

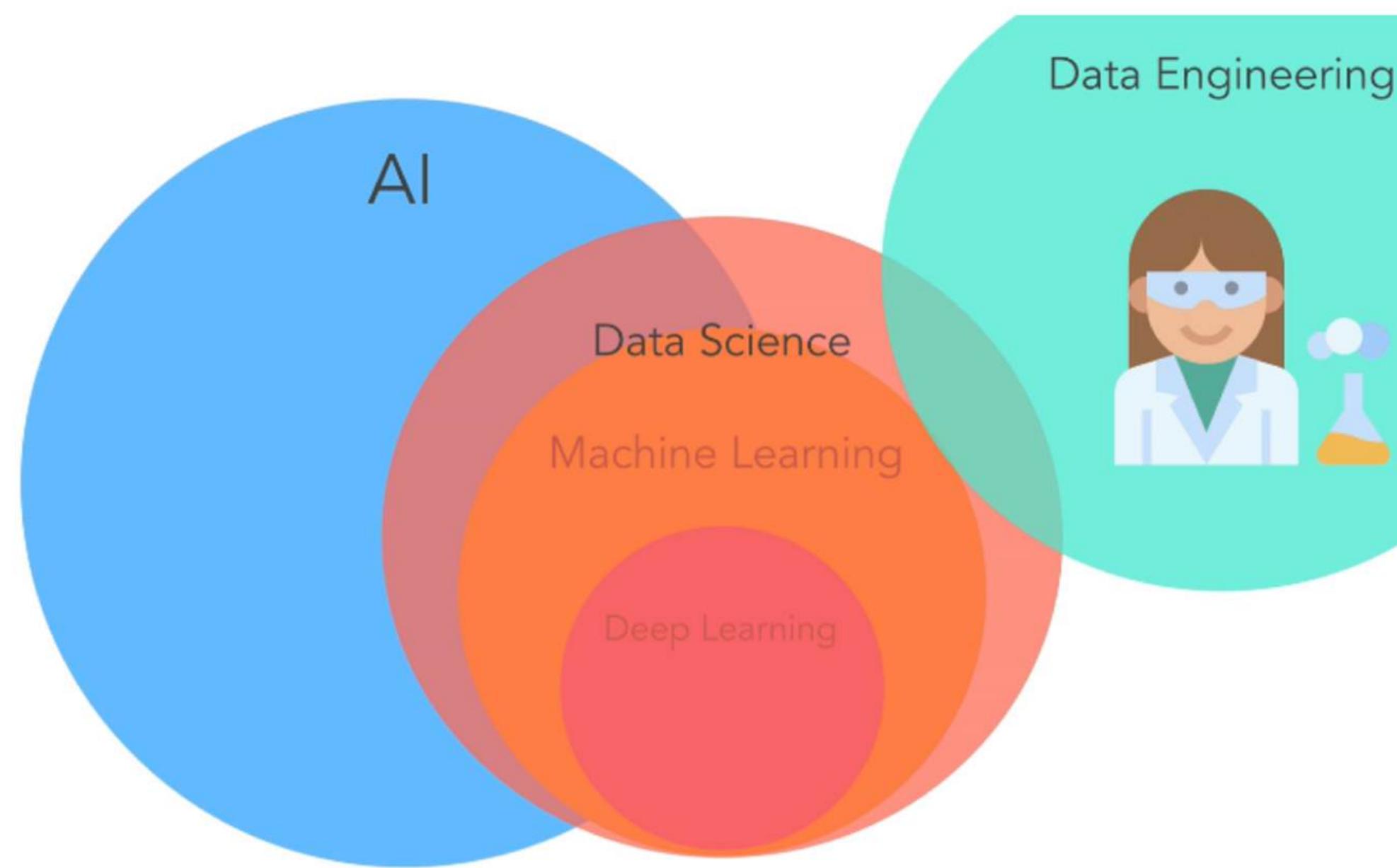
MACHINE LEARNING

MTRO. ALFONSO GREGORIO RIVERO DUARTE



MODELOS DE MACHINE LEARNING

¿QUÉ ES MACHINE LEARNING?



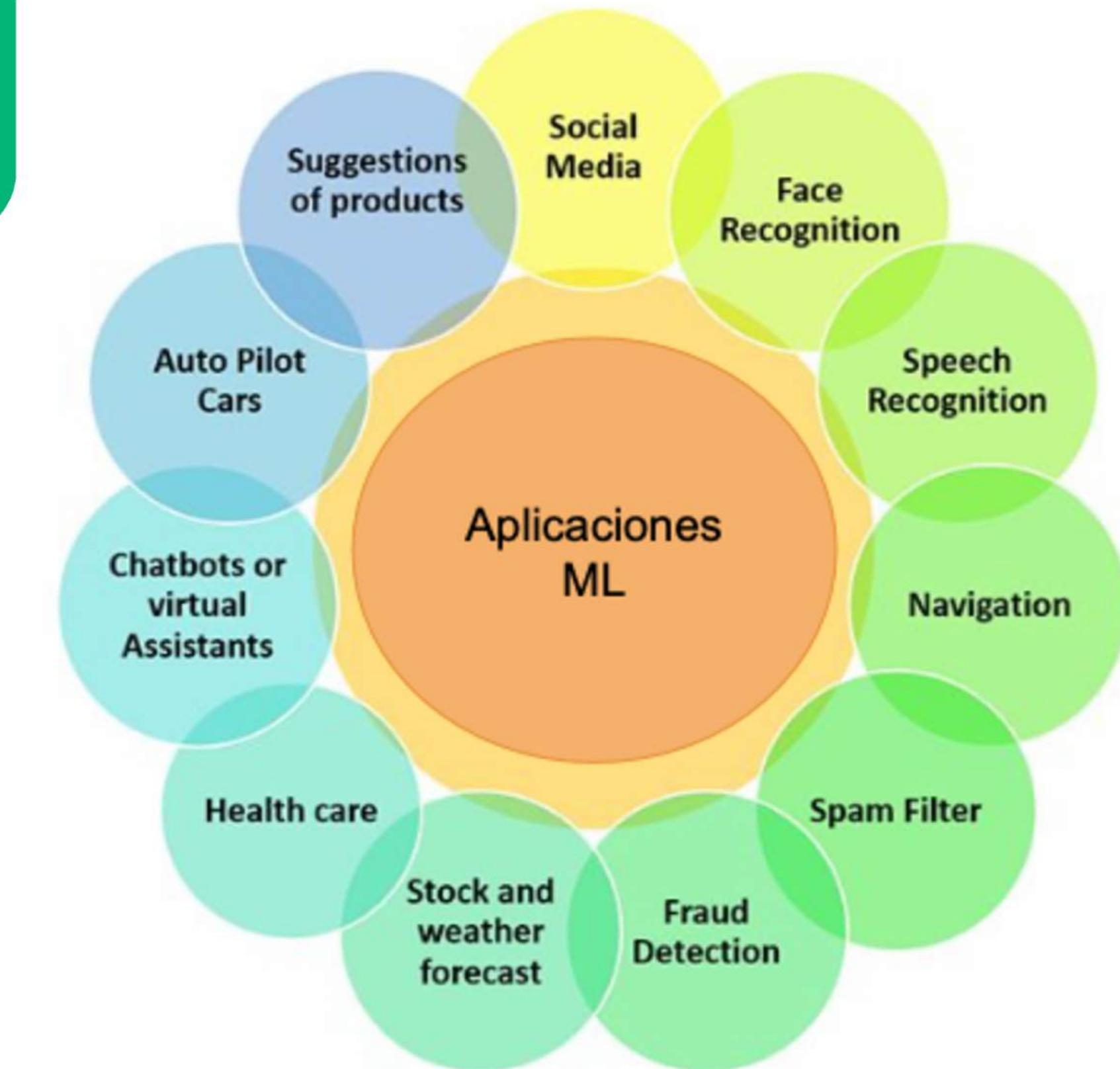
Machine Learning es un subconjunto de inteligencia artificial (IA)

- Se centra en enseñar a las computadoras a aprender de los datos y mejorar con la experiencia – en lugar de ser explícitamente programadas para hacerlo–
- En machine learning, los algoritmos se capacitan para encontrar patrones y correlaciones en grandes datasets y para tomar las mejores decisiones y previsiones basadas en ese análisis.

