Introducción al Machine Learning

It's tough to make predictions, especially about the future.
-Yogi Berra

Dra. Mariana Pereyra

¿Qué es el machine learning (ML)?

El aprendizaje automático es el proceso que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas.

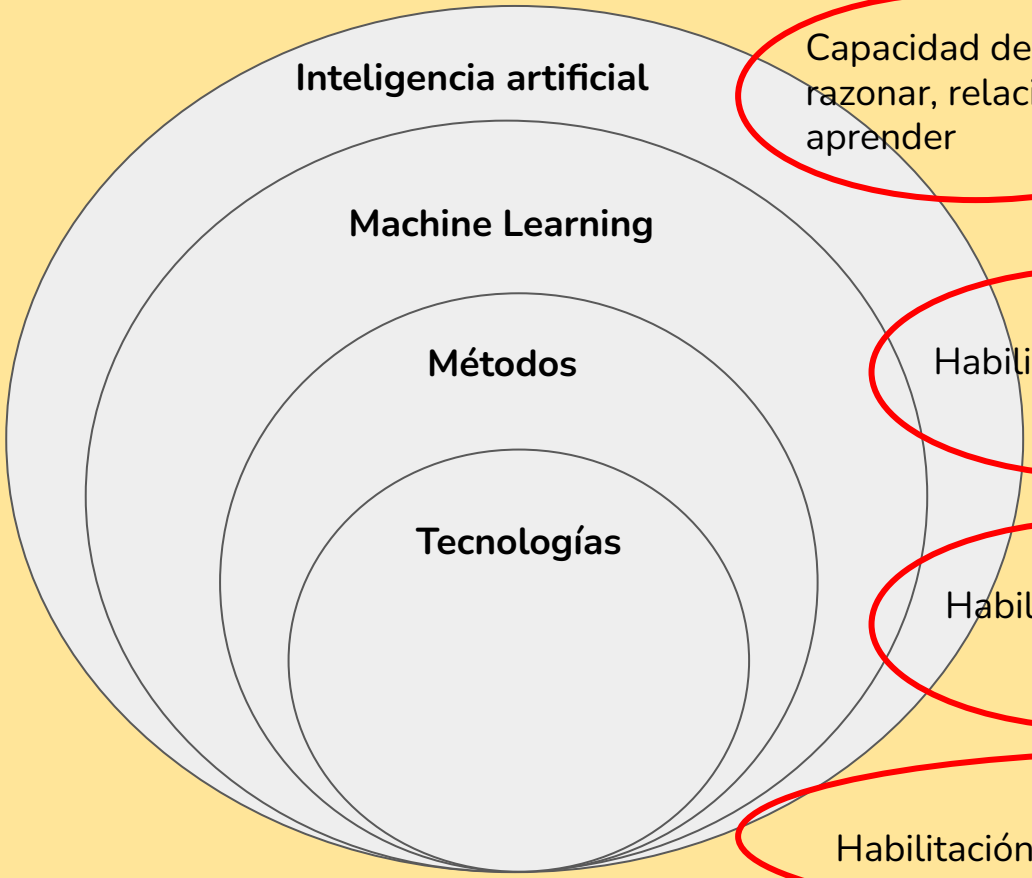
A.L Samuel (1959)

Se dice que un programa de computación aprende de la experiencia E con respecto a una tarea T y alguna medida de rendimiento P, si es que el rendimiento en T, medido por P, mejora con la experiencia

E.T.M Mitchell

https://www.youtube.com/watch?v=f_uwKZIAeM0

¿Cómo se relacionan la IA y el ML?



Capacidad de pensar, razonar, relacionar y aprender

- Visión por computadora
- Procesamiento del lenguaje natural
- Reconocimiento de voz
- Robótica y movimiento
- Planificación y optimización
- Simulación de conocimiento

