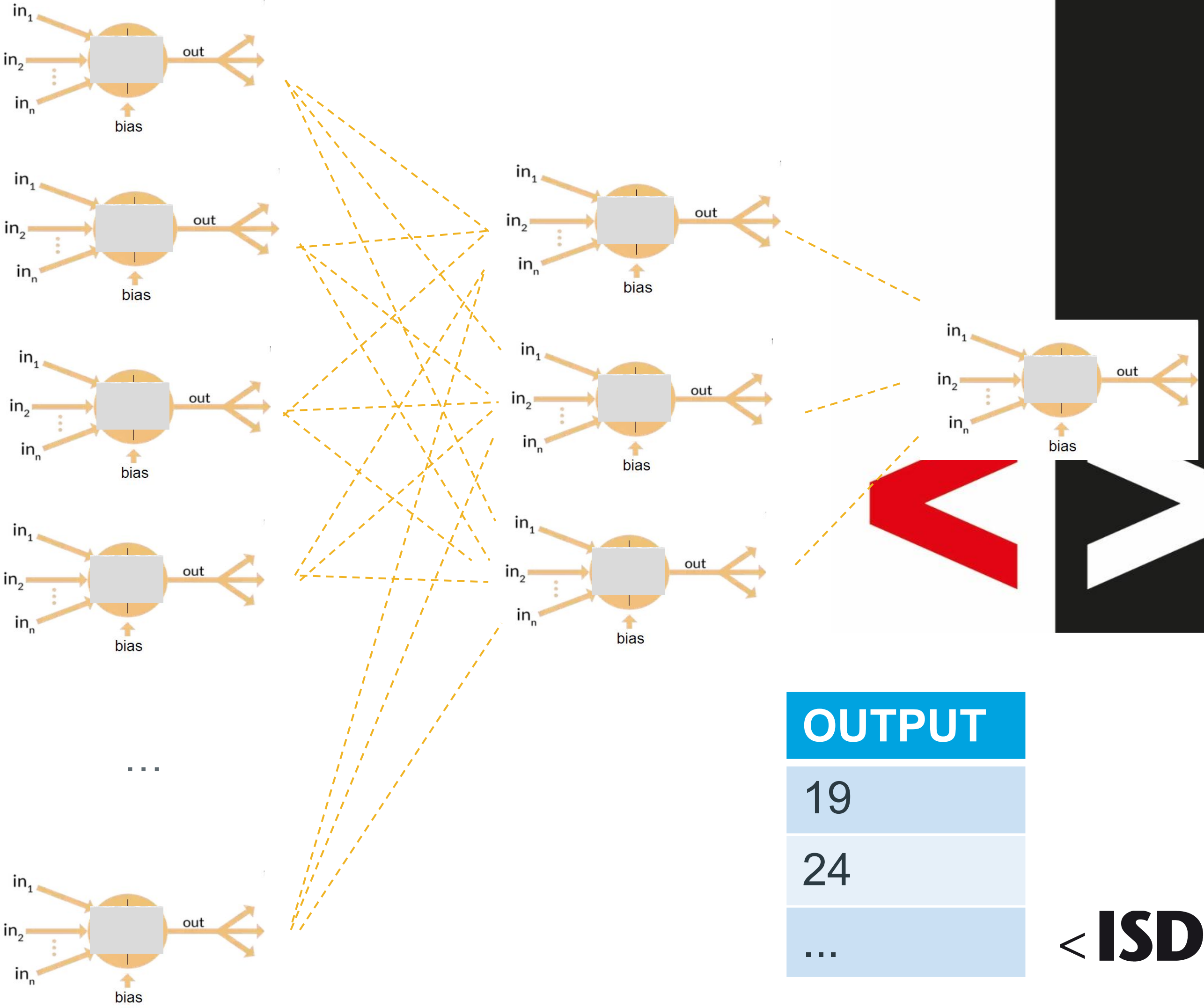


**Datos anotados (var. dep e indep)**

| MES | DIA | OUTPUT |
|---|---|---|
| Junio | 16 | 0.7 |
| Diciembre | 21 | 21 |
| ... | ... | ... |



| MES | DIA |
|---|---|
| Junio | 16 |
| Diciembre | 21 |
| ... | ... |




| OUTPUT |
|---|
| 19 |
| 24 |
| ... |

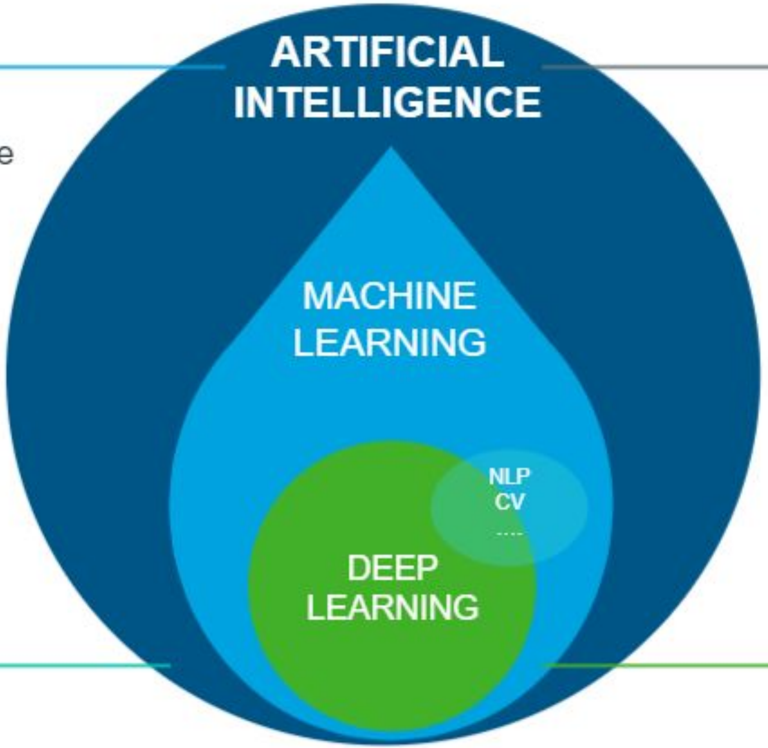
# Final del resumen

**Artificial Intelligence**

Any technique which enable computers to mimic intelligence

**Deep Learning**

Subset of machine learning that drives un/supervised learning from data that is unstructured or unlabeled




**Machine Learning**

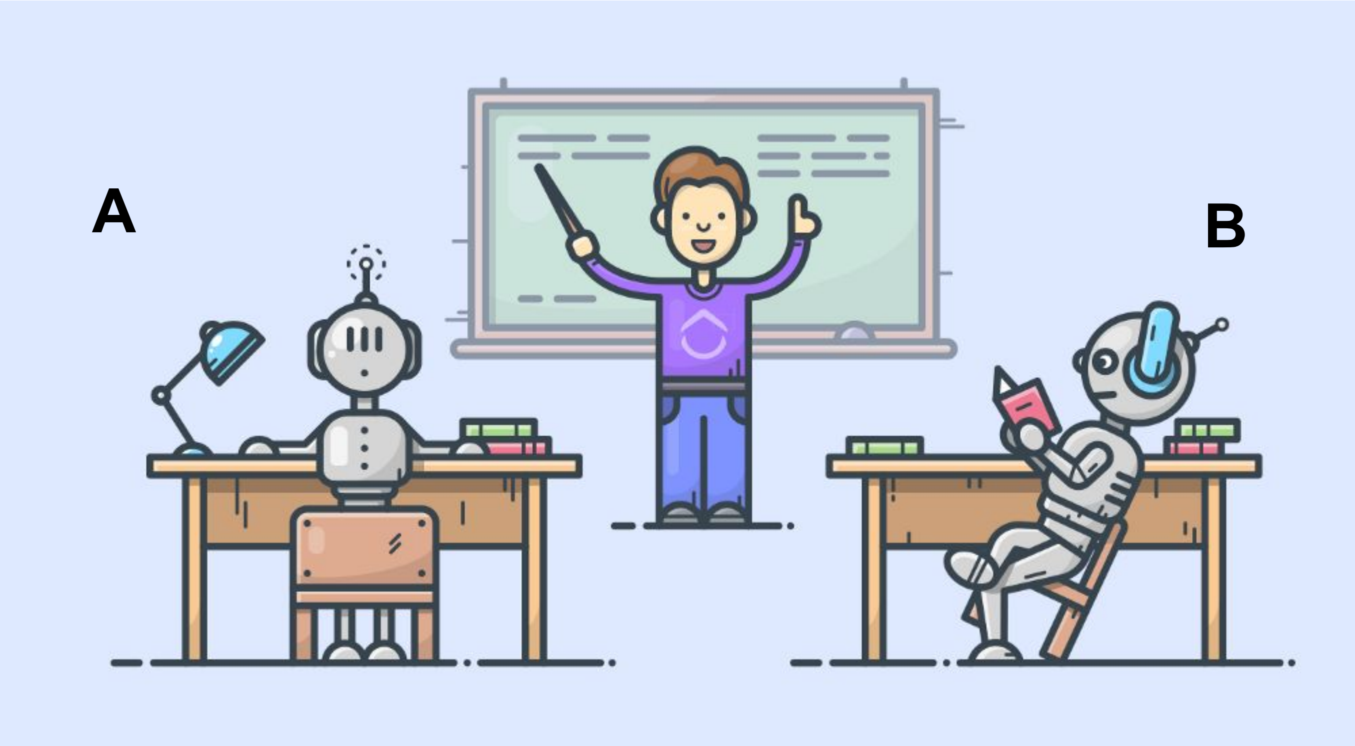
Subset of AI techniques which use statistical methods to enable machines to improve a result with experience

**Natural Language Processing and Computer Vision**

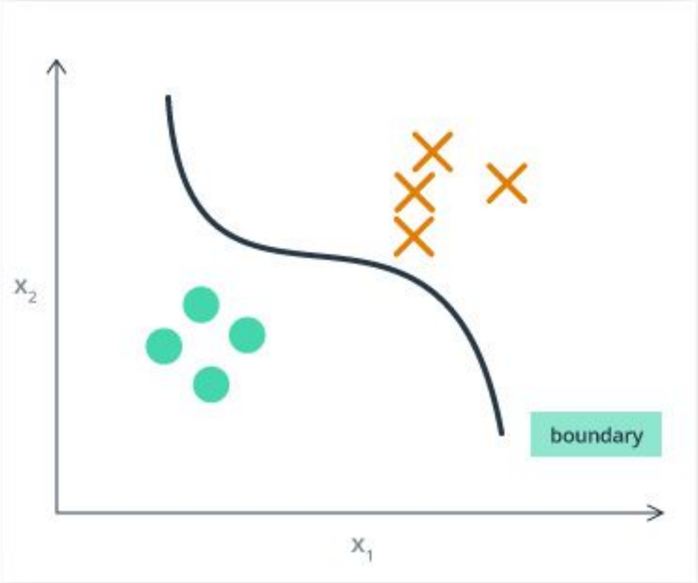
Programing that allows computers to understand human language or images

# Machine Learning supervisado y no supervisado

¿Quien está aprendiendo de forma no supervisada?