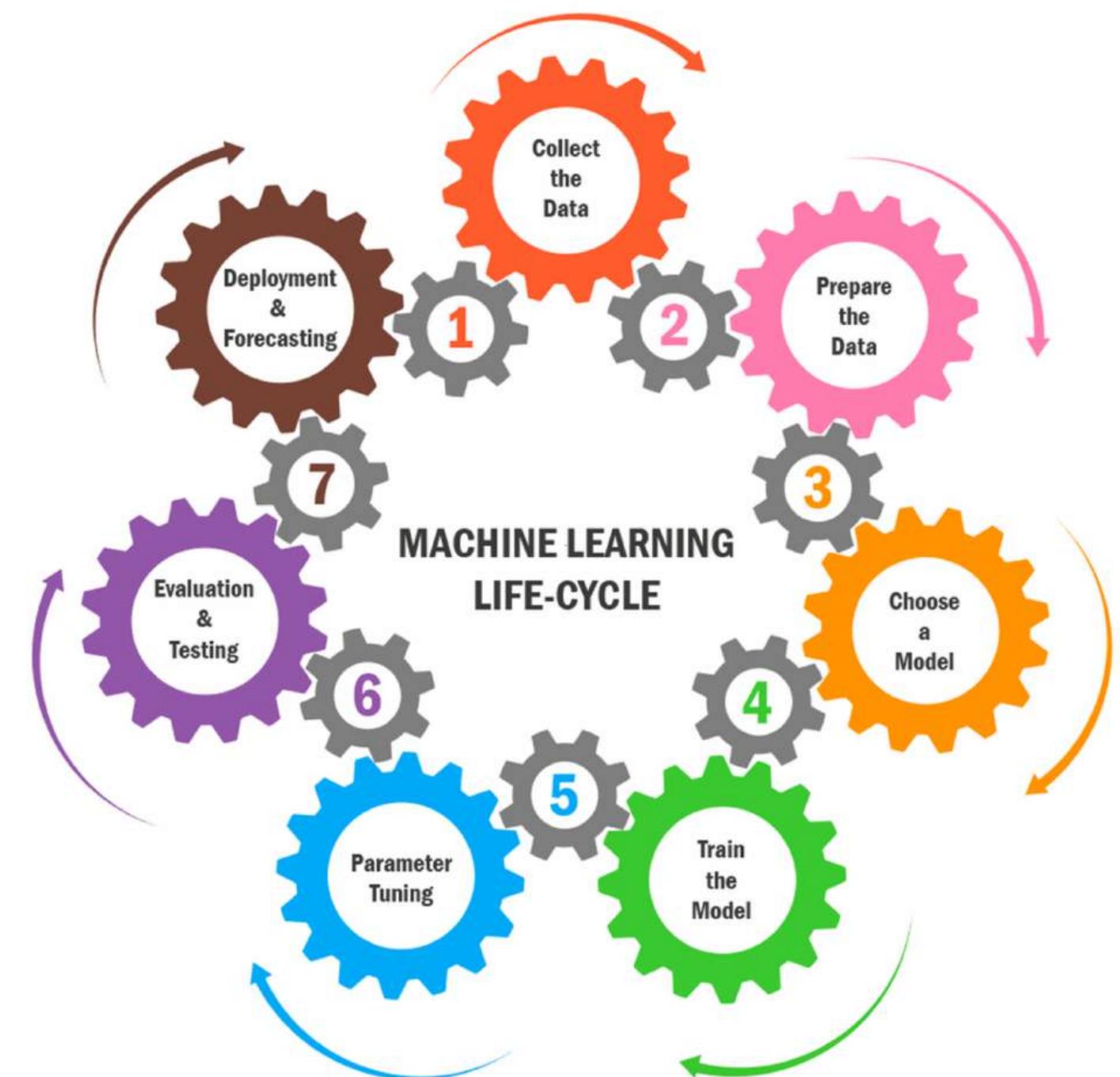
APLICACIONES DE MACHINE LEARNING

Las aplicaciones de machine learning mejoran con el uso y se vuelven más precisas a medida que tienen acceso a más datos

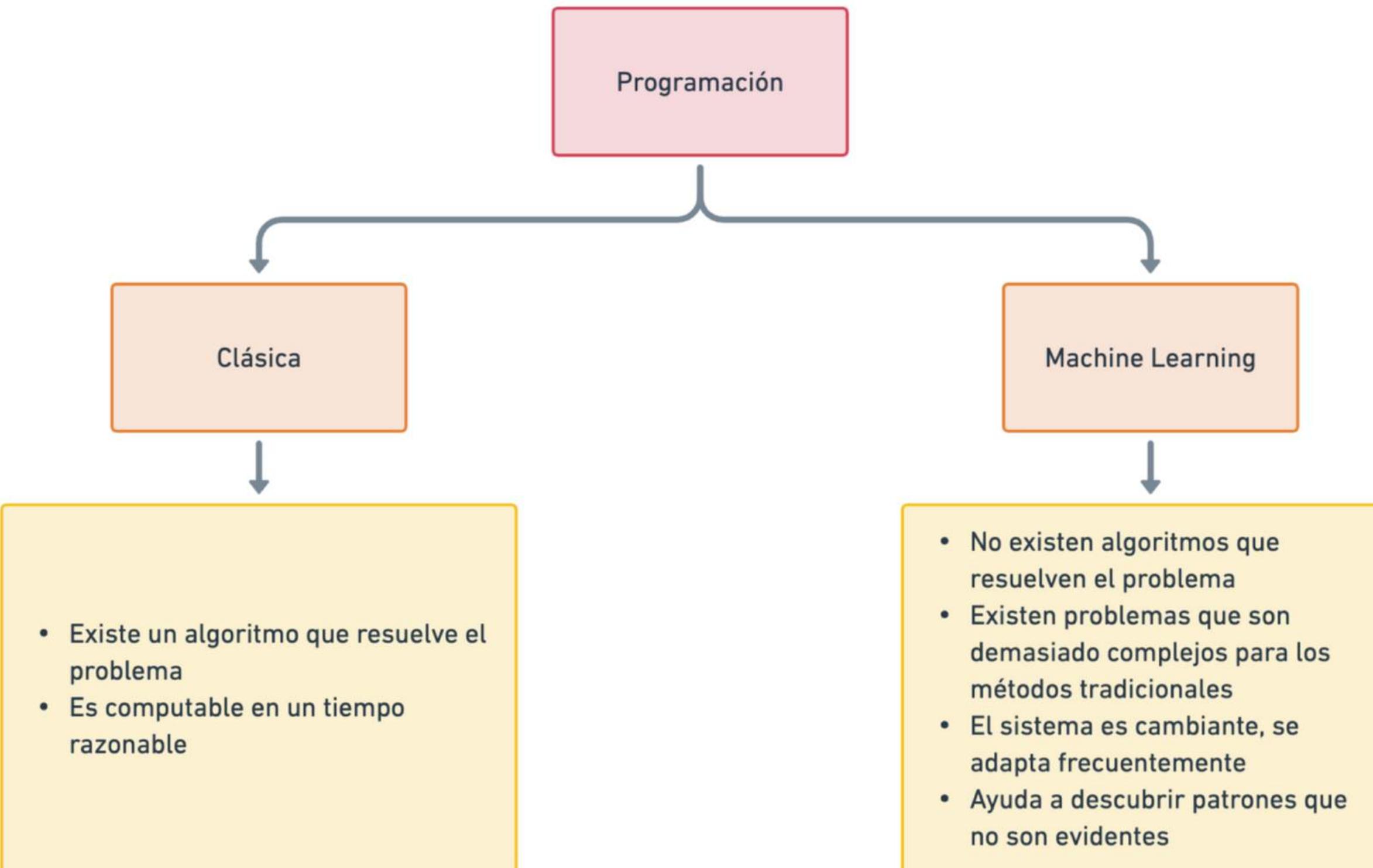
- Las aplicaciones de machine learning están en nuestras vida:
 - en nuestras casas
 - carritos de compra
 - medios de entretenimiento
 - cuidado de la salud
 - etc.