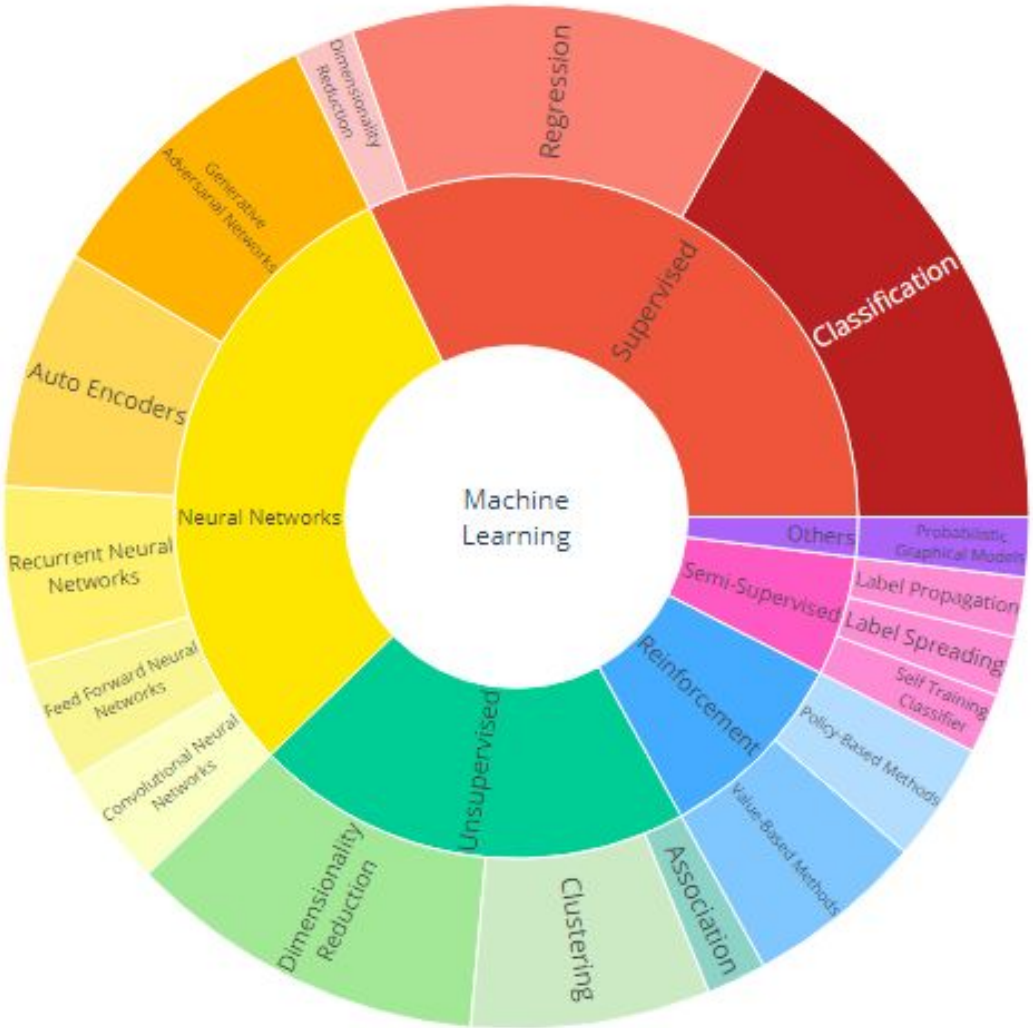
Habilidad de aprender

Aprendizaje supervisado,
Aprendizaje no supervisado
A. por refuerzo

Habilidad de razonar

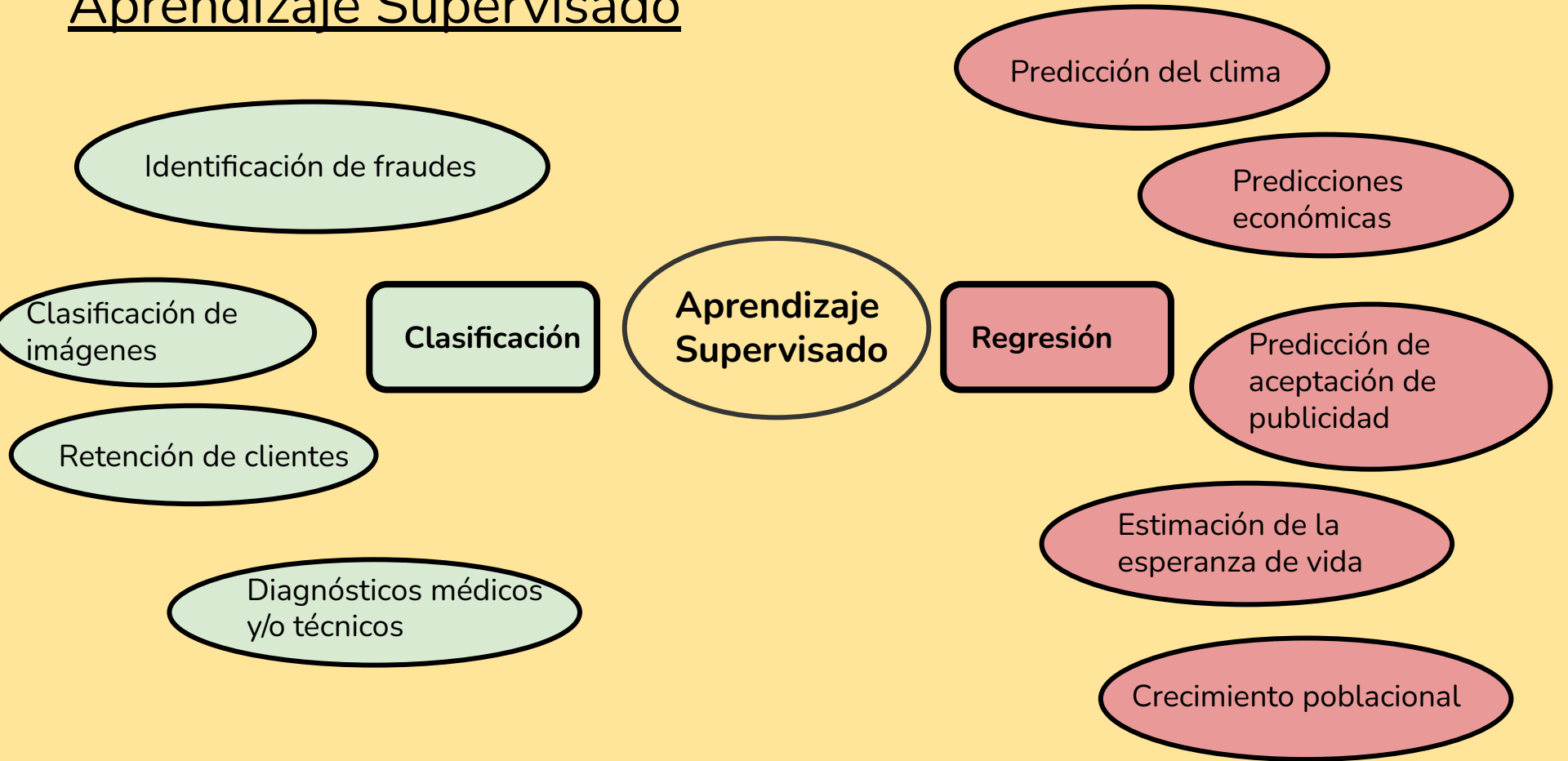
Regresión
Árboles de decisión
ANN
SVM
Ensemble Learning

Habilitación física de hardware y software



¿Qué tipos de aprendizaje automático o machine learning hay?

Aprendizaje Supervisado



Etapas en la aplicación del Aprendizaje Supervisado

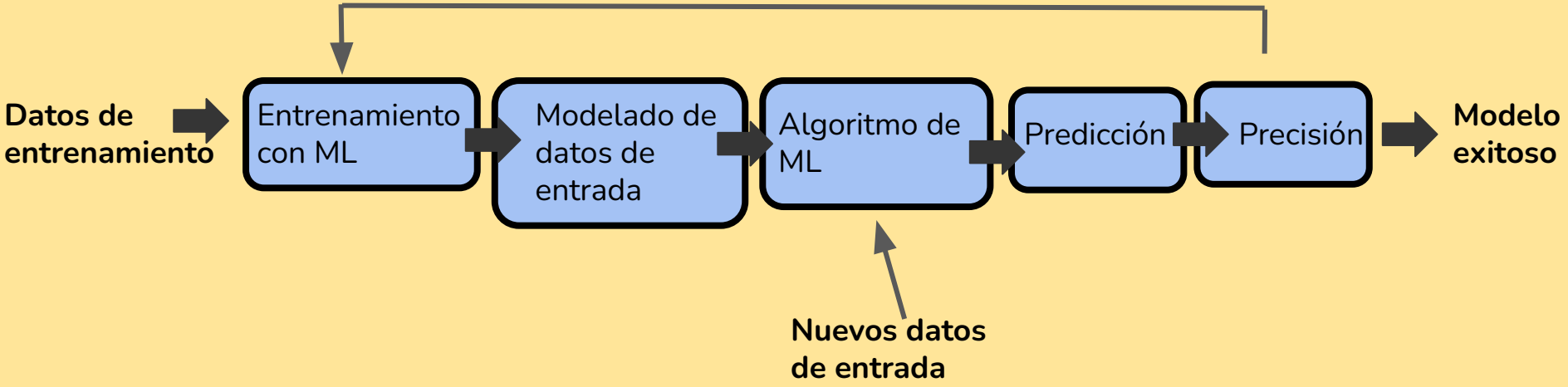
- Definición del problema
- Análisis detallado de los datos
- Definición de métricas de éxito y fracaso
- División del conjunto de datos (entrenamiento y test)
- Procesar los datos
- Diseñar un modelo "blando" para ver que ande
- Ajustar los parámetros del modelo a lo deseado
- Analizar la evolución del entrenamiento
- Entrenar con los requerimientos definidos
- Testear el sistema
- En caso de no obtener resultados aceptables, volver al principio

Descripción del problema: Aprendizaje Supervisado

Datos: Se dispone de un conjunto de registros (o ejemplos, o instancias) descritos por n atributos: A1, A2, ... , An y cada instancia está anotada con una etiqueta, pudiendo ser una clase o un valor numérico.