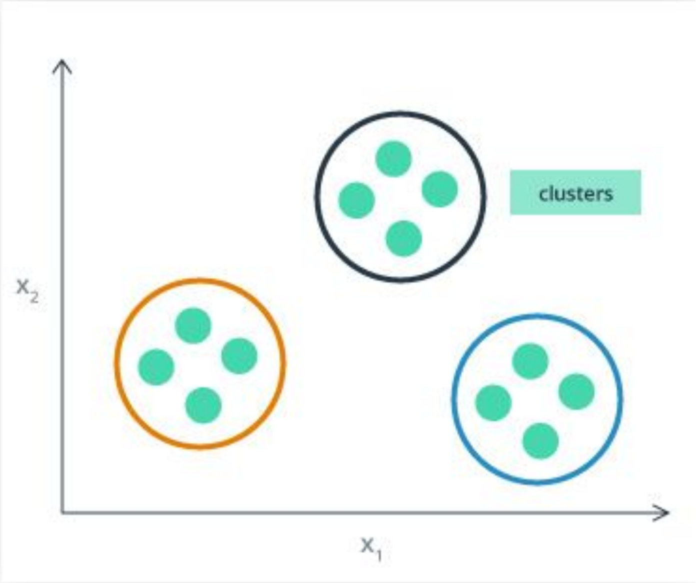


Supervised learning




Ej: Regresión, clasificación, etc

Unsupervised learning

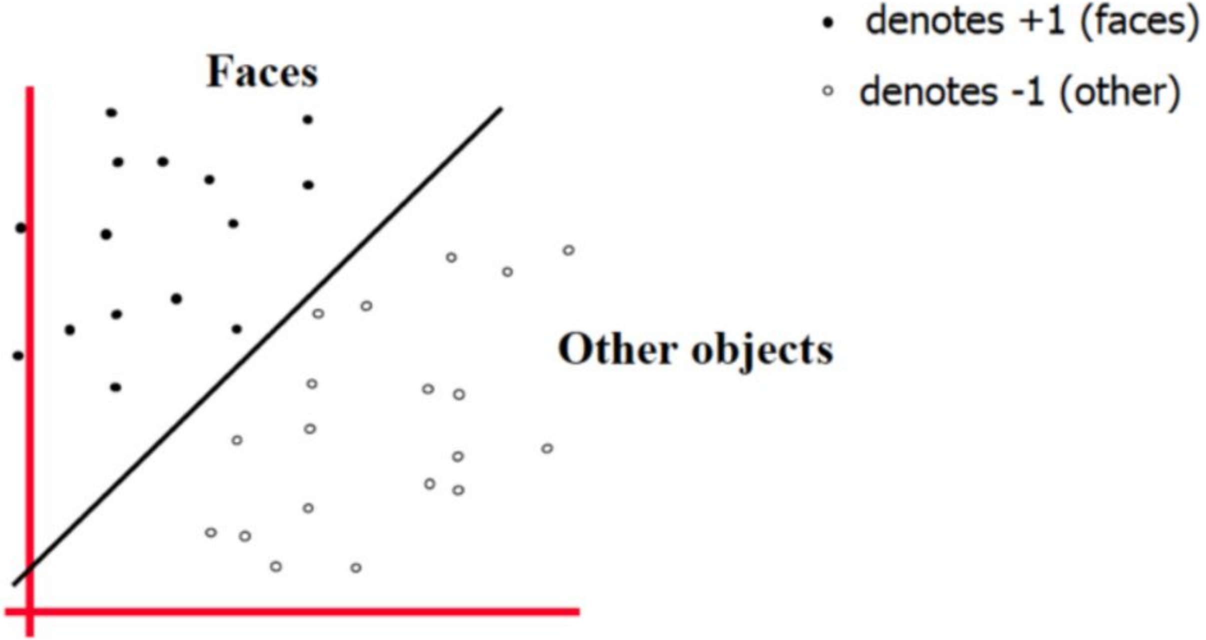



Ej: Clusterización, reducción de dimensionalidad, modelos de lengua en NLP. etc

| **Aprendizaje supervisado** | **Aprendizaje no supervisado** |
|---|---|
| Datos *etiquetados* a priori | Datos *no etiquetados* |
| Output *conocido* | Output *desconocido* |
| Enfoque *predictivo* | Enfoque *descriptivo* |
| *Regresión, clasificación* | *Clusterización, asociación, modelos de lengua* |
| *Más complejo de entrenar*, necesita etiquetado humano | *Menos complejo* |
| *Muchos casos de uso* | Menos casos de uso pero muy relevante ultimamente en NLP con modelos de lengua como Google BERT |

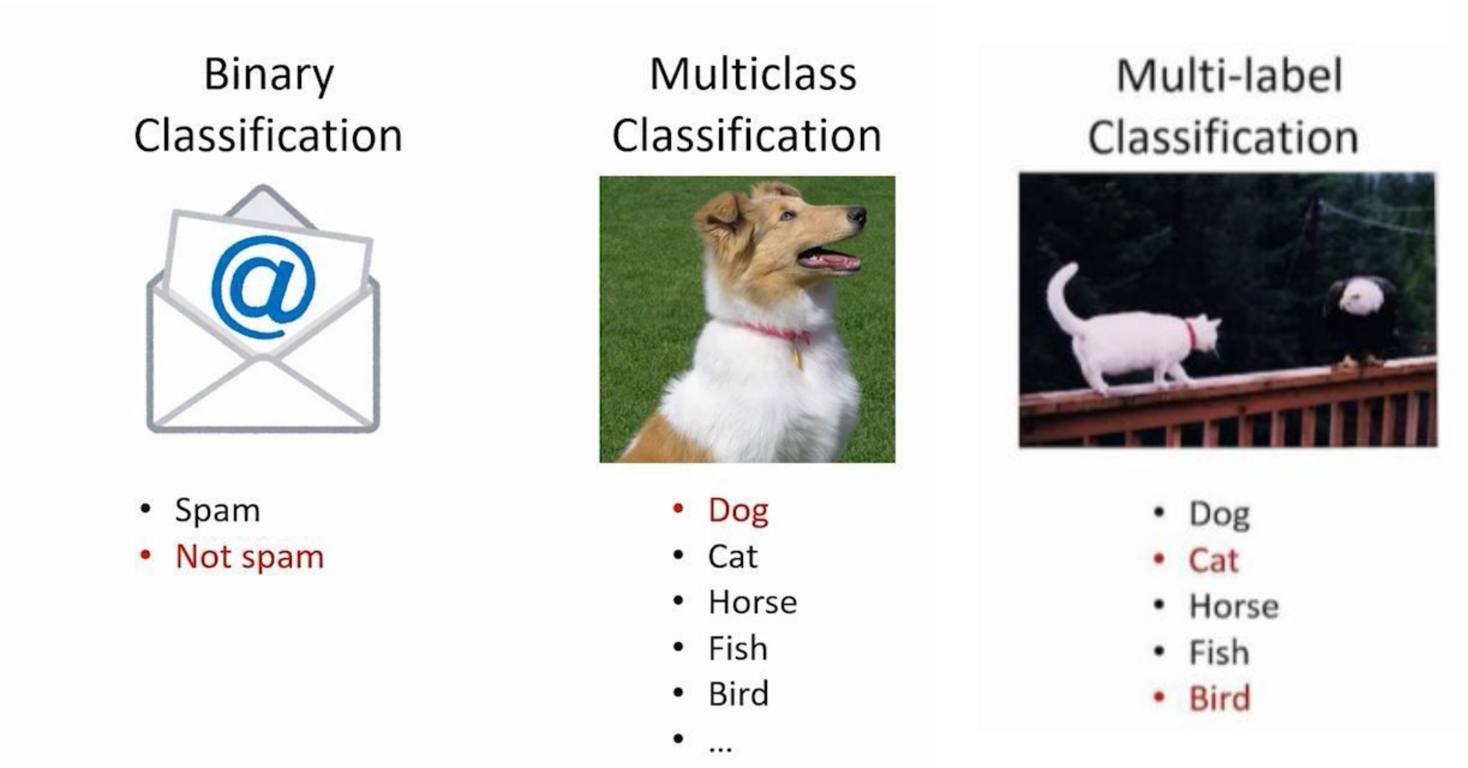
# Machine Learning supervisado: Clasificación

# Clasificación vs Regresión




Idea: Find a separating hyperplane (line in this case)