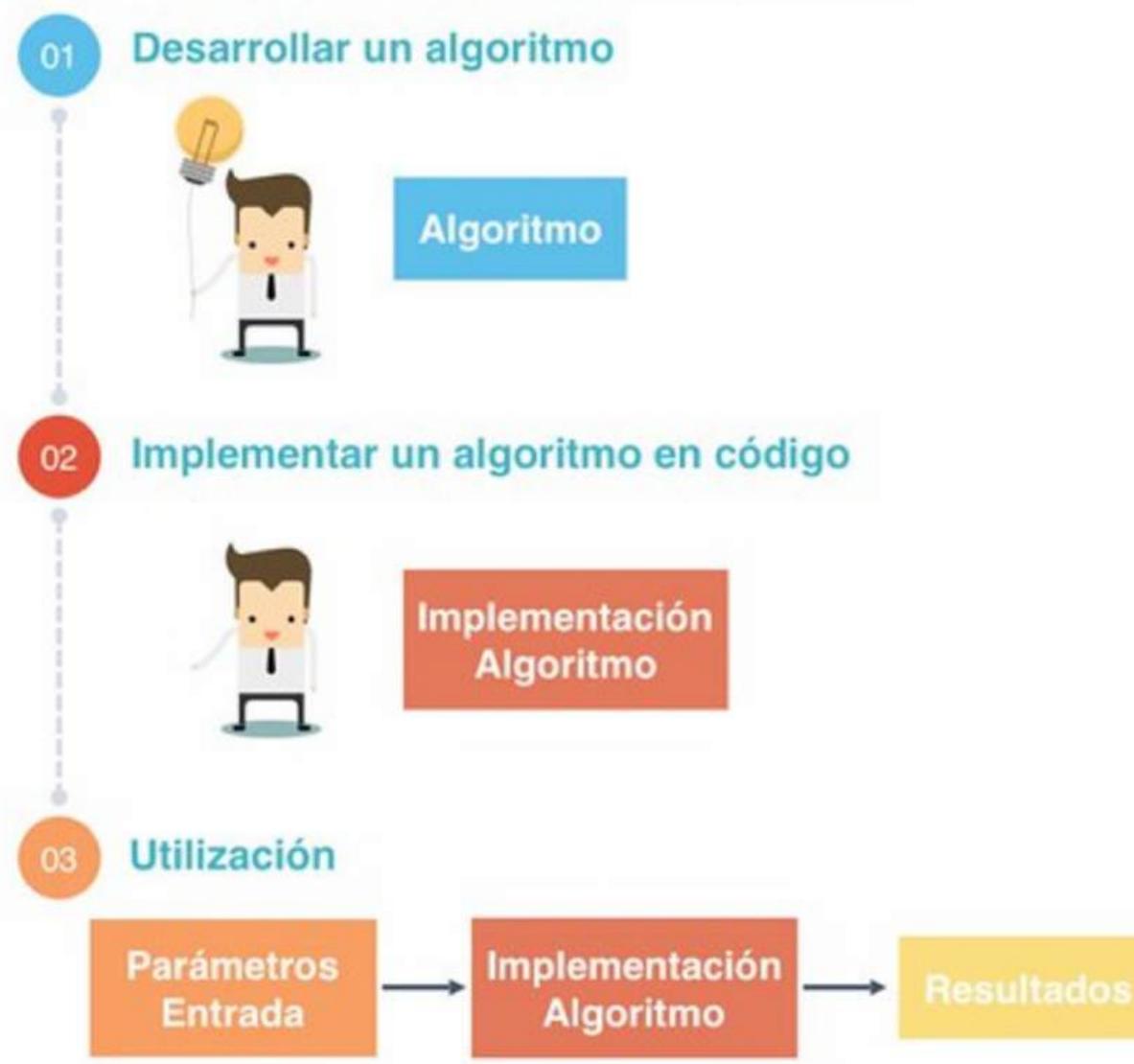
MACHINE LEARNING - CICLO DE VIDA



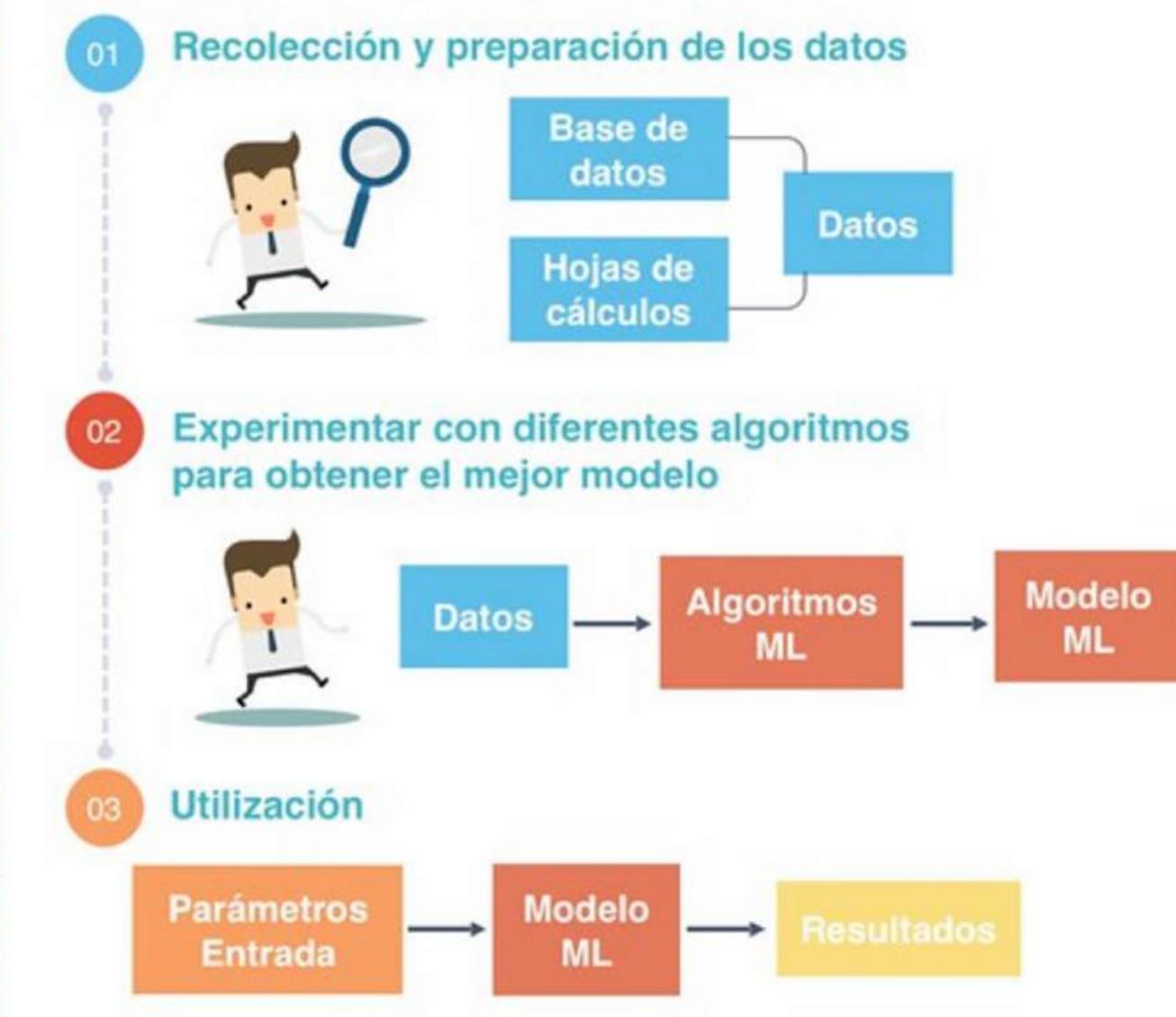
¿CUÁL ES LA DIFERENCIA ENTRE PROGRAMACIÓN CLÁSICA Y MACHINE LEARNING?



Programación Tradicional



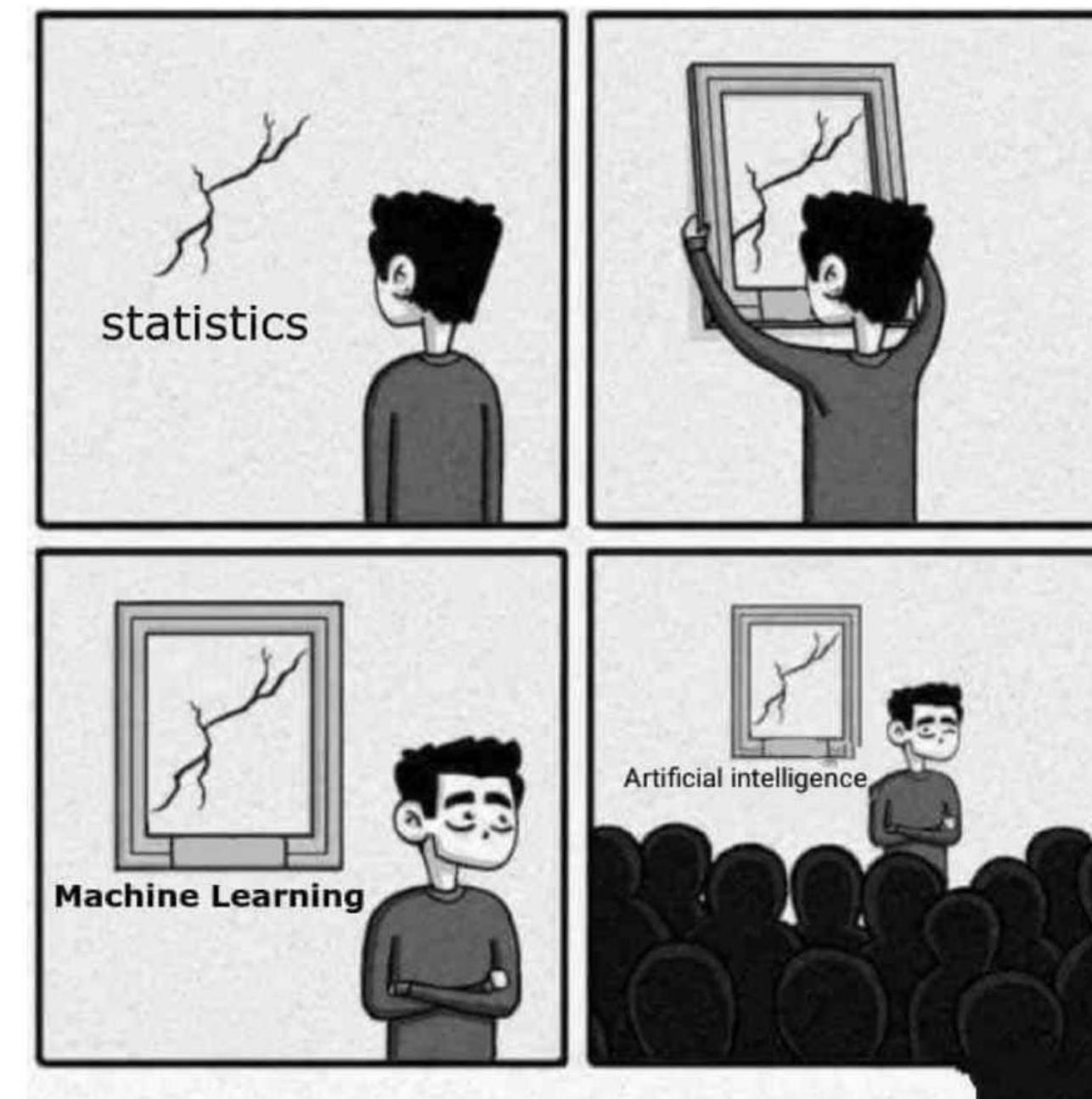
Machine Learning



En conclusión: La programación tradicional hay que formular y/o codificar manualmente las reglas, mientras que en Machine Learning los algoritmos formulaan automáticamente las reglas a partir de los datos, lo que es muy potente.

¿POR QUE USAR MACHINE LEARNING?