Objetivo: Aprender un modelo (o función) a partir de los datos, buscando predecir sus etiquetas a partir de los atributos. Este modelo puede ser utilizado para predecir las etiquetas de nuevos registros sin anotar.

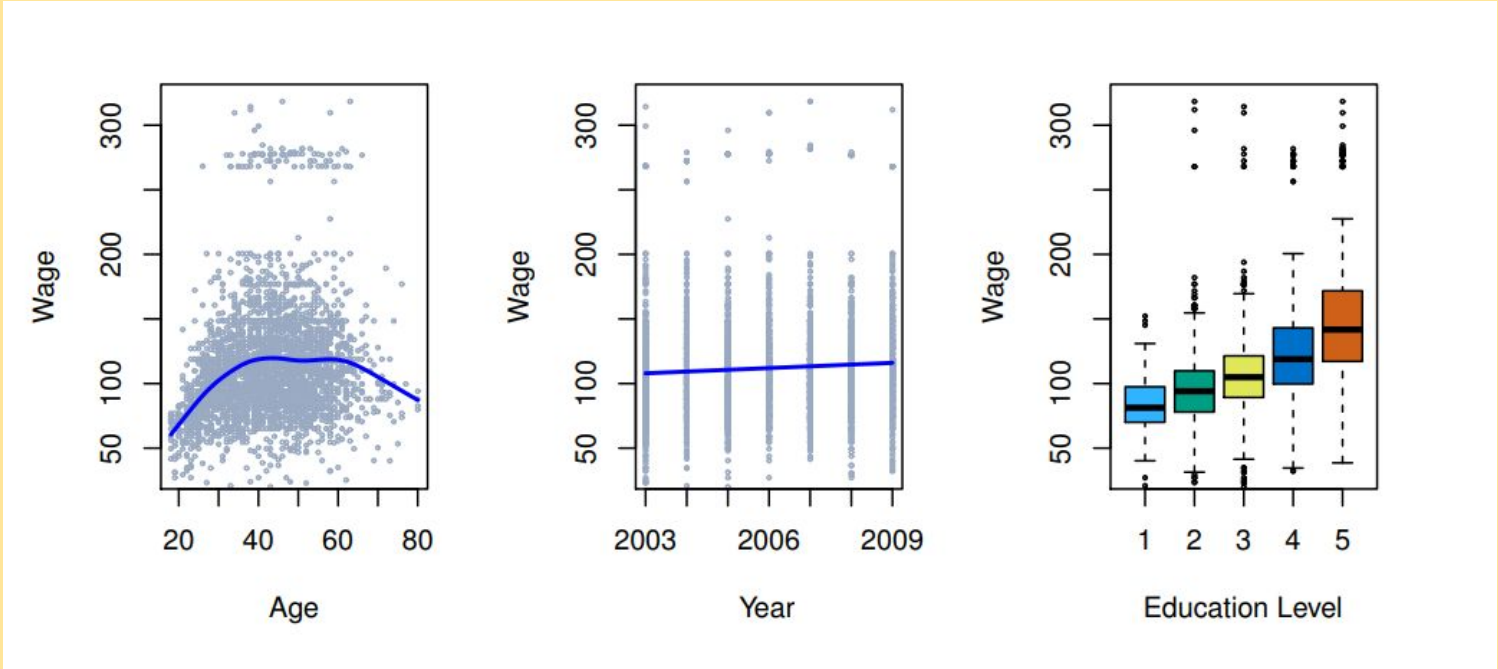
¿Cómo procedemos?



Aprendizaje Supervisado

Regresión

Dados (x1,y1), (x2, y2), ..., (xn, yn)
Aprender una f(x) que permita predecir y a partir de x
Si y ϵ ℝn: Es un problema de regresión.

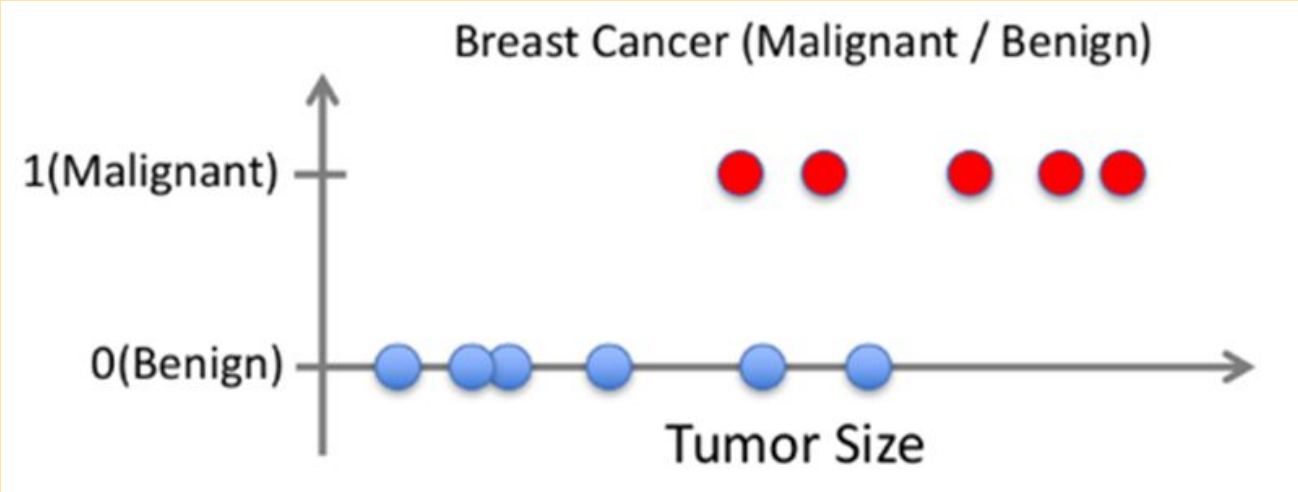


Clasificación

Dados (x1,y1), (x2, y2), ..., (xn, yn)

Aprender una f(x) que permita predecir y a partir de x

Si y es categórica: Es un problema de clasificación.