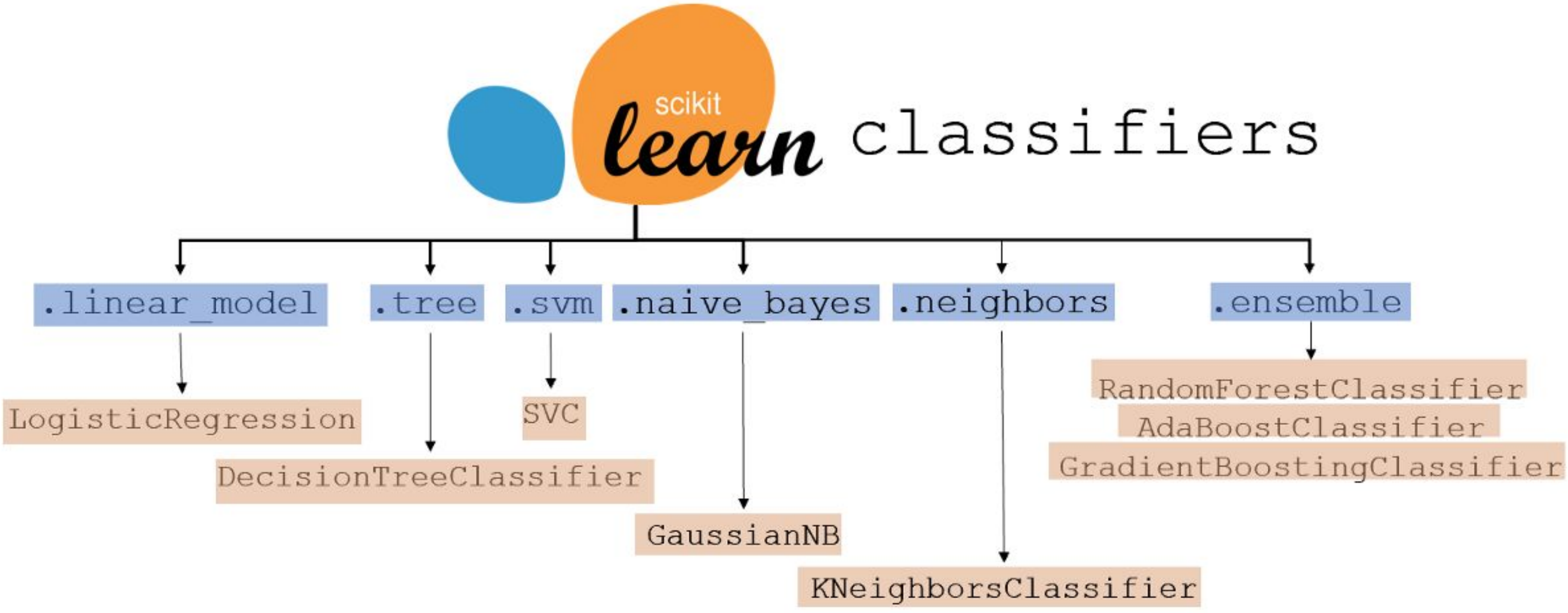
# Tipos de clasificación



# Tipos de clasificación


K = Total number of classes in the problem statement
C = Number of classes an item maybe assigned to

# Tipos de algoritmos



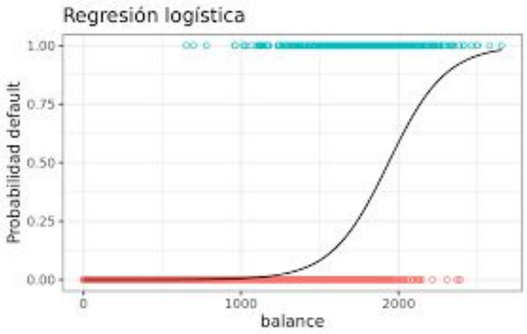

https://scikit-learn.org/stable/supervised_learning.html

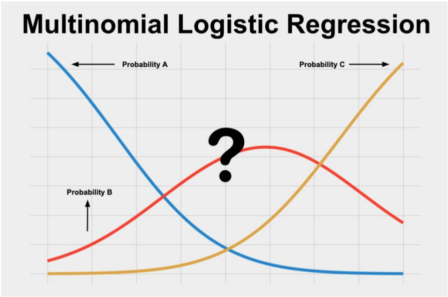
<ISDI>

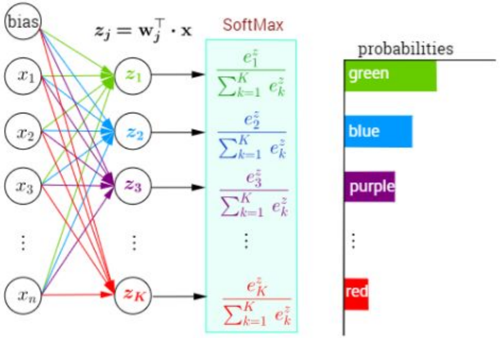
# Tipos de algoritmos

**No todos los algoritmos soportan las 3 modalidades (binaria, multiclase, multietiqueta).**

**Aunque muchos de ellos tienen adaptaciones.**




Para saber qué algoritmos están disponibles, acceder a:

https://scikit-learn.org/stable/modules/multiclass.html

<ISDI>

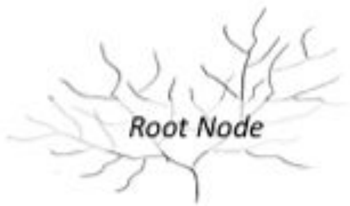
# Modelos lineares: Regresión logística

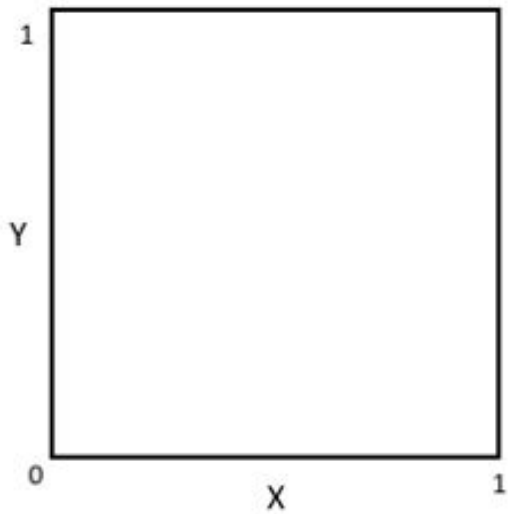
(Y)
cliente
crítico

cliente
NO
crítico


número de seguros con
nosotros (X)

# Árboles de decisión (Decision Trees)



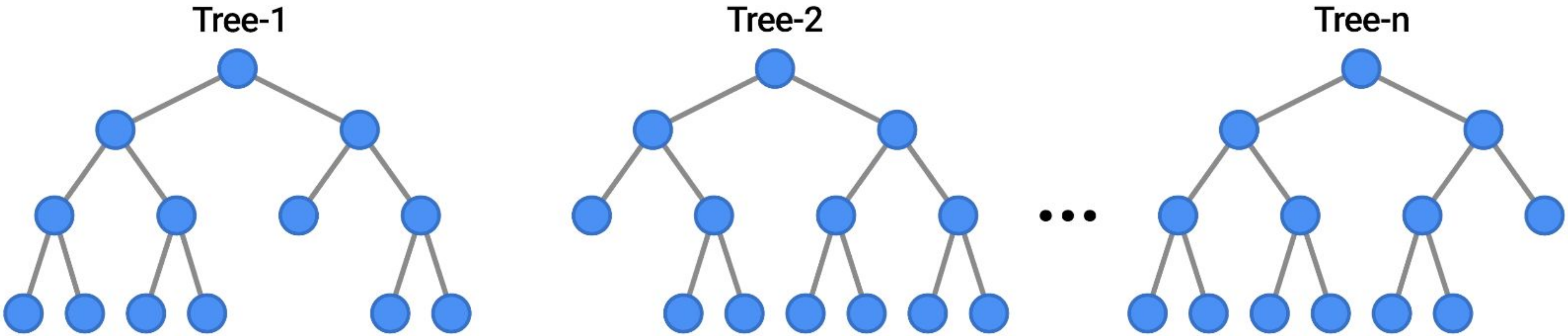
# Ensemble

Algoritmos de aprendizaje para obtener un mejor rendimiento predictivo que el que se podría obtener de cualquiera de los algoritmos de aprendizaje constituyentes
Se distinguen dos familias de métodos de conjunto:

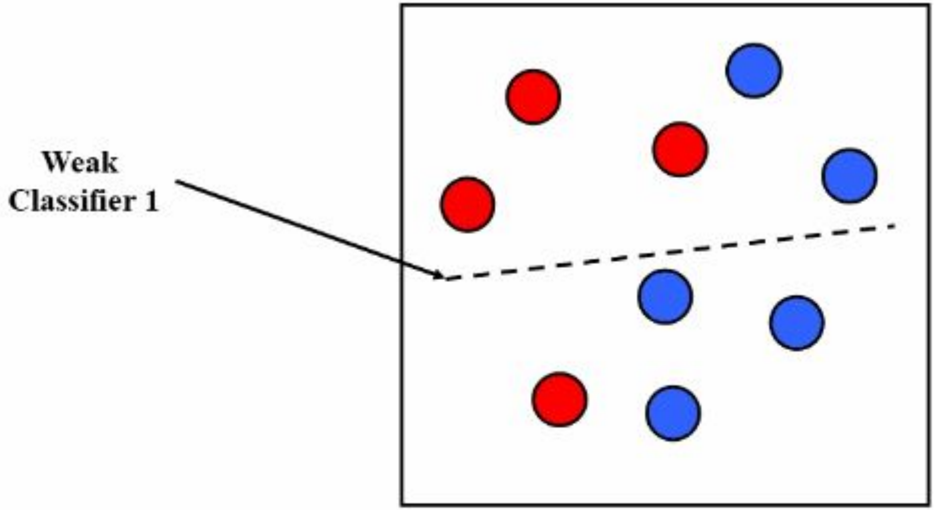
**En los métodos de votación (bagging):** el principio impulsor es construir varios estimadores de forma independiente y luego promediar sus prediccione. Ej: RandomForestTrees.

**Los métodos de impulso (boosting)**, los estimadores base se construyen secuencialmente y se intenta reducir el sesgo del estimador acumulado. La motivación es combinar varios modelos débiles para producir un modelo mejor. Ejemplo: AdaBoost, XGBoost.