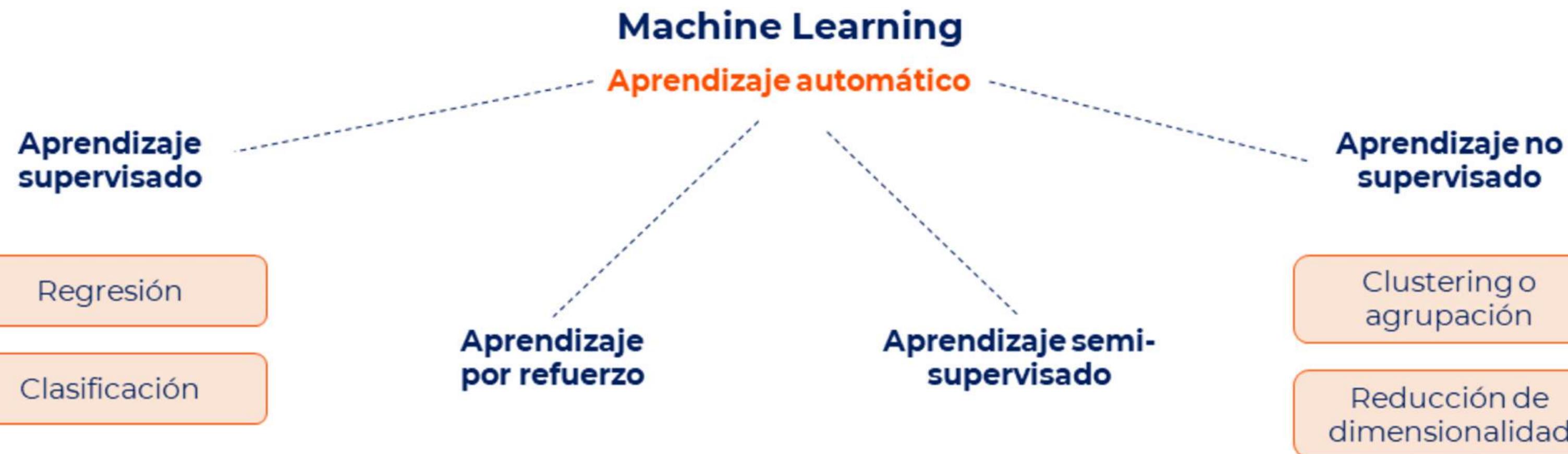
Machine Learning se utiliza en el caso de que la estrategia de programación tradicional se quede rezagada y no sea suficiente para implementar plenamente una determinada tarea





TIPOS DE APRENDIZAJE

TIPOS DE APRENDIZAJE



CLASIFICACIÓN DE LOS SISTEMAS DE MACHINE LEARNING

En función de la manera en que se entrenan:

- Aprendizaje Supervisado
- Aprendizaje No Supervisado
- Aprendizaje Semi Supervisado
- Aprendizaje Reforzado

En función de la manera en que aprenden con el tiempo:

- Aprendizaje online
- Aprendizaje batch

En función de la forma en la que realizan las predicciones:

- Aprendizaje basado en instancias
- Aprendizaje basado en modelos

APRENDIZAJE SUPERVISADO Y NO SUPERVISADO

Supervisado

Se cuenta con:

Un conjunto de datos X_m y de etiquetas Y_m
 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

Se obtiene:

Un modelo predictivo
Predecir un comportamiento o clase para
datos que no se han visto

$$Y = f(X)$$

Grupos de algoritmos:

- Clasificación
- Regresión

No supervisado

Se cuenta con:

Un conjunto de datos X_m
 $\{(x_1), (x_2), \dots, (x_m)\}$

Se obtiene:

Un modelo descriptivo
Obtener más información sobre los datos:

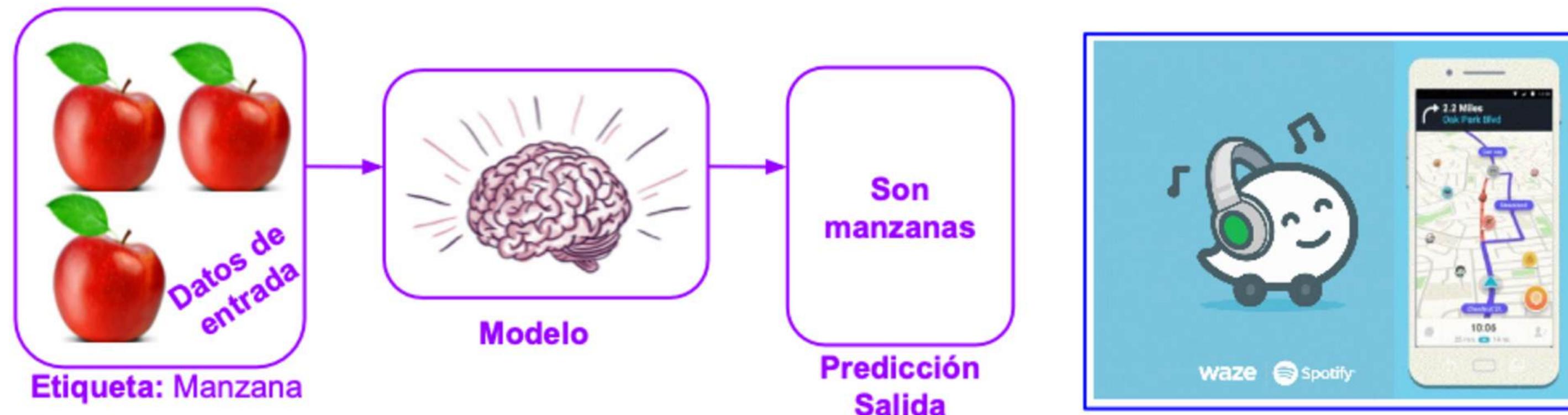
- Patrones
- Estructuras
- Distribuciones

Grupos de algoritmos:

- Agrupamiento
- Reglas de asociación

ALGORITMOS SUPERVISADOS

En los algoritmos de aprendizaje supervisado, **a la máquina se le enseña mediante ejemplos**. Consisten en pares de datos de "entrada" y "salida", donde la salida se etiqueta con el valor deseado



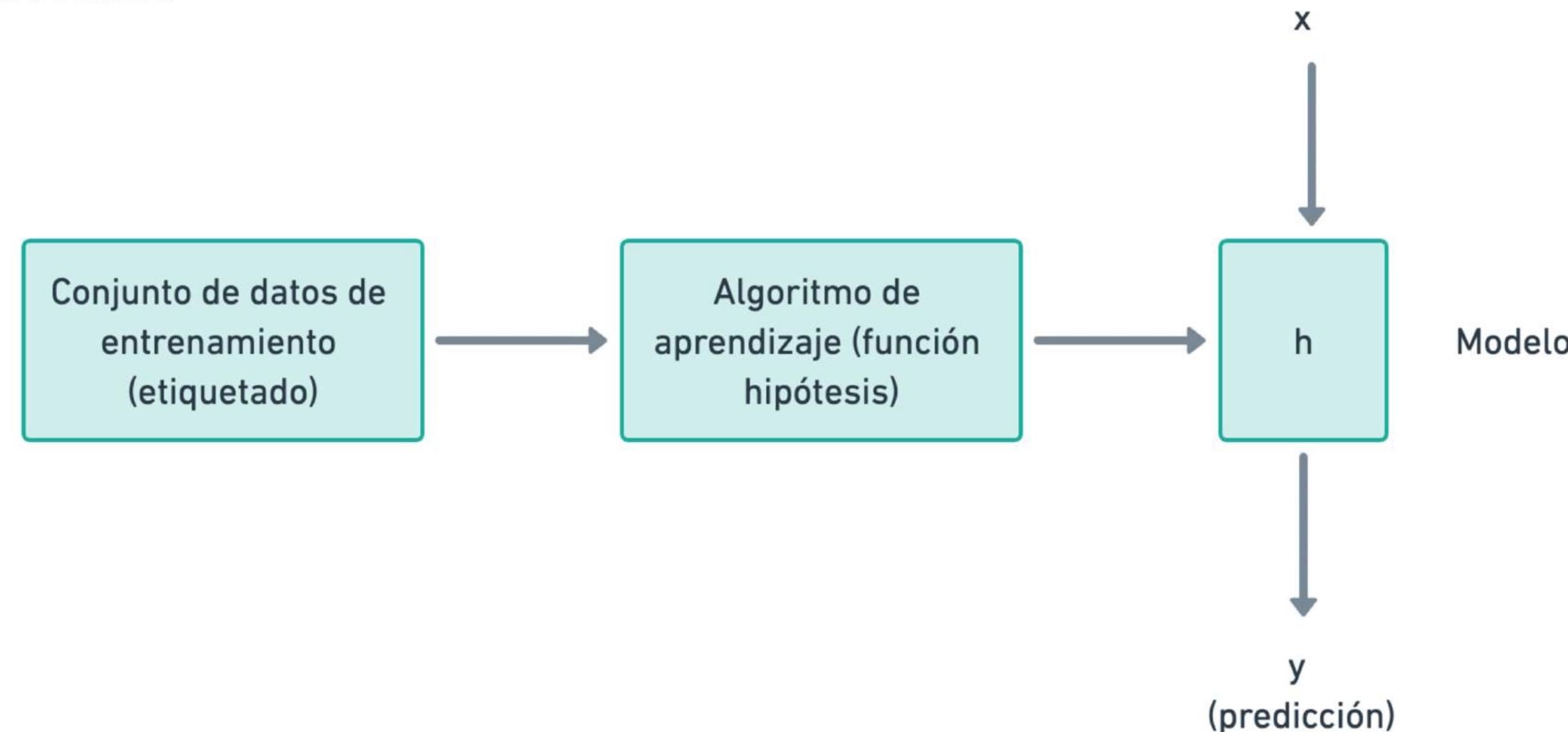
Ejemplo:

Los modelos de aprendizaje supervisado se utilizan en muchas de las aplicaciones con las que interactuamos todos los días, como motores de recomendación para productos y aplicaciones de análisis de tráfico como Waze, que prevén la ruta más rápida en diferentes horas del día

ALGORITMOS SUPERVISADOS

Es la tarea de aprendizaje automático que consiste en aprender una función que mapea una entrada a una salida basada en pares de entrada-salida de ejemplo.