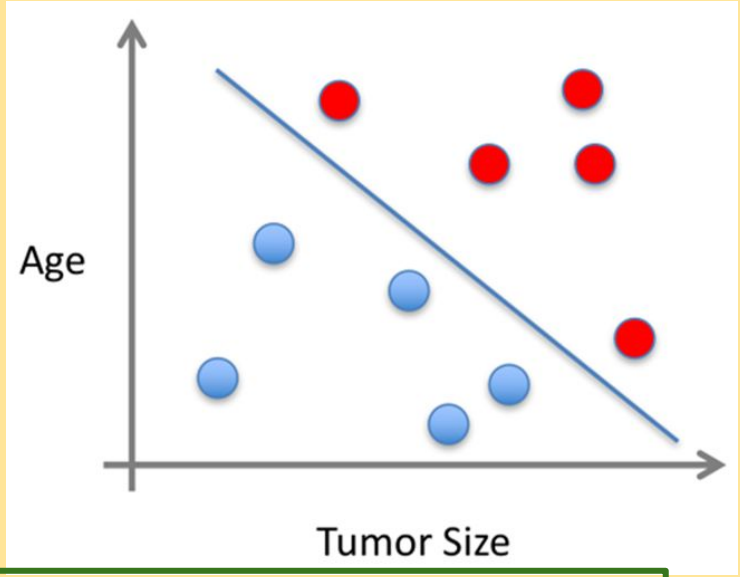


Aprendizaje Supervisado

La variable x puede ser multidimensional.

Cada dimensión corresponde a un atributo:

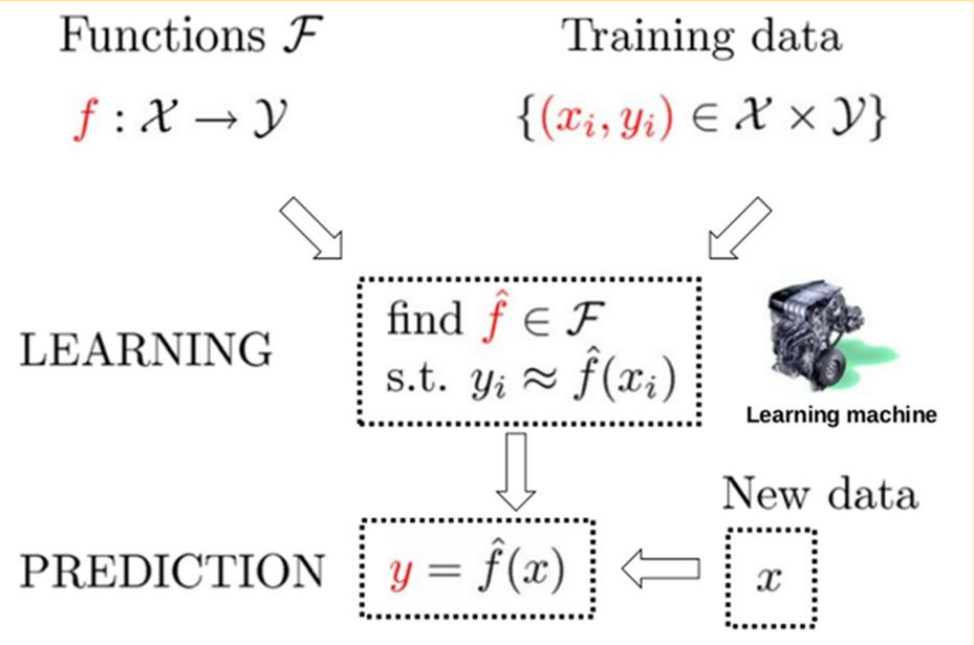
- Edad del paciente
- Tamaño del tumor
- Uniformidad en la forma de la célula
- Entre otros



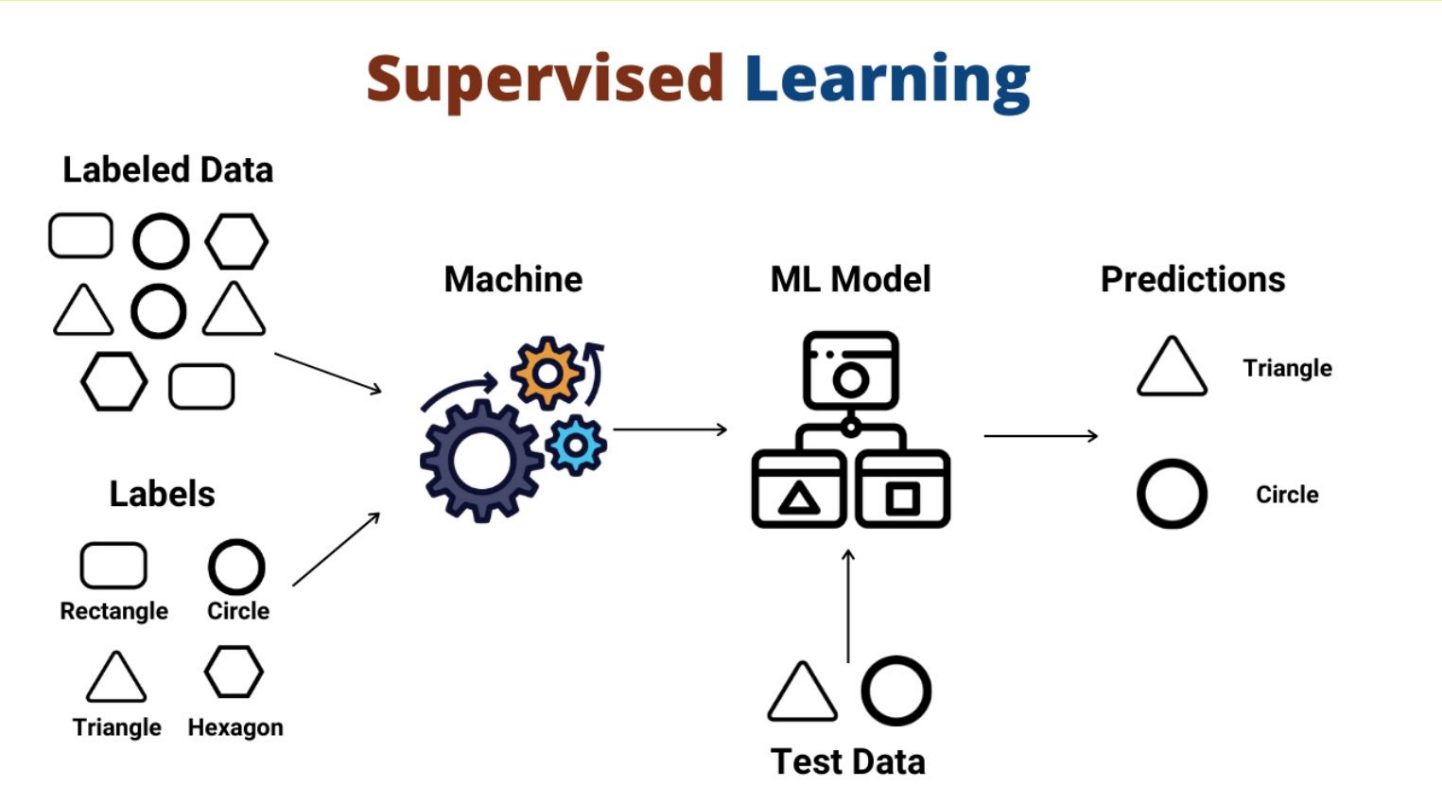
La regresión busca "acercar" los datos a una función (lineal, polinomial, etc.)