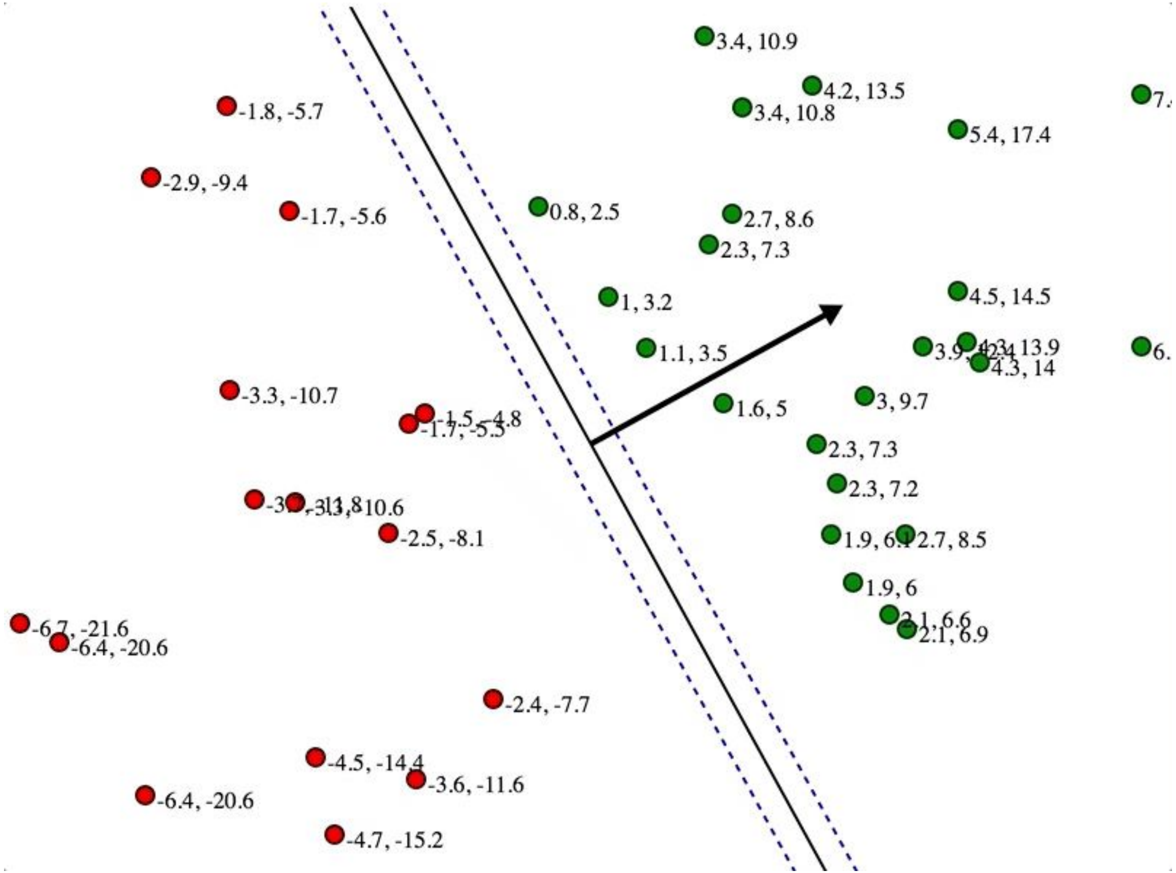
# Ensemble - Bagging (votación). Random Forest



# Ensemble - Boosting (impulso) - Adaboost



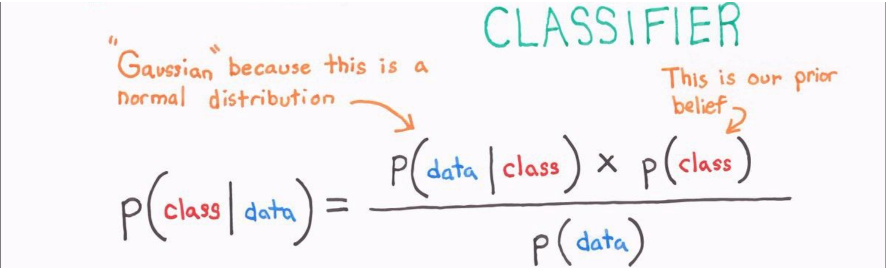
# Support Vector Machines

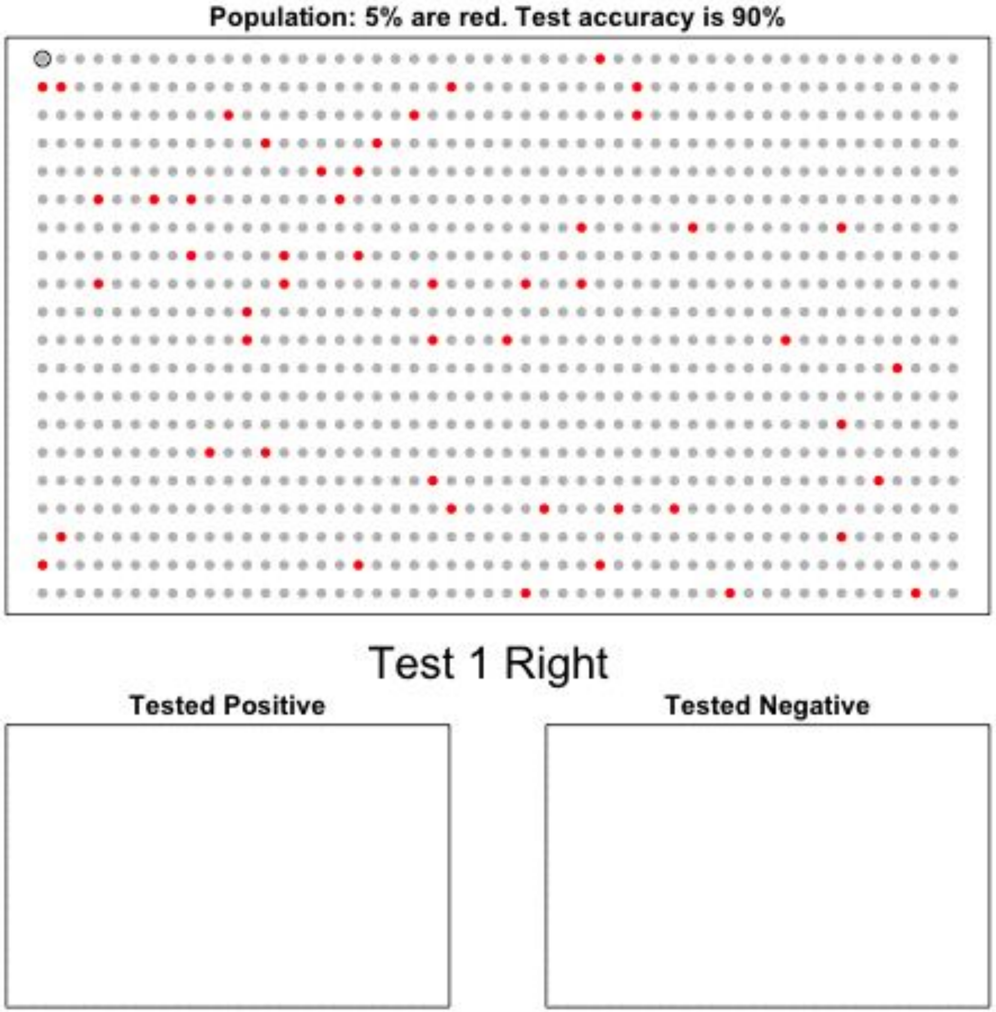



<ISDI>

# Modelos Bayesianos

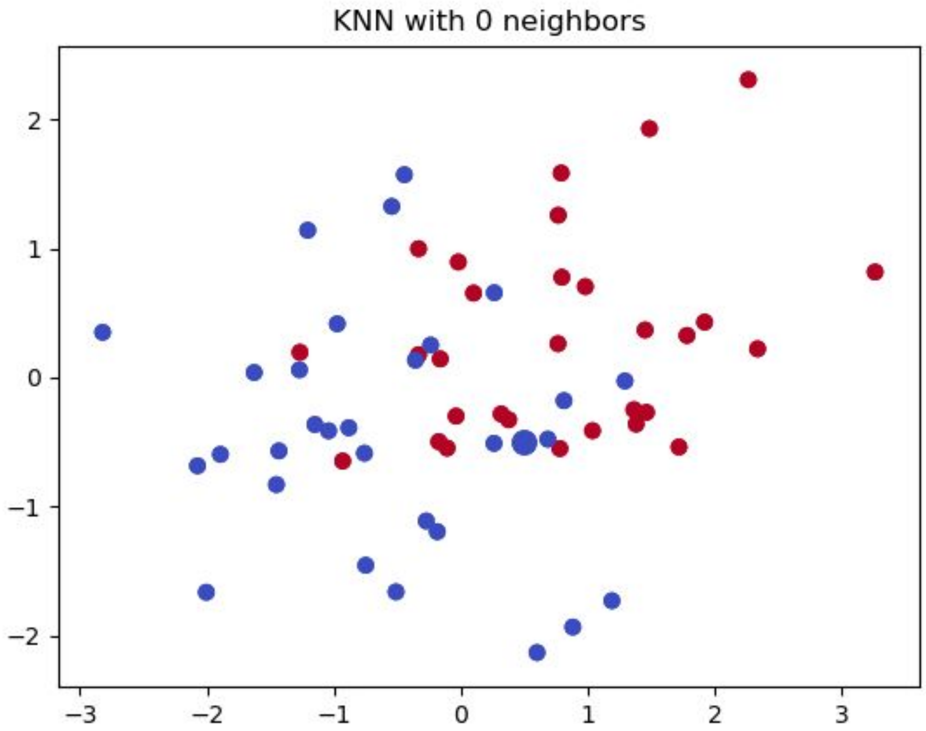
Modelos basados en el Teorema de Bayes, que calcula la probabilidad de darse una clase A teniendo una observacion B, en base a la probabilidad de haberse visto B y ser una clase A.



# Modelos de vecinos

**k-Nearest Neighbours:** Trata de buscar los K puntos / vecinos más cercanos a un punto concreto para poder inferir su valor.