La función resultante es utilizada posteriormente para predecir valores a partir de ejemplos de datos no etiquetados

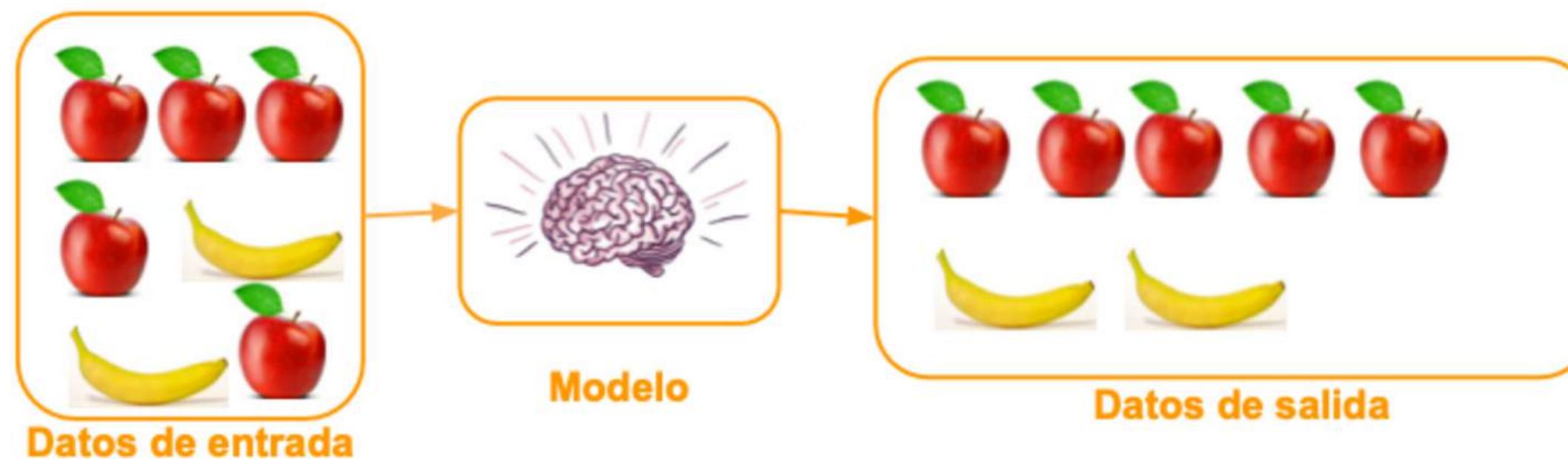


ALGORITMOS NO SUPERVISADOS

En los modelos de aprendizaje no supervisado, **no existe una clave de respuesta**.

La máquina estudia los datos de entrada –muchos de los cuales no están etiquetados ni estructurados– y comienza a identificar patrones y correlaciones, utilizando todos los datos relevantes y accesibles.

El aprendizaje no supervisado sigue el modelo de cómo los humanos observan el mundo. Utilizamos la intuición y la experiencia para agrupar cosas. A medida que experimentamos cada vez más ejemplos de algo, nuestra capacidad de categorizar e identificar se vuelve cada vez más precisa.

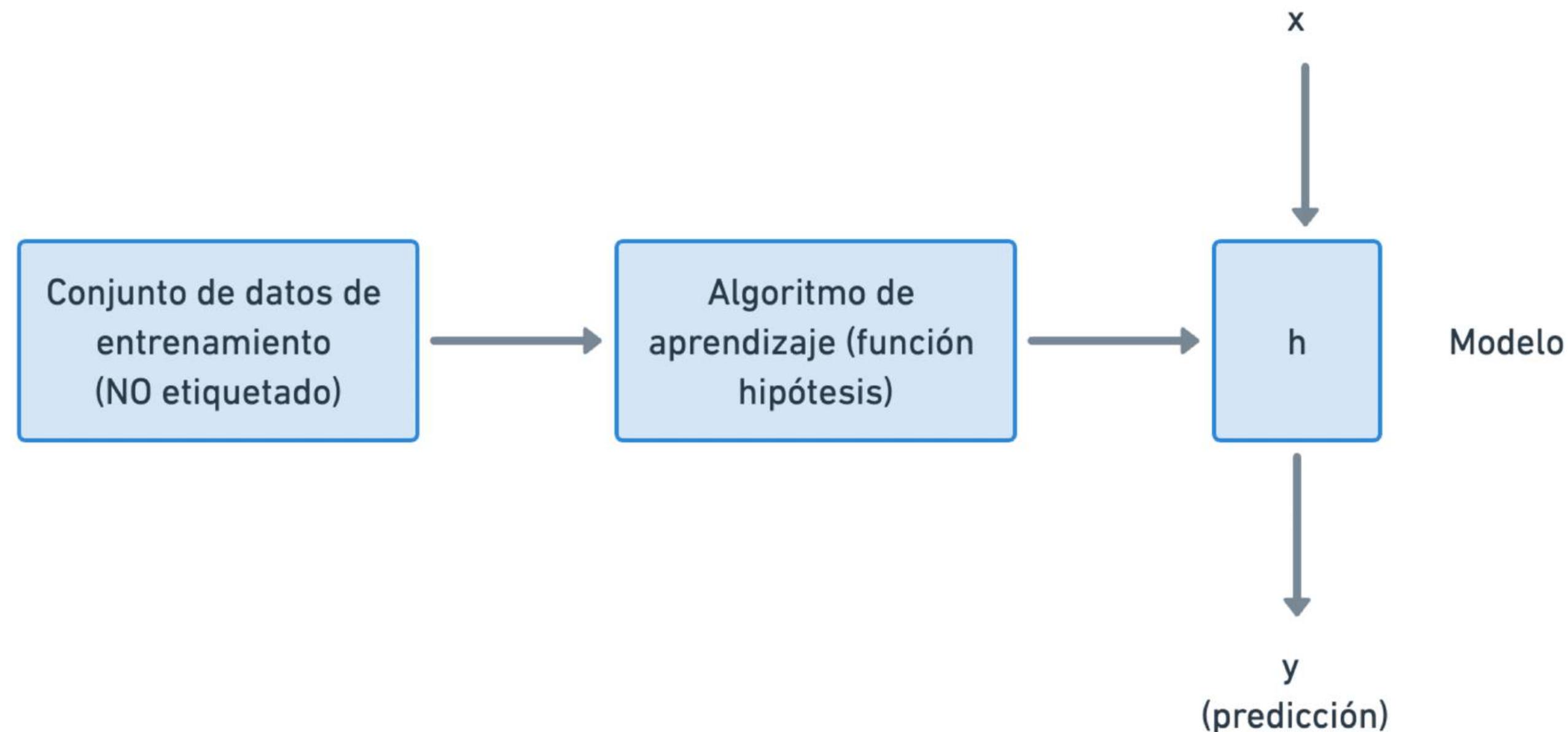


Ejemplo:

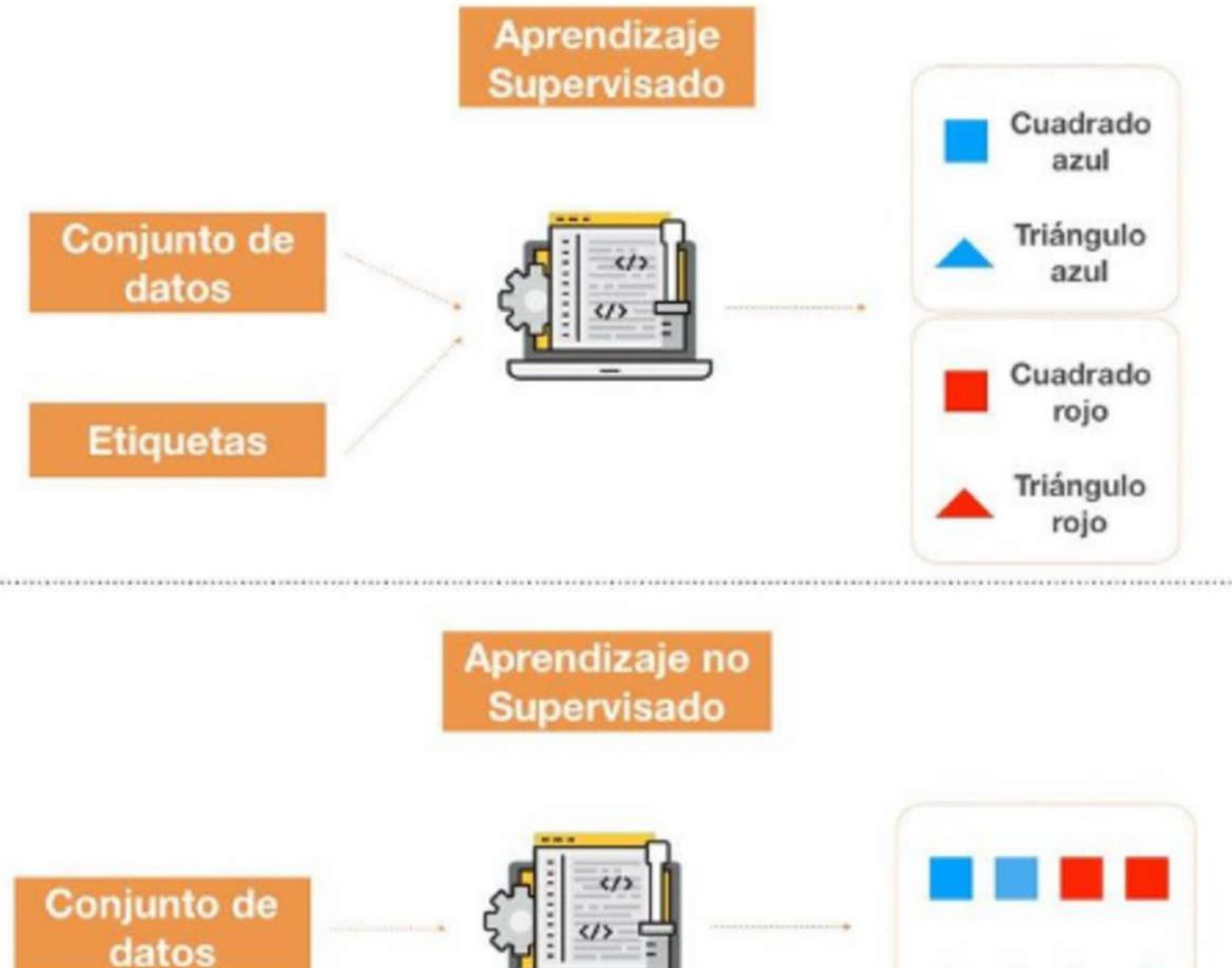
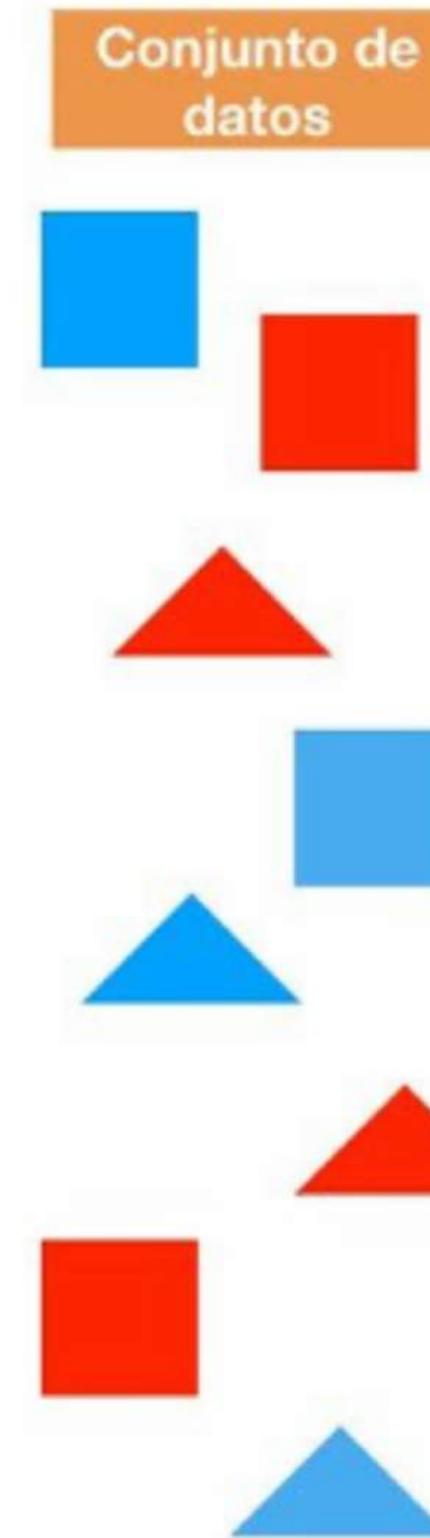
El reconocimiento facial, el análisis de secuencias genéticas, la investigación de mercado y la ciberseguridad

ALGORITMOS NO SUPERVISADOS

Es la tarea de aprendizaje automático que consiste en inferir una función que describe la estructura de un conjunto de datos sin etiquetar (es decir, datos que no se han clasificado ni categorizado)



COMPARANDO LOS TIPOS DE APRENDIZAJES



ALGORITMOS SEMI SUPERVISADOS

El aprendizaje semi supervisado se convierte en una solución viable cuando hay grandes cantidades de datos crudos y no estructurados

Este modelo consiste en

- Introducir pequeñas cantidades de datos etiquetados para aumentar los datasets sin etiquetar.
- Instruye a la máquina para que analice los datos etiquetados según propiedades correlativas que podrían aplicarse a los datos no etiquetados.



Ejemplo:

El aprendizaje semi supervisado se utiliza en el análisis del habla y del lenguaje, las investigaciones médicas complejas como la categorización de proteínas, y la detección de fraude de alto nivel

ALGORITMOS POR REFORZAMIENTO O REFUERZO

- En el modelo de aprendizaje por refuerzo no incluye una respuesta de referencia, sino que más bien introduce un conjunto de acciones permitidas, reglas y estados finales potenciales.
- Cuando el objetivo deseado del algoritmo es fijo o binario, las máquinas pueden aprender mediante el ejemplo.
 - Pero en los casos en los que el resultado deseado es mutable, el sistema debe aprender por experiencia y recompensa. En los modelos de aprendizaje por refuerzo, la "recompensa" es numérica y se programa dentro del algoritmo como algo que el sistema busca recopilar.
- Este modelo es análogo a enseñarle a alguien a jugar ajedrez. Sin duda, sería imposible intentar mostrarle cada movimiento potencial. En cambio, se le explican las reglas, y la persona aumenta su habilidad a través de la práctica. Las recompensas provienen no solo de ganar el juego, sino también de adquirir las piezas del oponente.

Ejemplo:

Las aplicaciones de aprendizaje por refuerzo incluyen las gráficas de precios automatizados para los compradores de publicidad on-line, el desarrollo de juegos de computadora, y la negociación bursátil de alto riesgo.

APRENDIZAJE POR REFUERZO

Se cuenta con:

- Agentes, ambientes, estados, acciones y recompensas

Se obtiene:

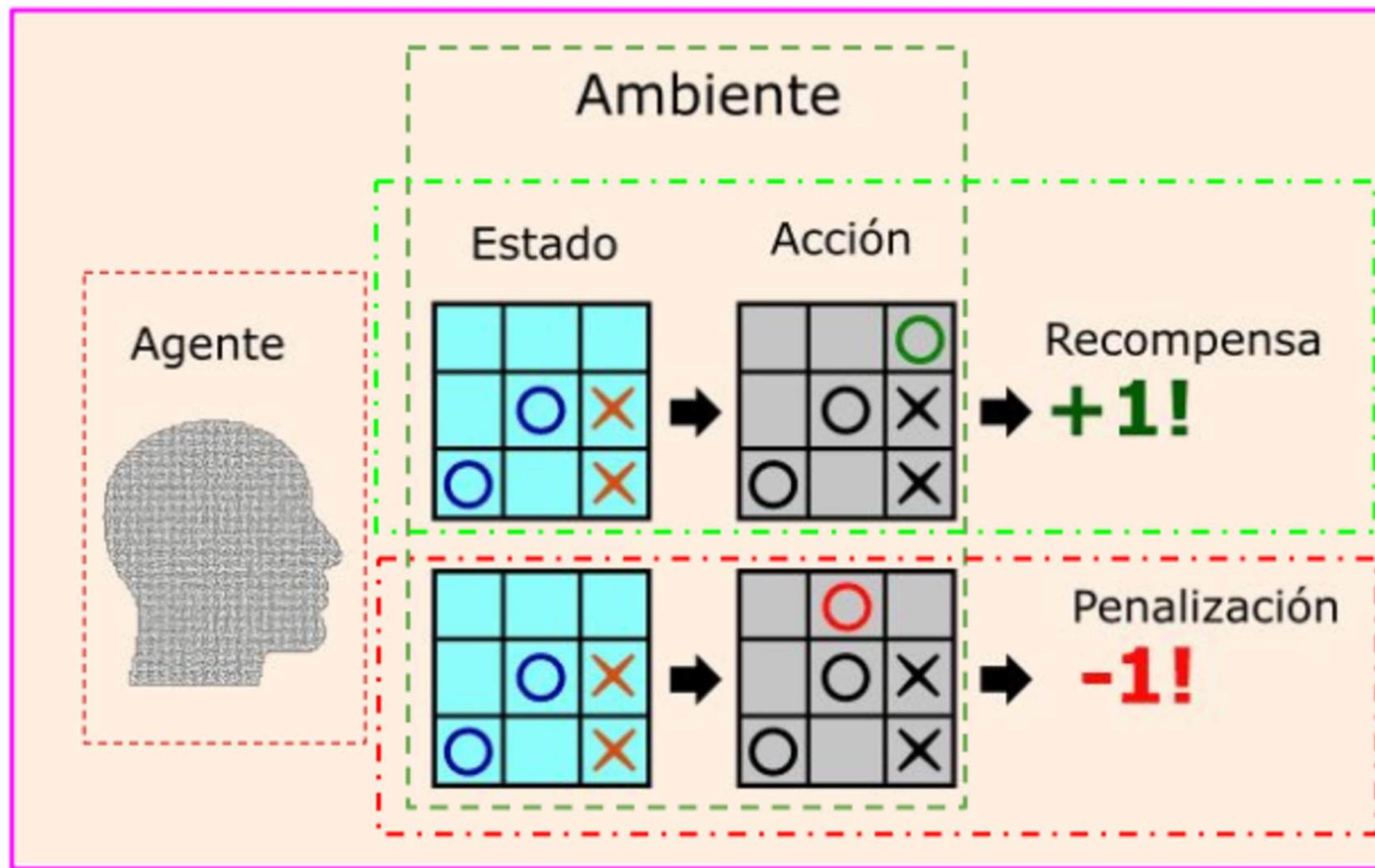
- Un comportamiento del agente ante su ambiente, controlado por políticas auto aprendidas que le dicen qué acciones tomar para maximizar recompensas (positivas) y/o minimizar riesgos (negativas)