La clasificación busca separar los datos mediante ciertos "bordes".

Aprendizaje supervisado



Datos etiquetados



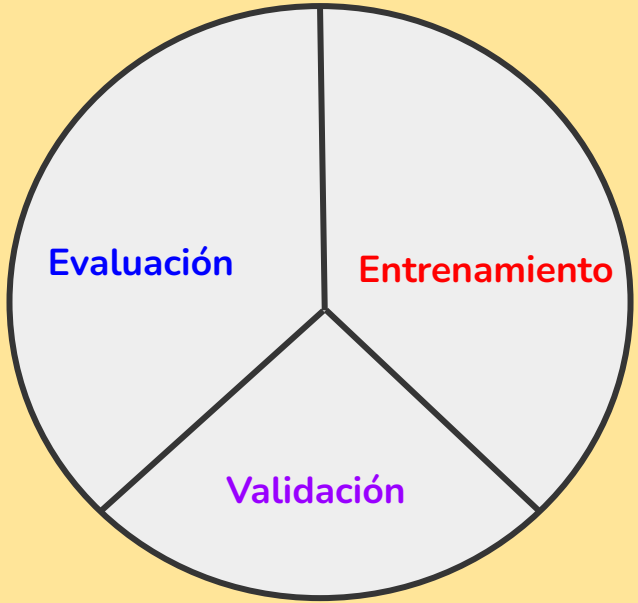
Elección de hiperparámetros

Se tiene la base de datos con toda los datos. Dividir el conjunto total de ejemplos en tres subconjuntos:

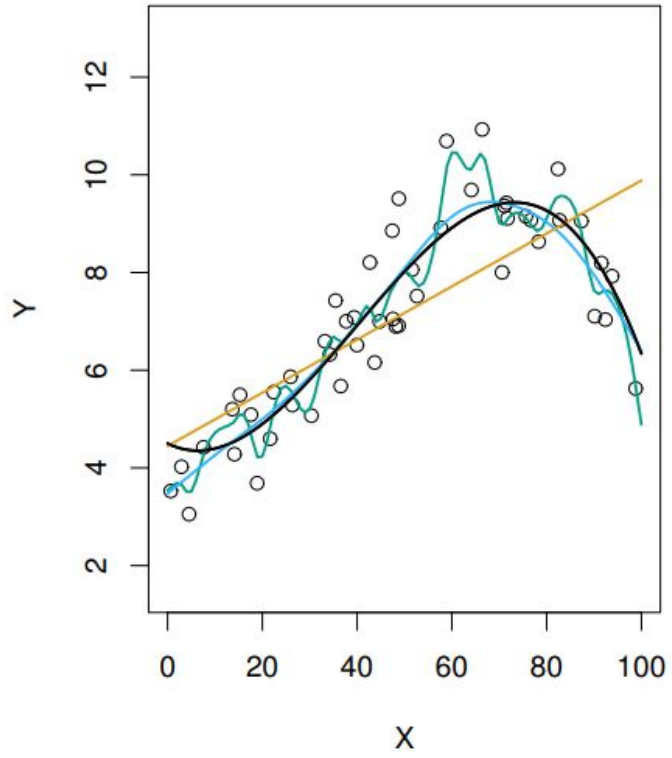
Entrenamiento: aprendizaje de variables del modelo.

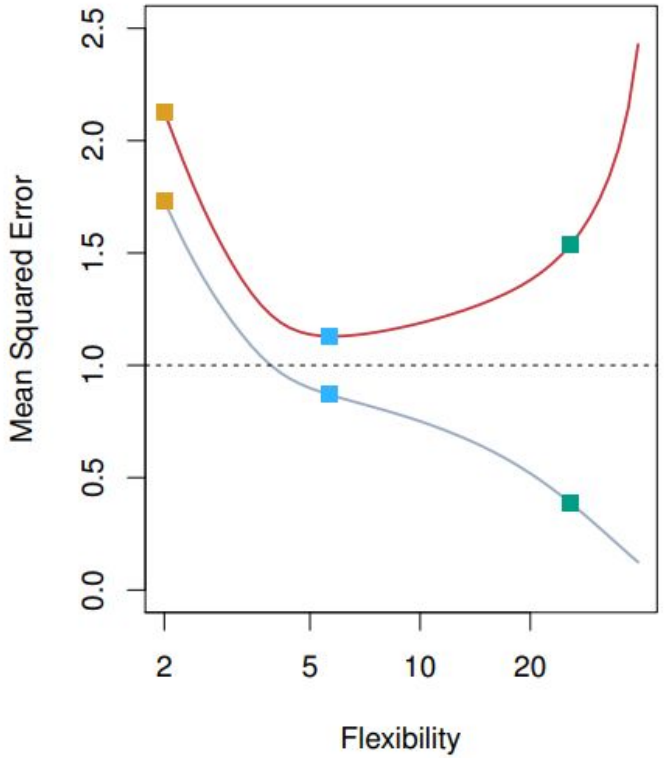
Validación: ajuste/elección de hiperparámetros.

Evaluación: estimación final del desempeño del modelo entrenado (y con hiperparámetros elegidos adecuadamente).



Recordando...

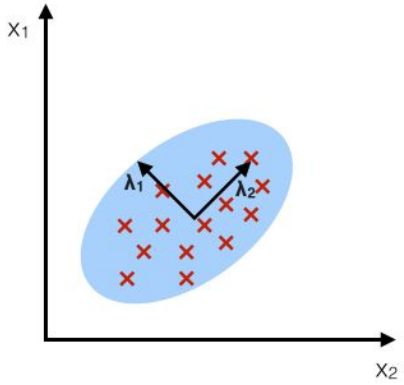

Y
2
4
6
8
10
12
0
20
40
60
80
100
X


Mean Squared Error
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility

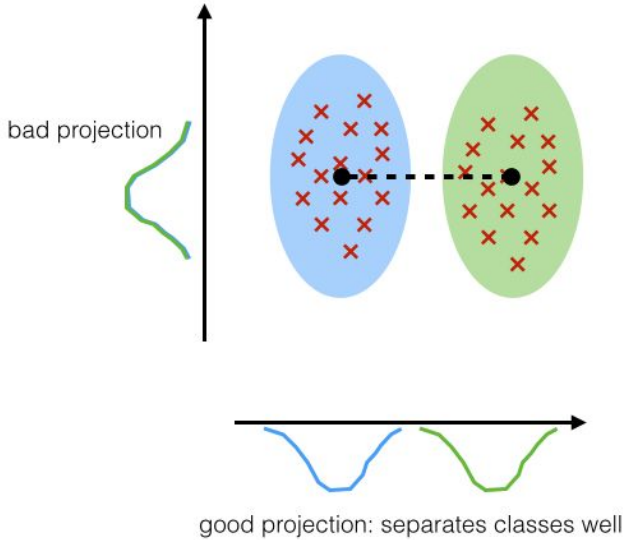

Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Deep Learning
High
Low
Interpretability
Low
High
Flexibility