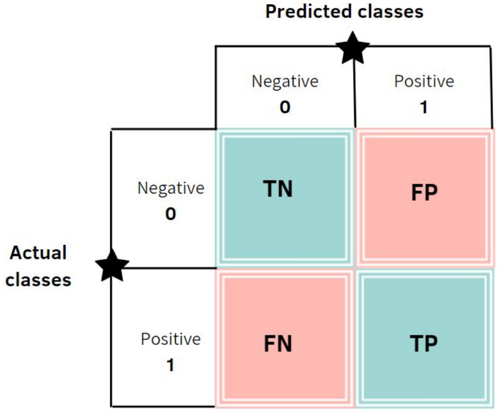
# Criterios de éxito

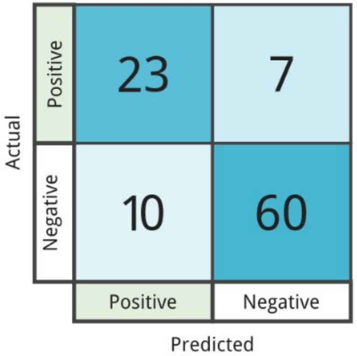
Queremos:

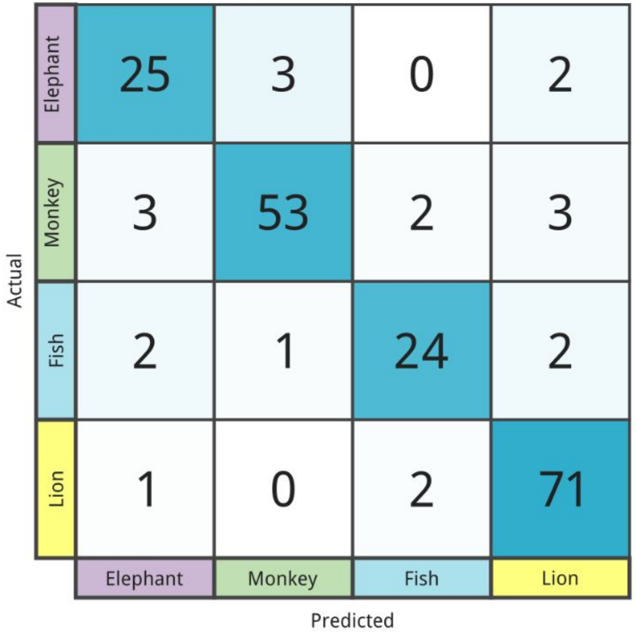
- **Maximizar la precisión** de la predicción
- **Minimizar el tamaño de la hipótesis**
- **Maximizar la adecuación** de la hipótesis a los datos de entrada
- **Maximizar la interpretabilidad** de la hipótesis
- **Minimizar la complejidad** temporal de la predicción

# Métricas de rendimiento

- **Matriz de confusión:** resumen de los resultados de predicción (Falsos Positivos **FP**, Falsos Negativos **FN**, Verdaderos Positivos **TP**, Verdaderos Negativos **TN**) por cada combinacion 1 a 1 de clases.



# Métricas de rendimiento

**Recall:** Cuántos positivos consigo detectar del total de positivos existentes (cuánto abarco)

**Precisión:** Qué tasa de positivos correctos vs positivos incorrectos tengo (cuanto aprieto)

**F1:** Promedio (media armónica) de Recall y Precisión

> *El que mucho abarca poco aprieta*

# Métricas de rendimiento

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

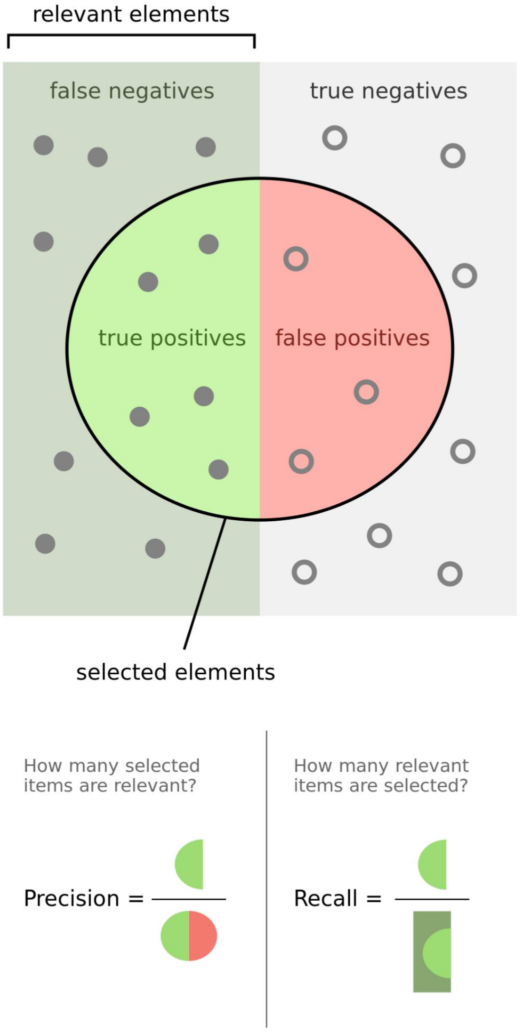
$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$TP$ = True positive

$TN$ = True negative

$FP$ = False positive

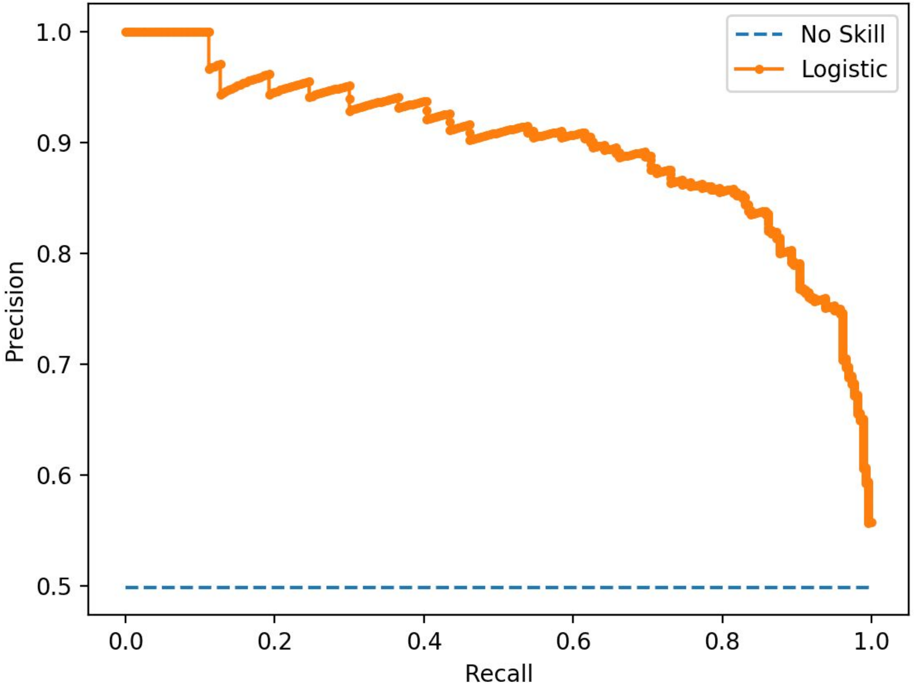
$FN$ = False negative



# Métricas de rendimiento

**Curva Precisión / Recall:** Como evoluciona la precisión cuanto más abarcamos (aumentamos recall)

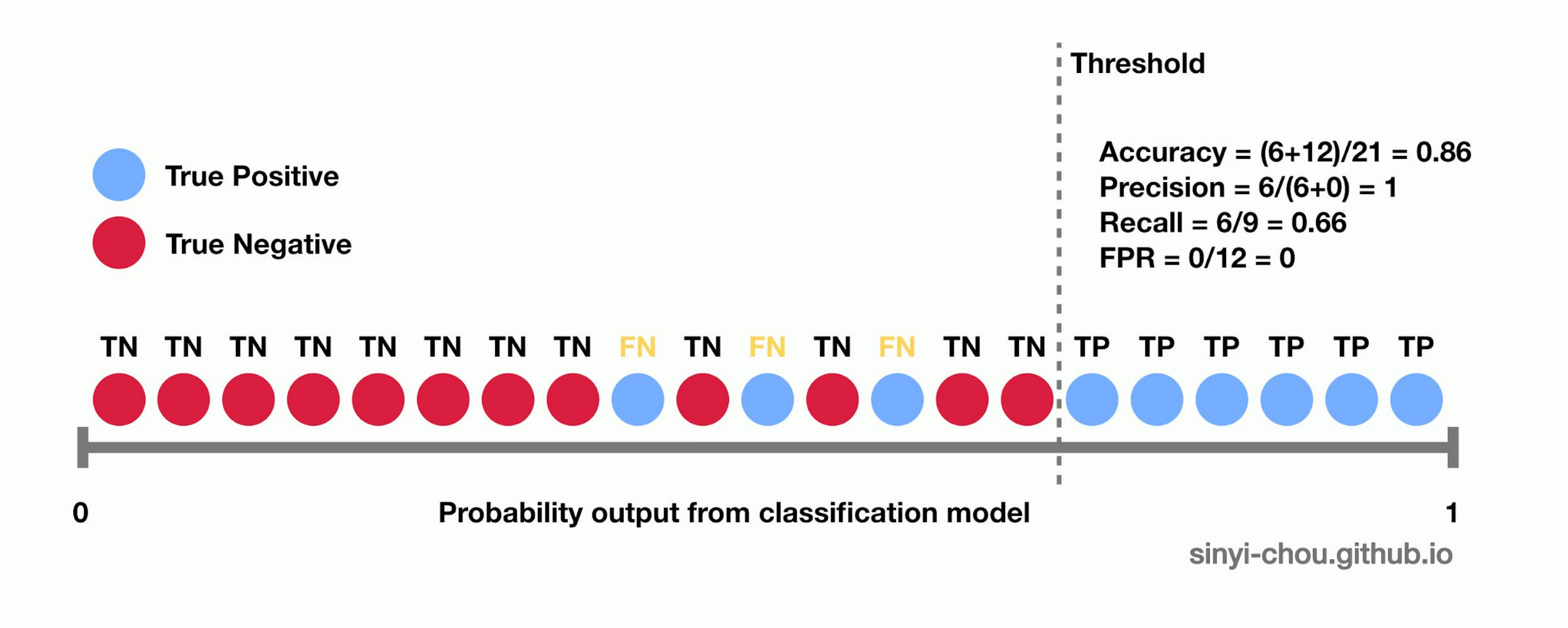
*“El que mucho abarca poco aprieta”*



# Métricas de rendimiento

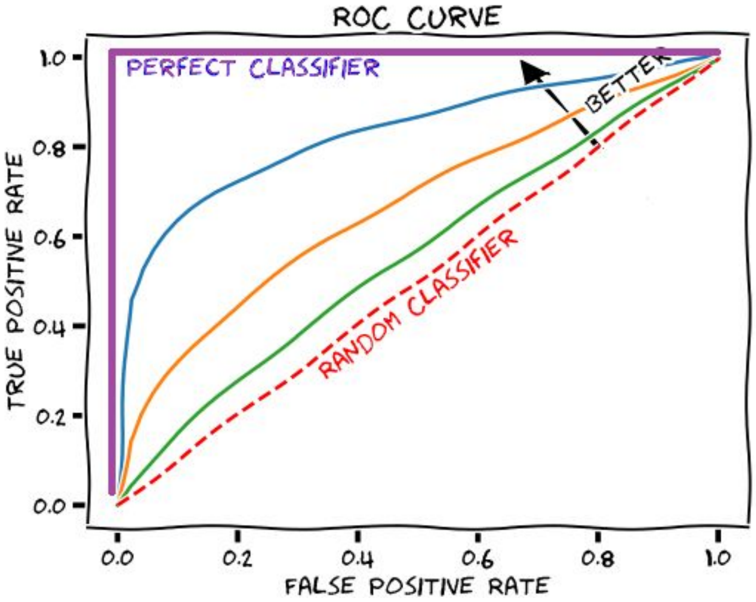
**ROC:** Como evoluciona la tasa Verdaderos Positivos (TP) vs Falsos Positivos (FP) en diferentes **thresholds**.