Aplicaciones:

- Juegos (de mesa, videojuegos)
- Brazos robóticos
- Vehículos autónomos

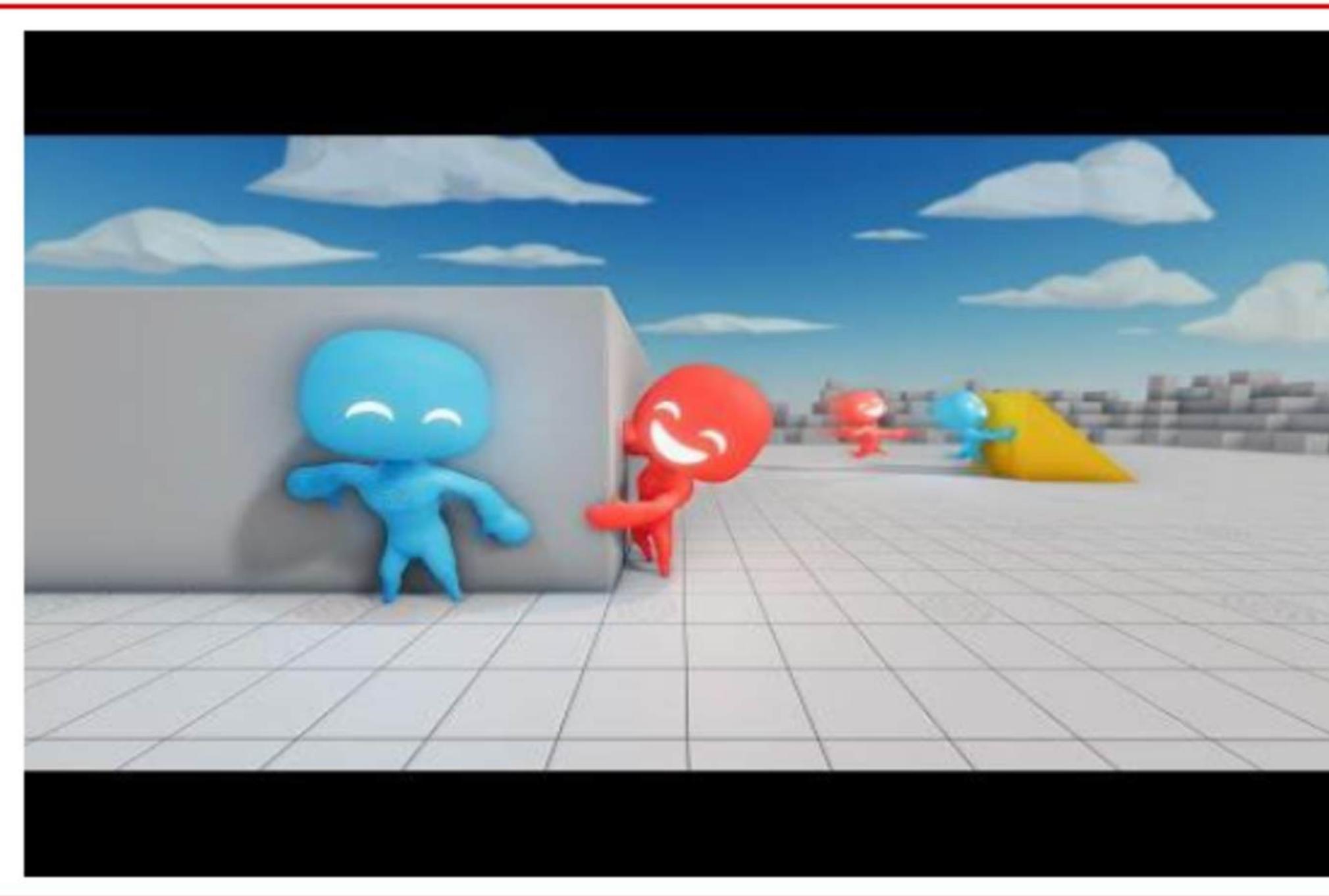


APRENDIZAJE POR REFUERZO



APRENDIZAJE POR REFUERZO

Aprender a esconderse y a buscar



Emergent Tool Use

APLICACIONES DE LAS EMPRESAS BASADAS EN ML

El campo de aplicación práctica depende de la imaginación y de los datos que estén disponibles en la empresa. Estos son algunos ejemplos:

- Detectar fraude en transacciones
- Predecir fallos en equipos tecnológicos
- Prever qué empleados serán más rentables el año que viene (el sector de los Recursos Humanos está apostando seriamente por el Machine Learning)
- Seleccionar clientes potenciales basándose en comportamientos en las redes sociales, interacciones en la web...
- Predecir el tráfico urbano
- Saber cuál es el mejor momento para publicar tuits, actualizaciones de Facebook o enviar las newsletter
- Hacer pre diagnósticos médicos basados en síntomas del paciente
- Cambiar el comportamiento de una app móvil para adaptarse a las costumbres y necesidades de cada usuario
- Detectar intrusiones en una red de comunicaciones de datos
- Decidir cuál es la mejor hora para llamar a un cliente

PLAYGROUND

Machine Learning Playground

The screenshot shows the Machine Learning Playground interface. At the top right, there are buttons for "Upload Data" (orange), "Save Data" (purple), "Clear all" (red), and a red "X". Below these are four model cards: "K Nearest Neighbors" (selected, orange border), "Perceptron" (grey), "Support Vector Machine" (grey), and "Artificial Neural Network" (grey). A fifth card, "Decision Tree", is partially visible below them. On the left, there is a large white area for visualizations. At the bottom right, there is a "Parameters:" section with a "K:" input field containing "3" and a "Train" button.

Upload Data Save Data

Clear all

K Nearest Neighbors

Perceptron

Support Vector Machine

Artificial Neural Network

Decision Tree

Parameters:

K:

Train

<https://ml-playground.com/>



**¿CÓMO LLEGAMOS
HASTA ACA?**

Hojas de Cálculo



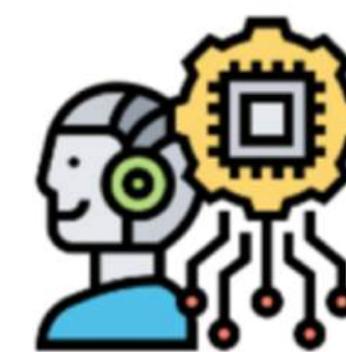
BD Relacionales



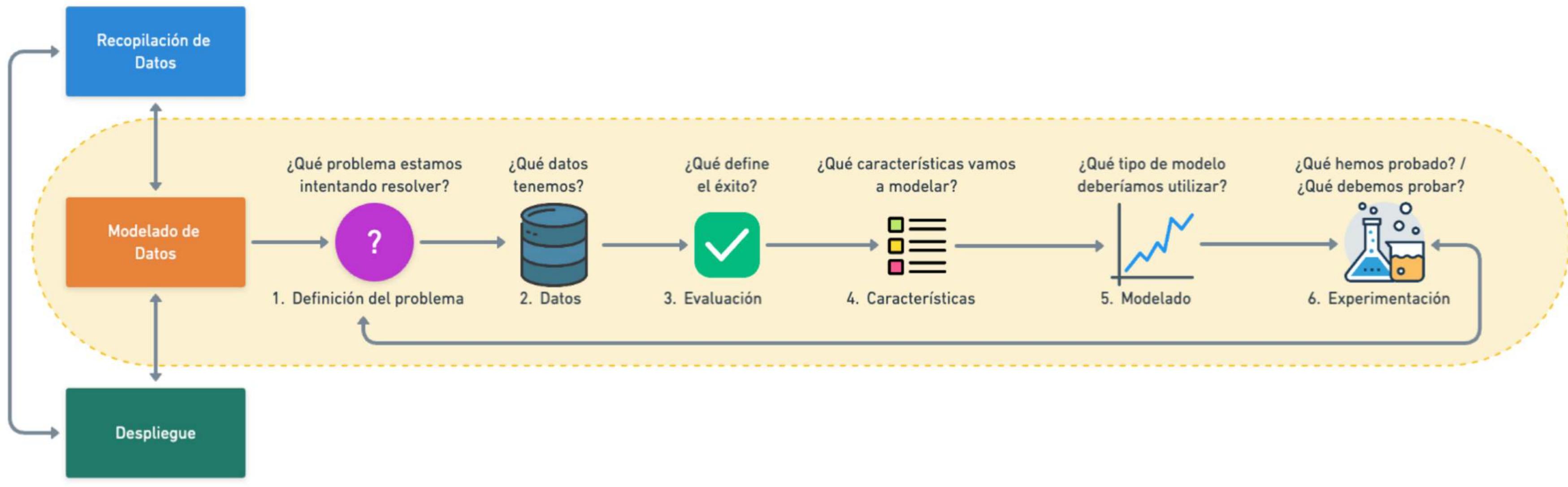
BD No Relacionales



Machine Learning



PASOS DE UN PROYECTO



PASOS DE UN PROYECTO

Una guía de campo de 6 pasos para crear proyectos de aprendizaje automático

Definición del problema: ¿Qué problema empresarial estamos intentando resolver? ¿Cómo se puede expresar como un problema de aprendizaje automático?