Análisis Linear Discriminante (LDA)

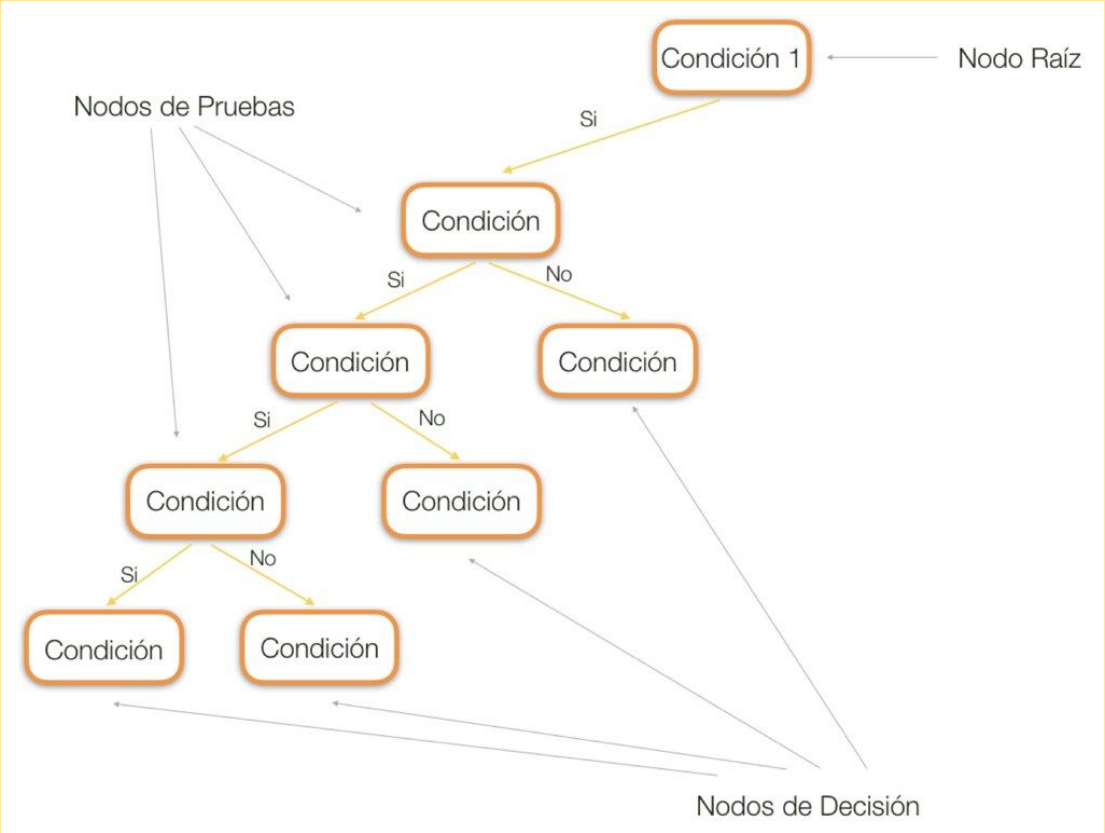
PCA:
component axes that maximize the variance