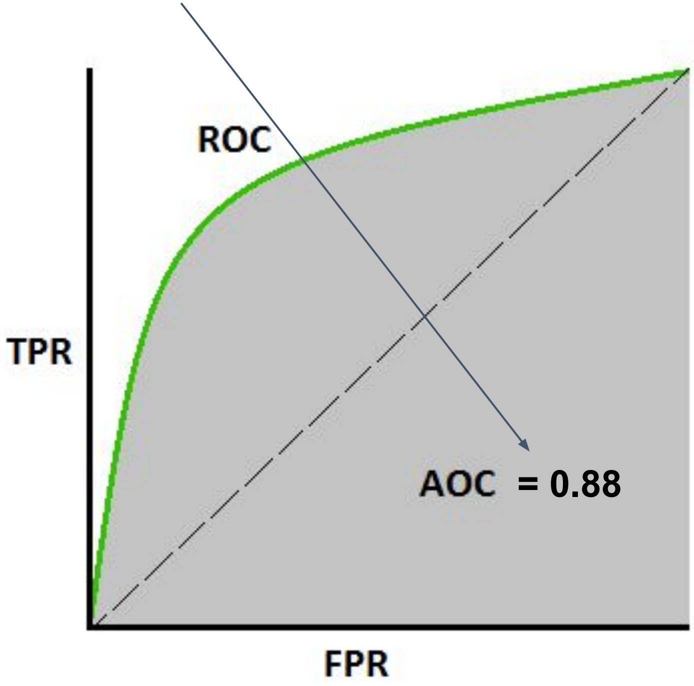
 Coche: 0.9 No coche 0.1  Coche: 0.7 No coche 0.3  Coche: 0.4 No coche 0.6



# Métricas de rendimiento

**ROC:** Como evoluciona la tasa Verdaderos Positivos (TP) vs Falsos Positivos (FP) en diferentes **thresholds**.
**AUC**: Area Under the Curve. Versión numérica que cuantifica el ROC



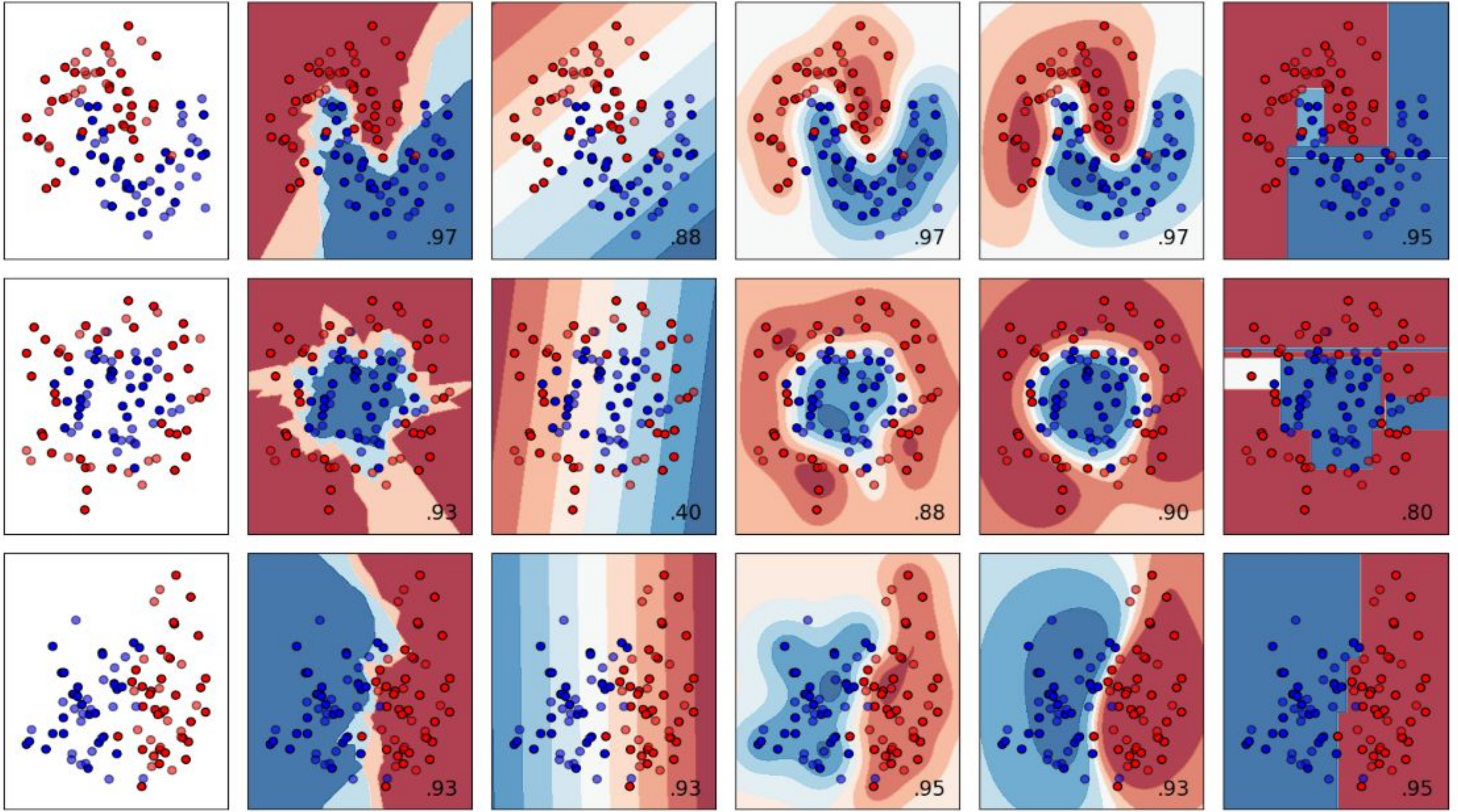



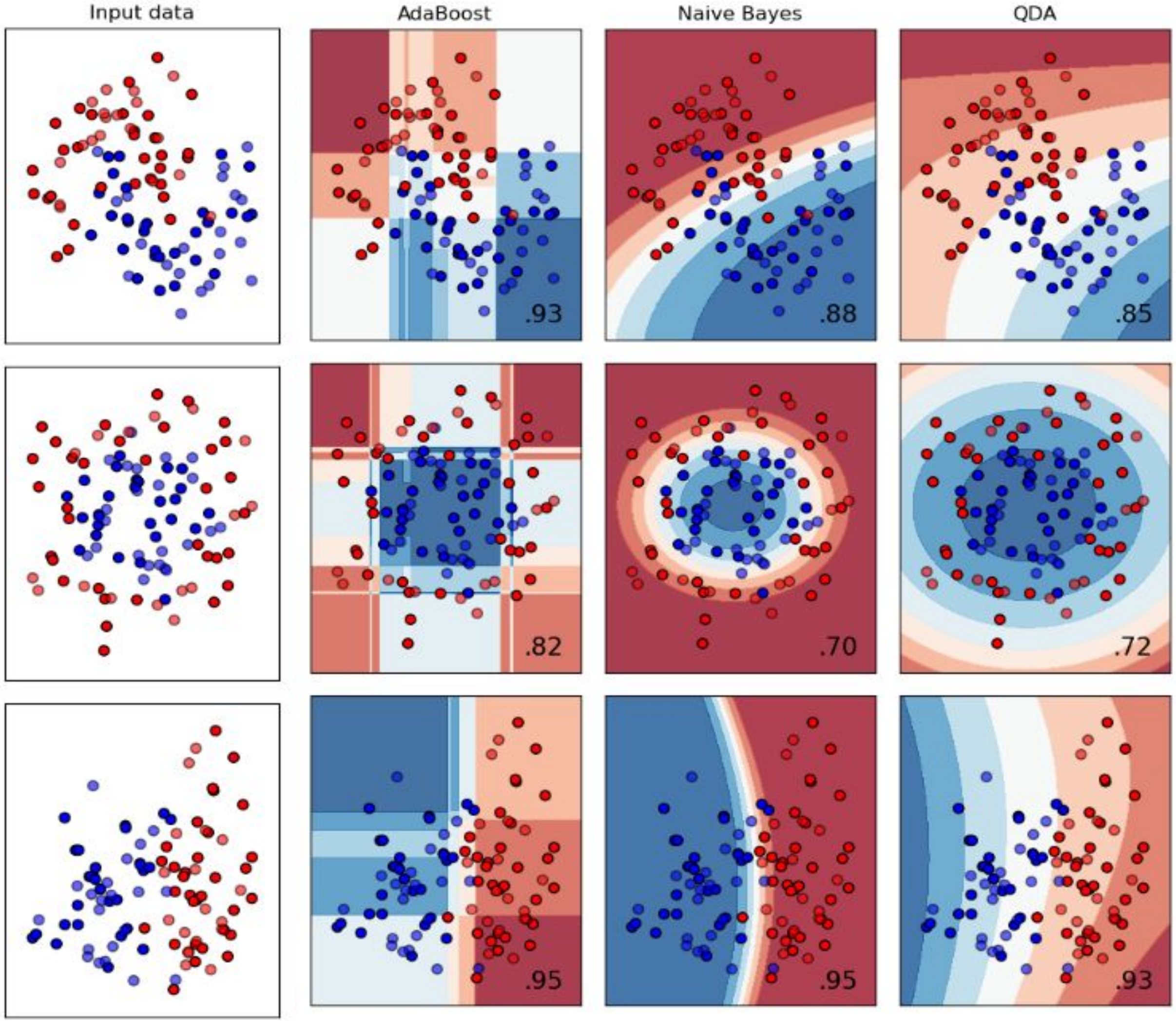
# Práctica clasificación

# Práctica

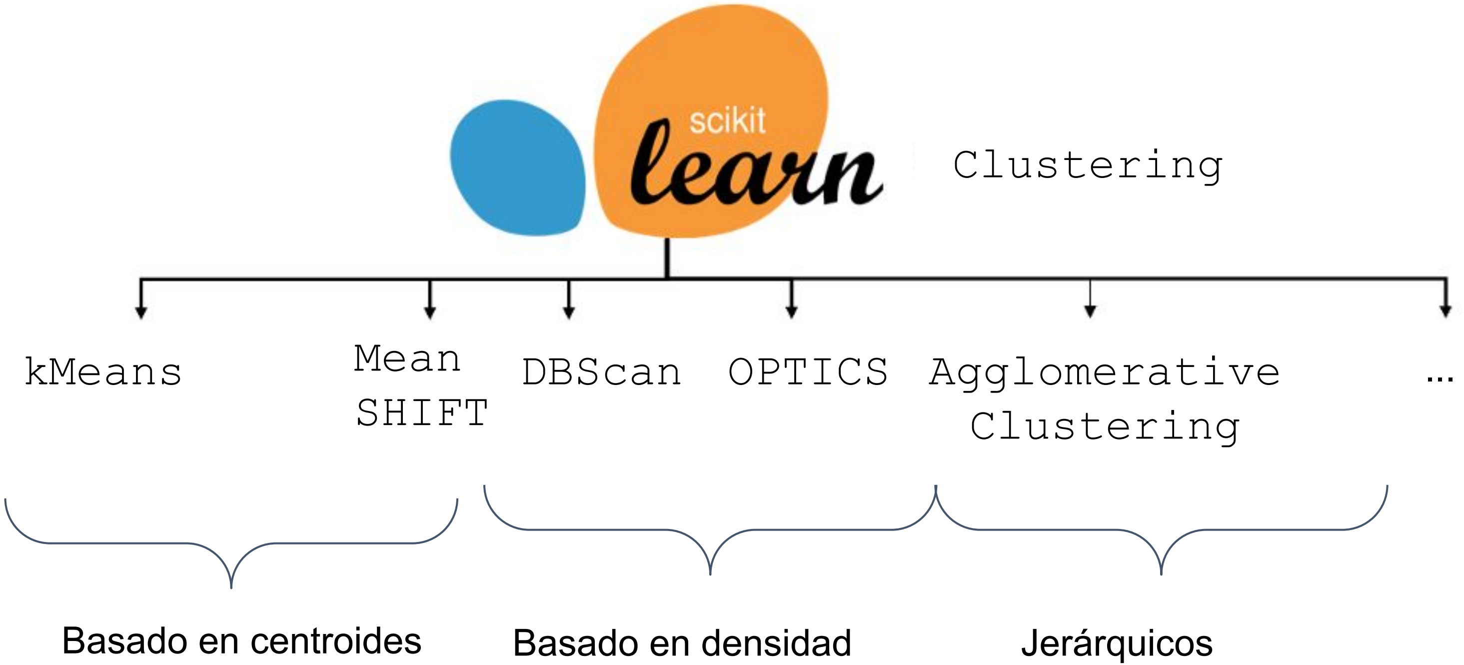
Vamos a poner en práctica todo el conocimiento que hemos adquirido en la sesión teórica, y vamos a :

1) "Dibujar" nuestro propio dataset utilizando https://drawdata.xyz/
2) Completar codigo para crear clasificadores
3) Utilizar algoritmos de clasificación diferentes
4) Visualizar las clasificaciones
5) Calcular métricas
6) Decidir cuál es el mejor para cada caso

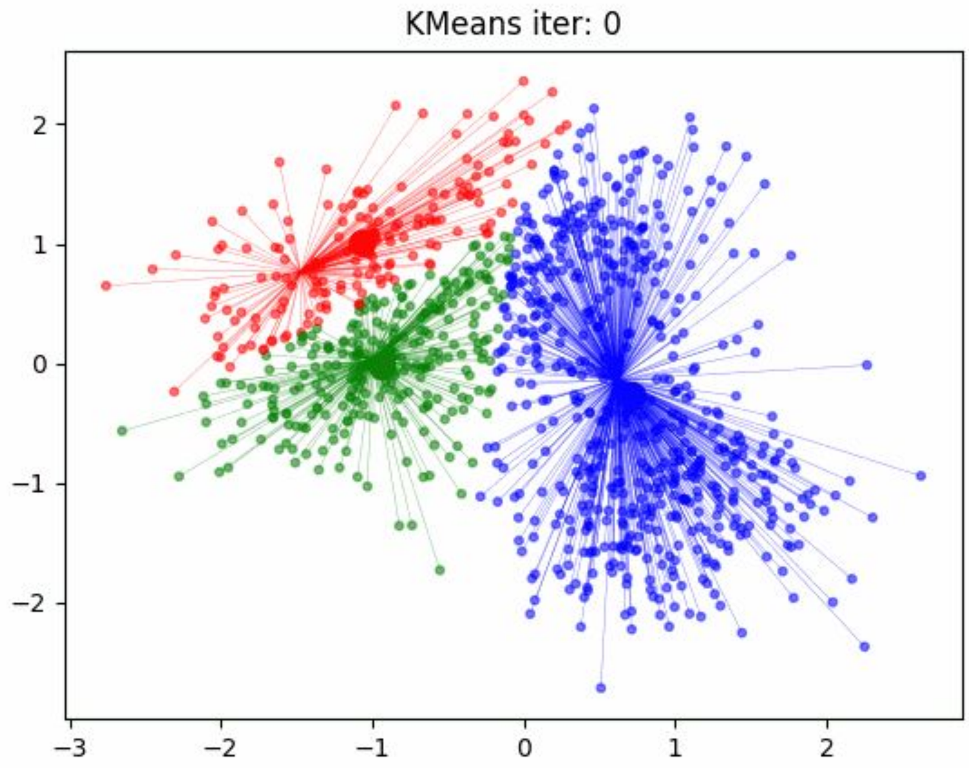

.97
.88
.97
.97
.95
.93
.40
.88
.90
.80
.93
.93
.95
.93
.95


Input data
AdaBoost
Naive Bayes
QDA
.93
.88
.85
.82
.70
.72
.95
.95
.93

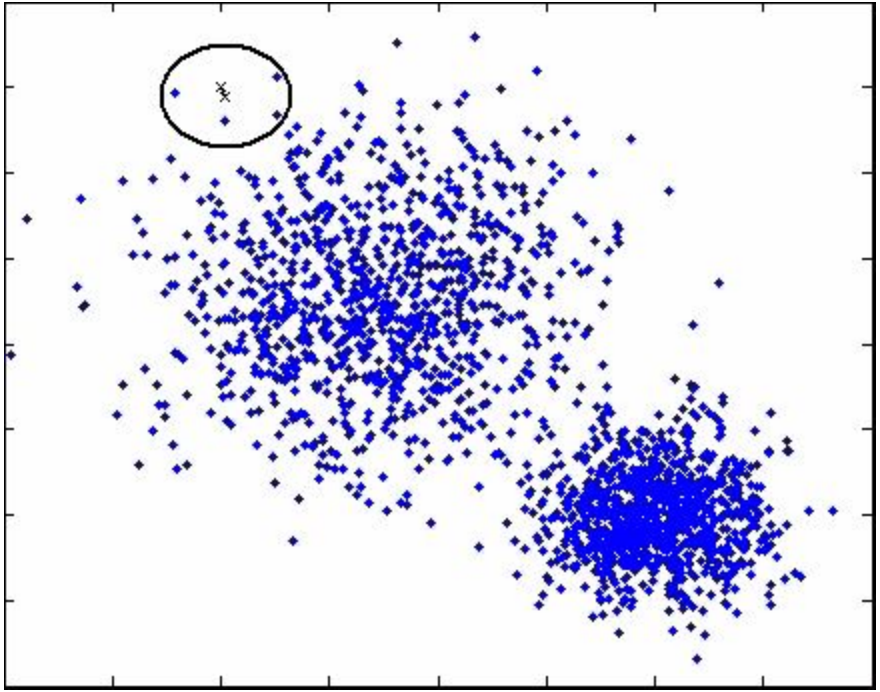
# Machine Learning supervisado: Clusterización


scikit
learn
Clustering
kMeans
Mean SHIFT
DBScan
OPTICS
Agglomerative Clustering
...
Basado en centroides
Basado en densidad
Jerárquicos