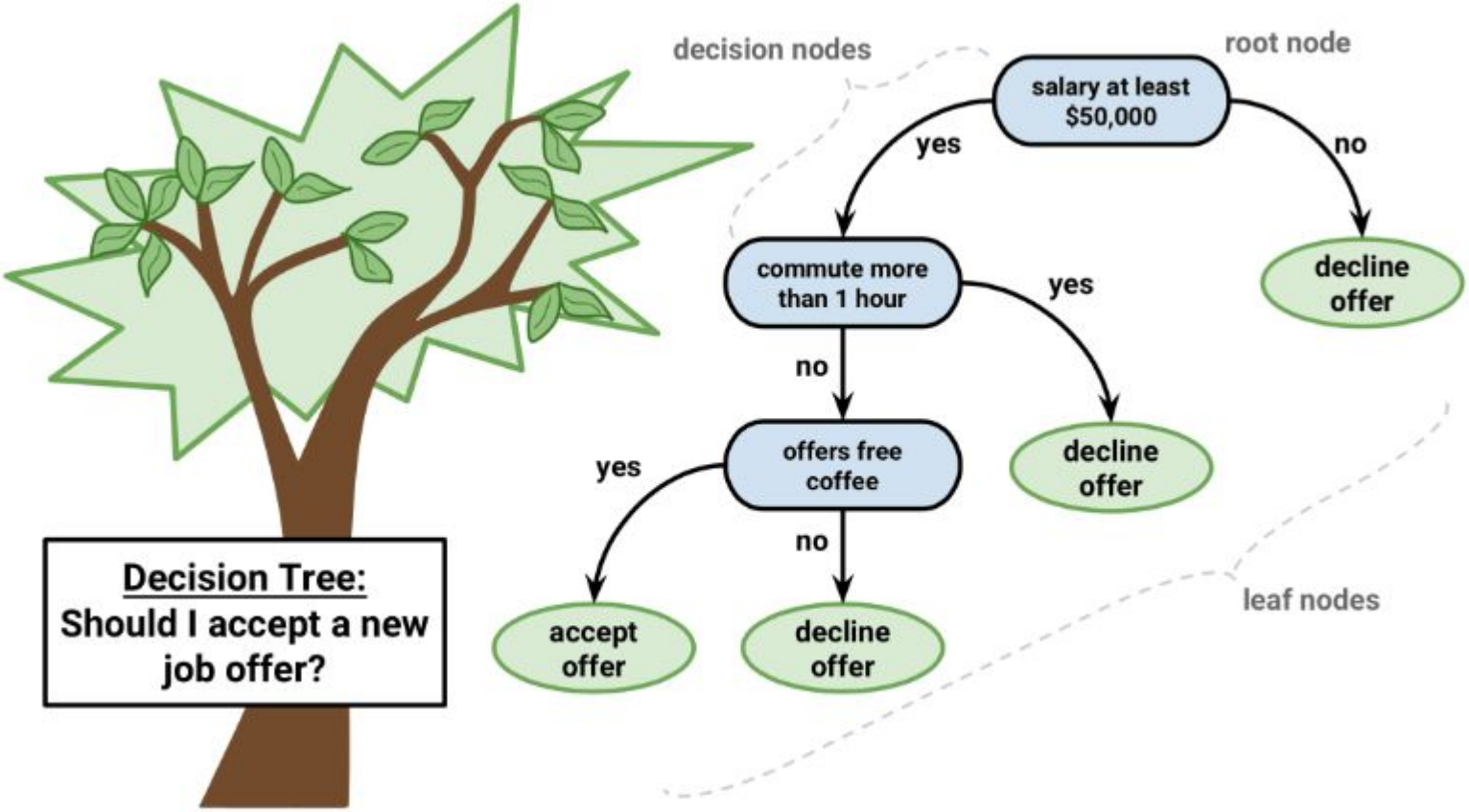


LDA:
maximizing the component axes for class-separation



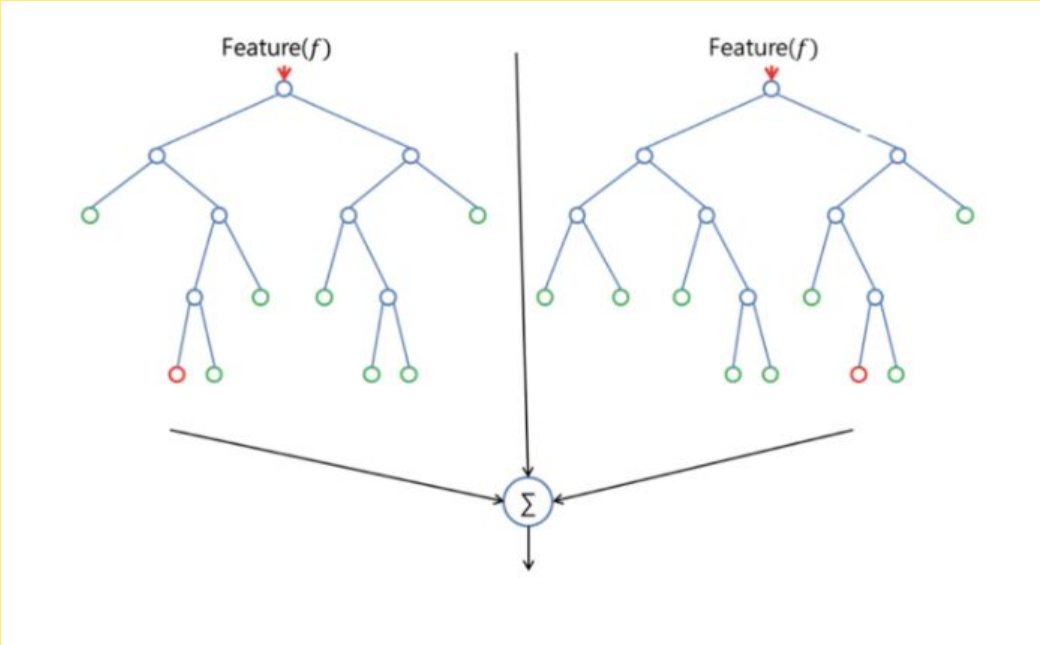
Árboles de clasificación




Decision Tree:
Should I accept a new job offer?
decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
yes
decline offer
no
offers free coffee
yes
decline offer
no
accept offer
decline offer
leaf nodes

Random forests

Definición simple: los Random forests construyen múltiples árboles de decisión y los fusionan para obtener una predicción más precisa y estable.



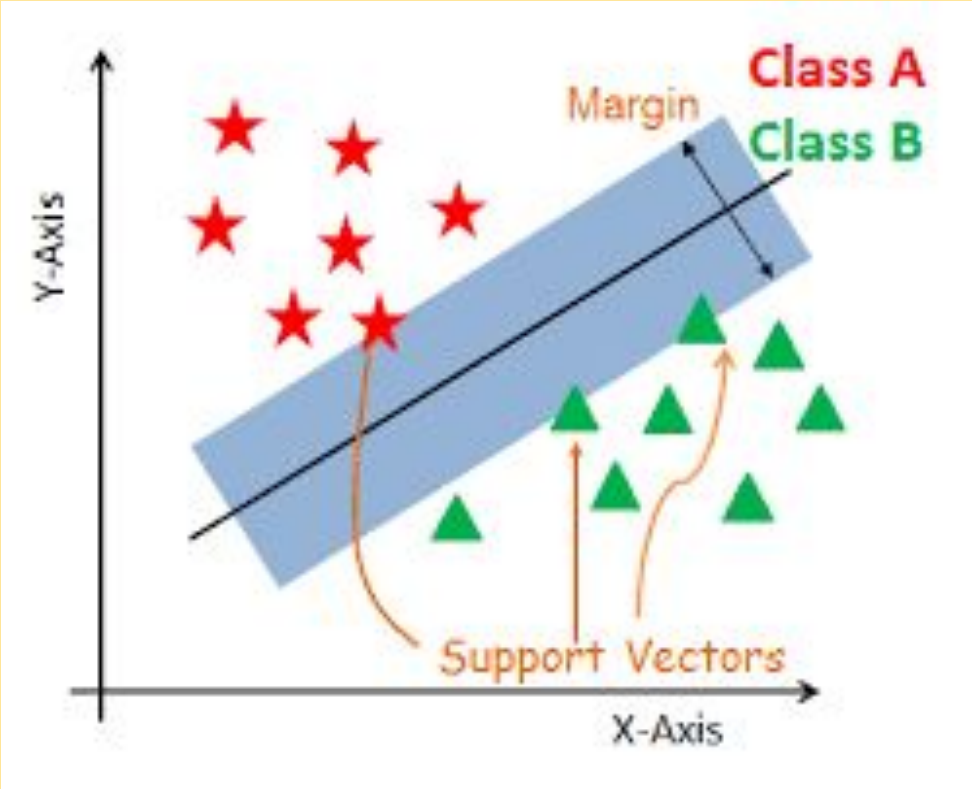
Super Vector Machines (SVM)

Es un algoritmo que busca separar los datos mediante la mejor frontera de decisión. Esta frontera de decisión es conocida como hiperplano.

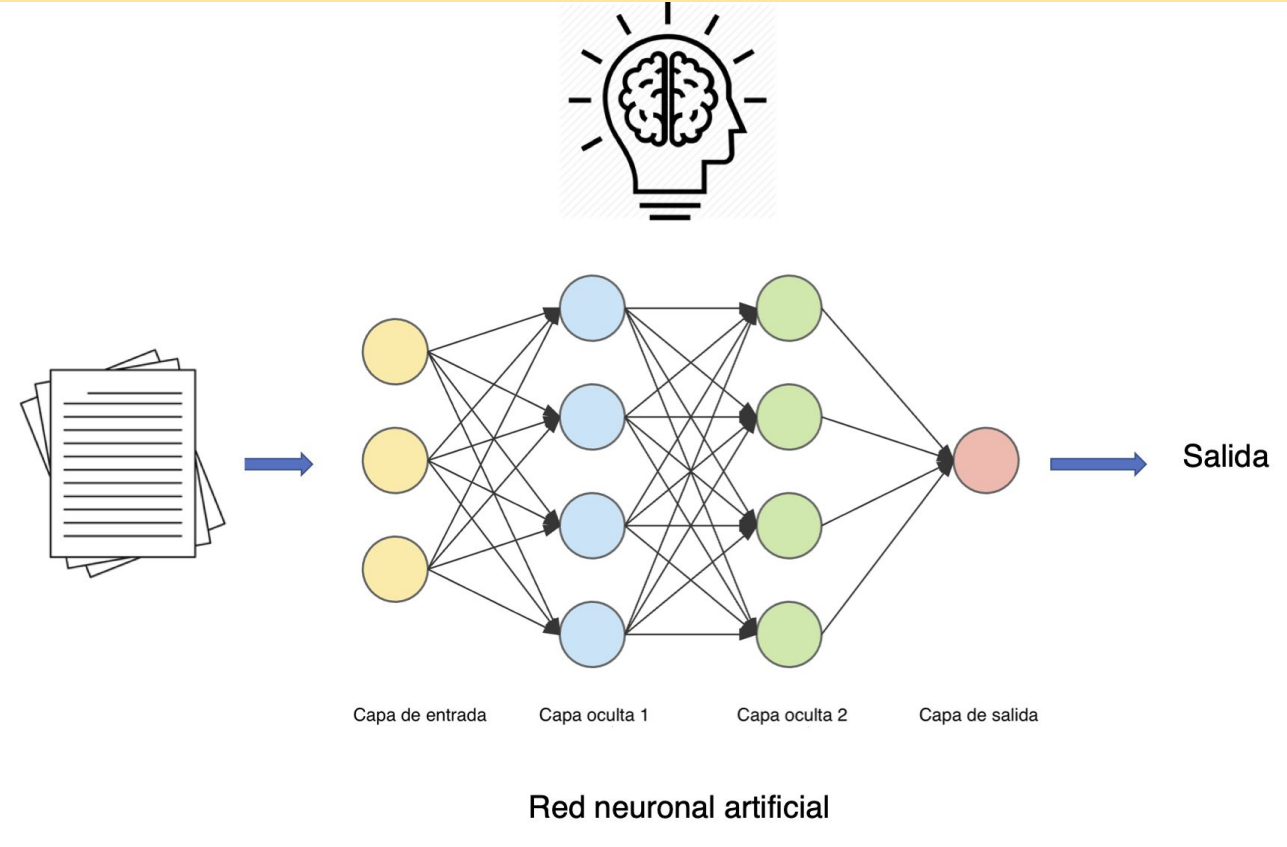
- En este caso, “mejor” se refiere a aquella que esté lo más separada posible de los puntos más cercanos a ella. Estos puntos son conocidos como vectores de soporte, y el espacio entre ellos y el hiperplano se conoce como margen.

- En términos más técnicos, un algoritmo de SVM encuentra el hiperplano que devuelva el mayor margen entre sí mismo y los vectores de soporte.

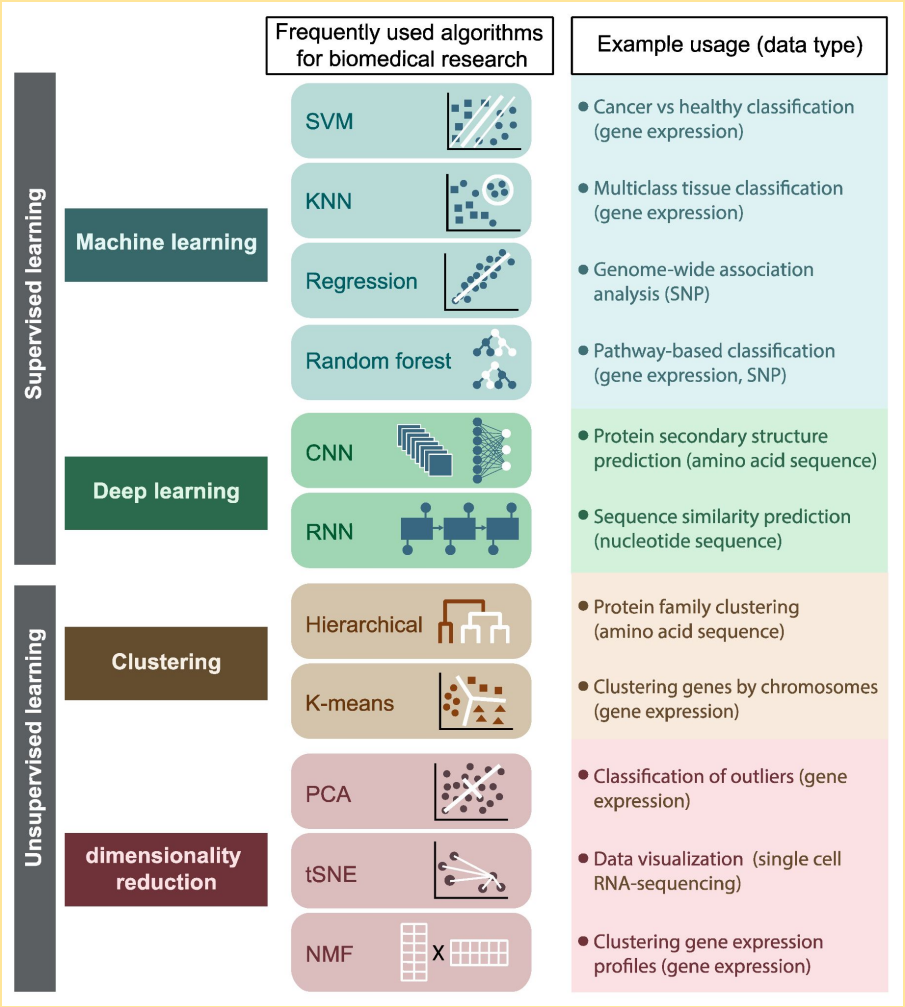
- Este tipo de clasificador a veces es conocido como “clasificador por márgenes” (margin classifier).


Class A
Class B
Margin
Y-Axis
X-Axis
Support Vectors

Redes neuronales (Deep Learning)



FIN


Frequently used algorithms for biomedical research
Example usage (data type)
Supervised learning
Machine learning
SVM
KNN
Regression
Random forest
Deep learning
CNN
RNN
Unsupervised learning
Clustering
Hierarchical
K-means
dimensionality reduction
PCA
tSNE
NMF
X
Cancer vs healthy classification (gene expression)
Multiclass tissue classification (gene expression)
Genome-wide association analysis (SNP)
Pathway-based classification (gene expression, SNP)
Protein secondary structure prediction (amino acid sequence)
Sequence similarity prediction (nucleotide sequence)
Protein family clustering (amino acid sequence)
Clustering genes by chromosomes (gene expression)
Classification of outliers (gene expression)
Data visualization (single cell RNA-sequencing)
Clustering gene expression profiles (gene expression)