# k- Means: basada en centroides

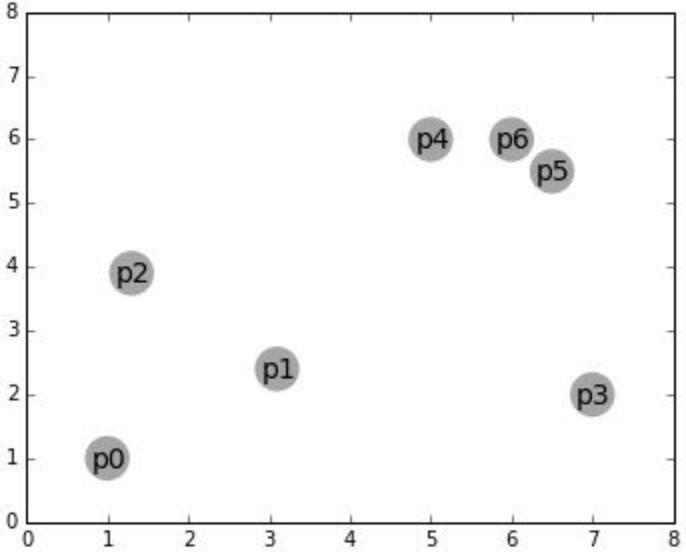


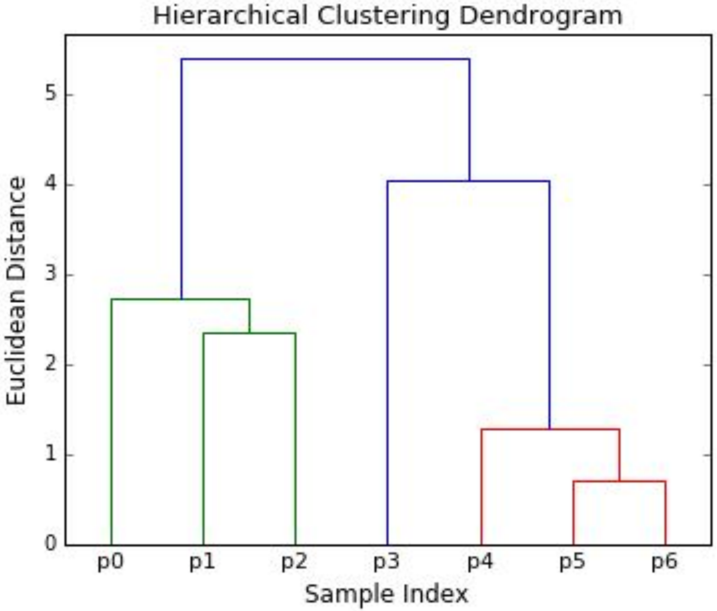
# Mean Shift



# DBScan: Basado en densidad

# Clustering jerárquico



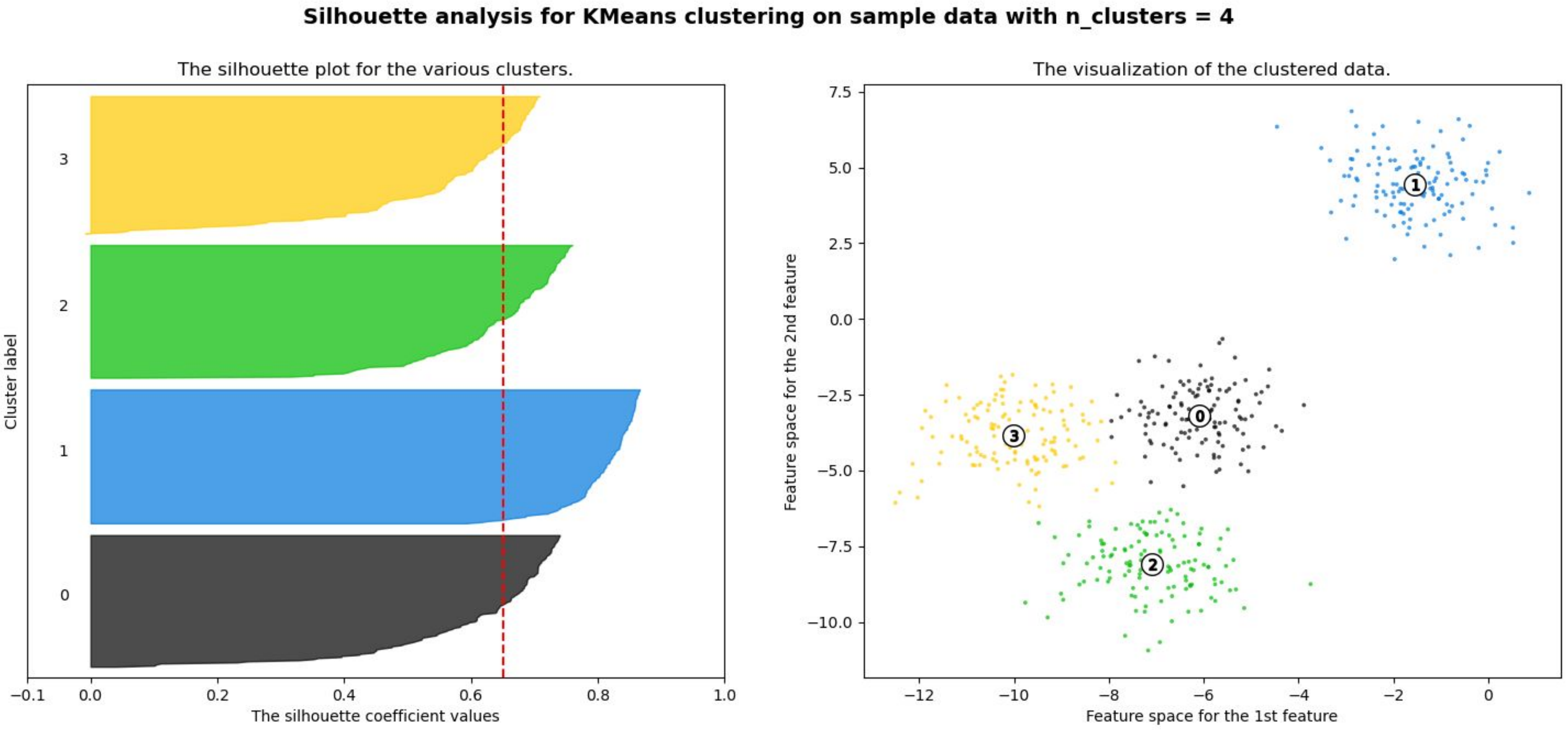



# Métricas en clusterización

En regresión y clasificación (aprendizaje supervisado), **los datos están etiquetados**, y podemos calcular la diferencia con el dato esperado.

En aprendizaje no supervisado, **no hay datos etiquetados**.

# Métricas: Silhouette (centroides y BIRCH)




Anchura: número de elementos en el cluster

Valores cercanos a 1: Alejamiento de los clústeres vecinos

# Otras métricas (índices)

**Sin datos etiquetados**
*Davis-Boulding*: El score se define como la media de la similitud de cada cluster con su cluster más similar. Entendemos similitud como la relación entre las distancias dentro del grupo y las distancias entre grupos.

Los grupos que estén más separados y menos dispersos darán como resultado una mejor puntuación. Valor óptimo: 0

**Con datos etiquetados**
*Mutual Information Score:* Similitud entre la información compartida entre los clusteres y los datos etiquetados.

Valor óptimo: 1

Perfect labelings are both homogeneous and complete, hence have score 1.0:

```
>>> from sklearn.metrics.cluster import adjusted_mutual_info_score
>>> adjusted_mutual_info_score([0, 0, 1, 1], [0, 0, 1, 1])
...
1.0
>>> adjusted_mutual_info_score([0, 0, 1, 1], [1, 1, 0, 0])
...
1.0
```

If classes members are completely split across different clusters, the assignment is totally in-complete, hence the AMI is null:

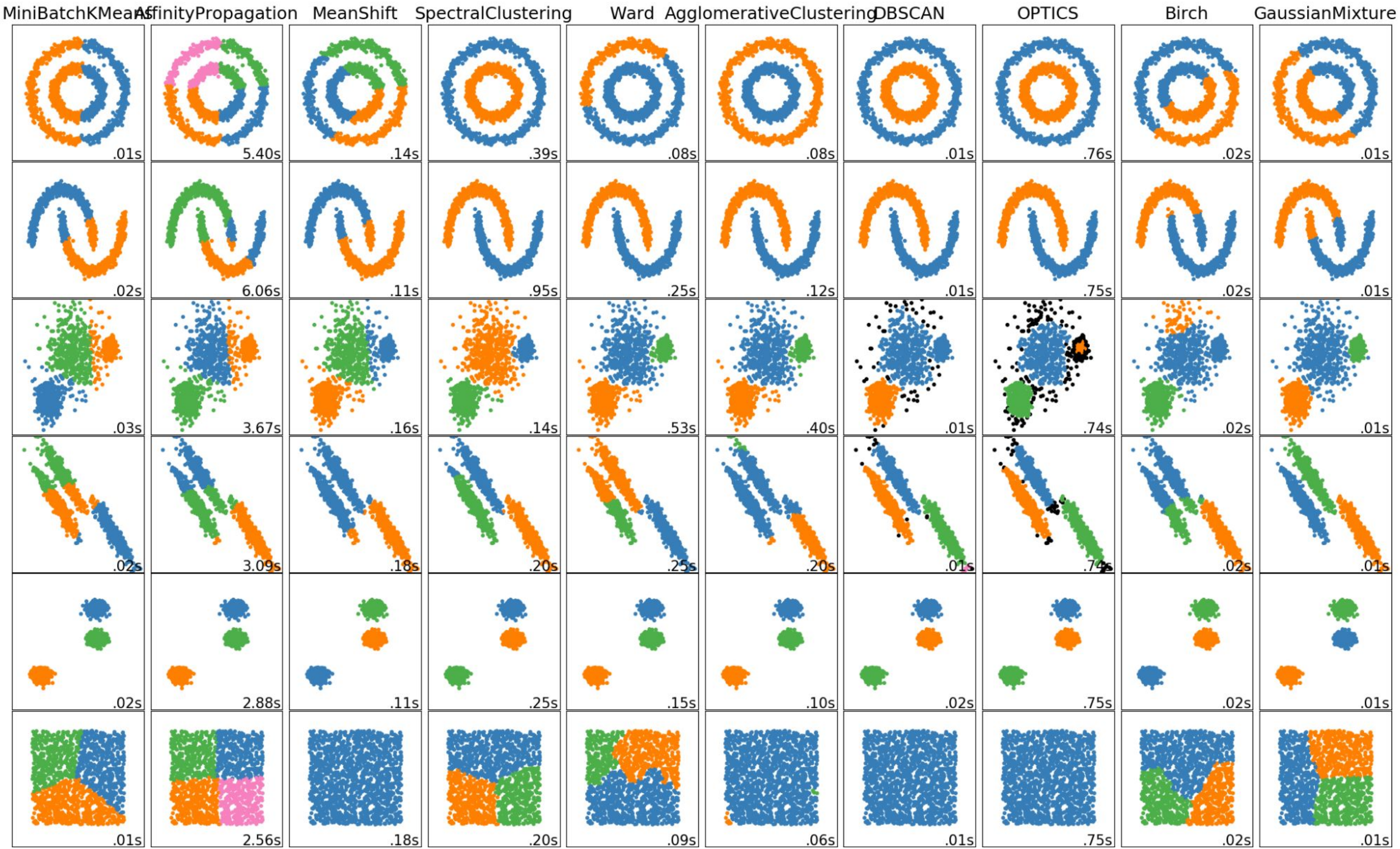
```
>>> adjusted_mutual_info_score([0, 0, 0, 0], [0, 1, 2, 3])
...
0.0
```

# Práctica clusterización

# Práctica

Vamos a poner en práctica todo el conocimiento que hemos adquirido en la sesión teórica, y vamos a :

1) “Dibujar” nuestro propio dataset utilizando https://drawdata.xyz/
2) Completar codigo para crear clusterizadores
3) Utilizar algoritmos de clusterización diferentes
4) Visualizar las clasificaciones
5) Calcular métricas
6) Decidir cuál es el mejor para cada caso


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s 5.40s .14s .39s .08s .08s .01s .76s .02s .01s
.02s 6.06s .11s .95s .25s .12s .01s .75s .02s .01s
.03s 3.67s .16s .14s .53s .40s .01s .74s .02s .01s
.02s 3.09s .18s .20s .25s .20s .01s .74s .02s .01s
.02s 2.88s .11s .25s .15s .10s .02s .75s .02s .01s
.01s 2.56s .18s .20s .09s .06s .01s .75s .02s .01s