Datos: Si el aprendizaje automático consiste en obtener información a partir de los datos, ¿qué datos tenemos? ¿Cómo coincide con la definición del problema? ¿Nuestros datos están estructurados o no estructurados? ¿Estático o en streaming?

Evaluación: ¿Qué define el éxito? ¿Es suficiente un modelo de aprendizaje automático con una precisión del 95%?

Características: ¿Qué partes de nuestros datos vamos a utilizar para nuestro modelo? ¿Cómo puede influir en esto lo que ya sabemos?

Modelado: ¿Qué modelo debería elegir? ¿Cómo puedes mejorarlo? ¿Cómo lo comparas con otros modelos?

Experimentación: ¿Qué más podríamos probar? ¿Nuestro modelo implementado funciona como esperábamos? ¿Cómo cambian los otros pasos según lo que hemos encontrado?

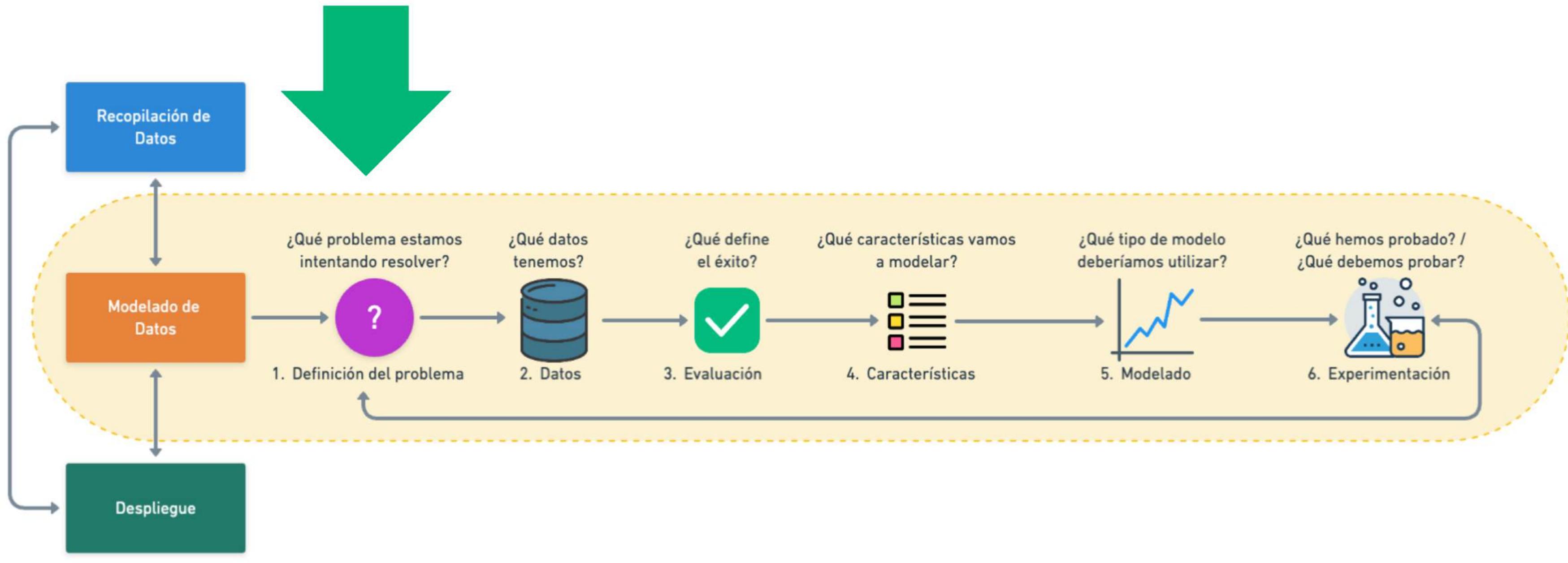


FRAMEWORK MACHINE LEARNING

DEFINICIÓN DEL PROBLEMA



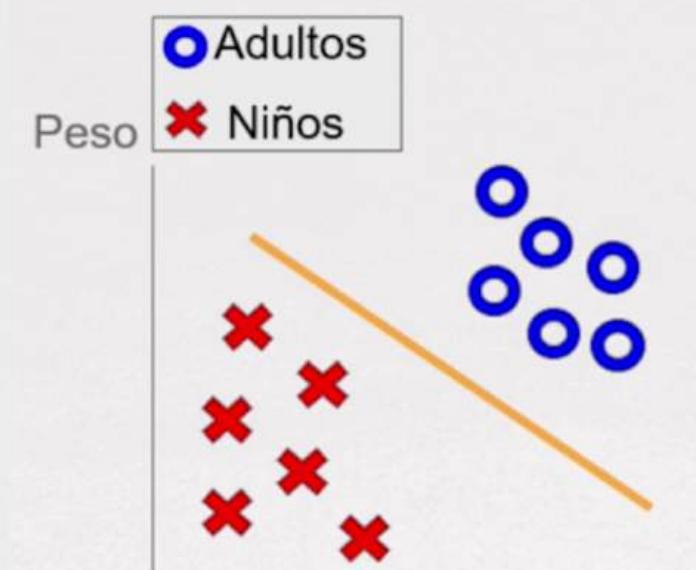
FRAMEWORK



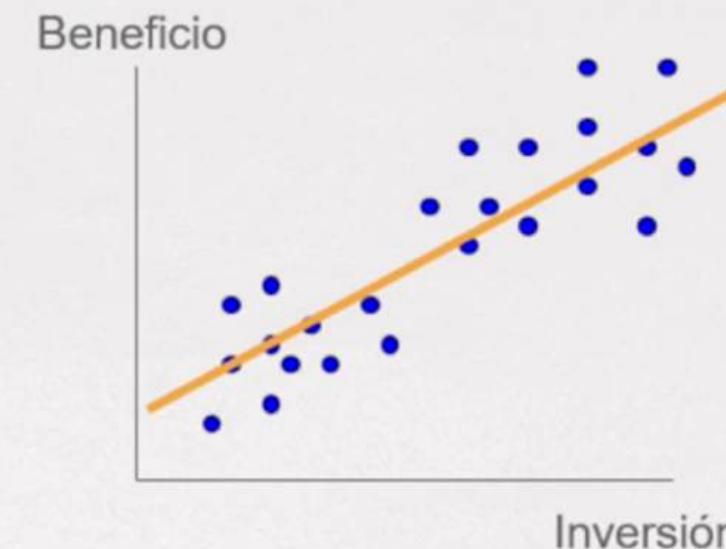
¿CUÁLES SON LOS TIPOS DE PROBLEMAS?

Técnicas de Machine Learning

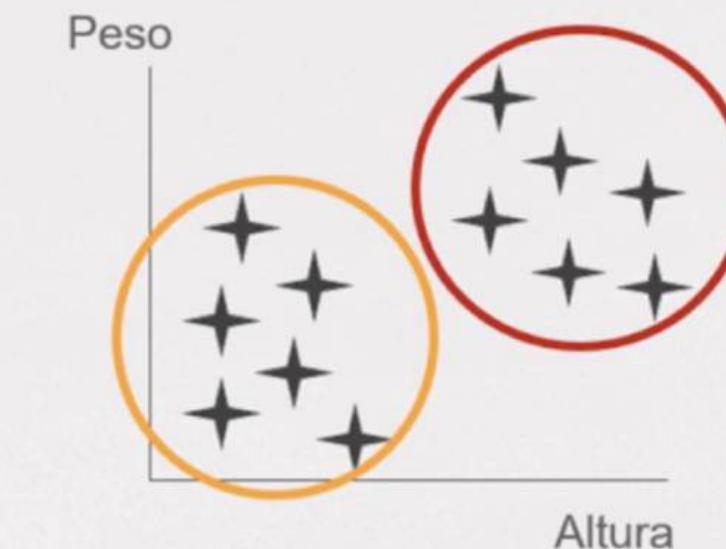
Clasificación



Regresión



Agrupación (*clustering*)

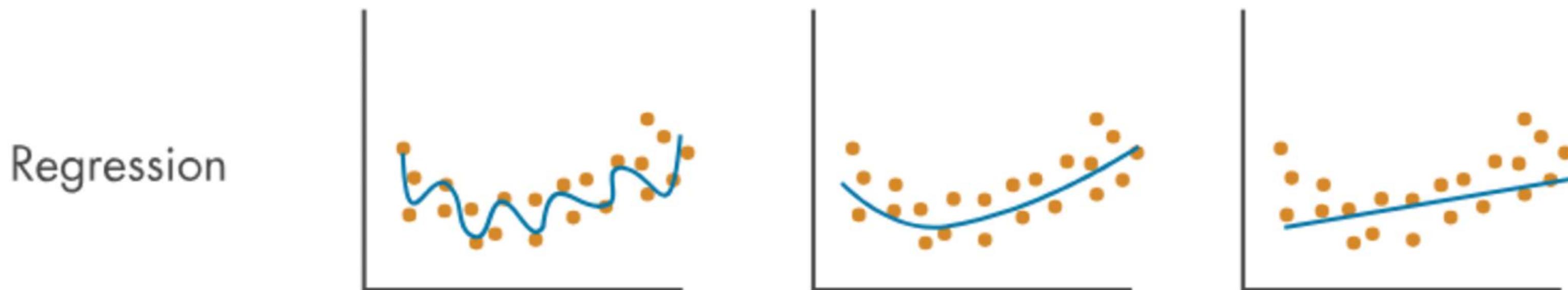


APRENDIZAJE SUPERVISADO

APRENDIZAJE NO SUPERVISADO

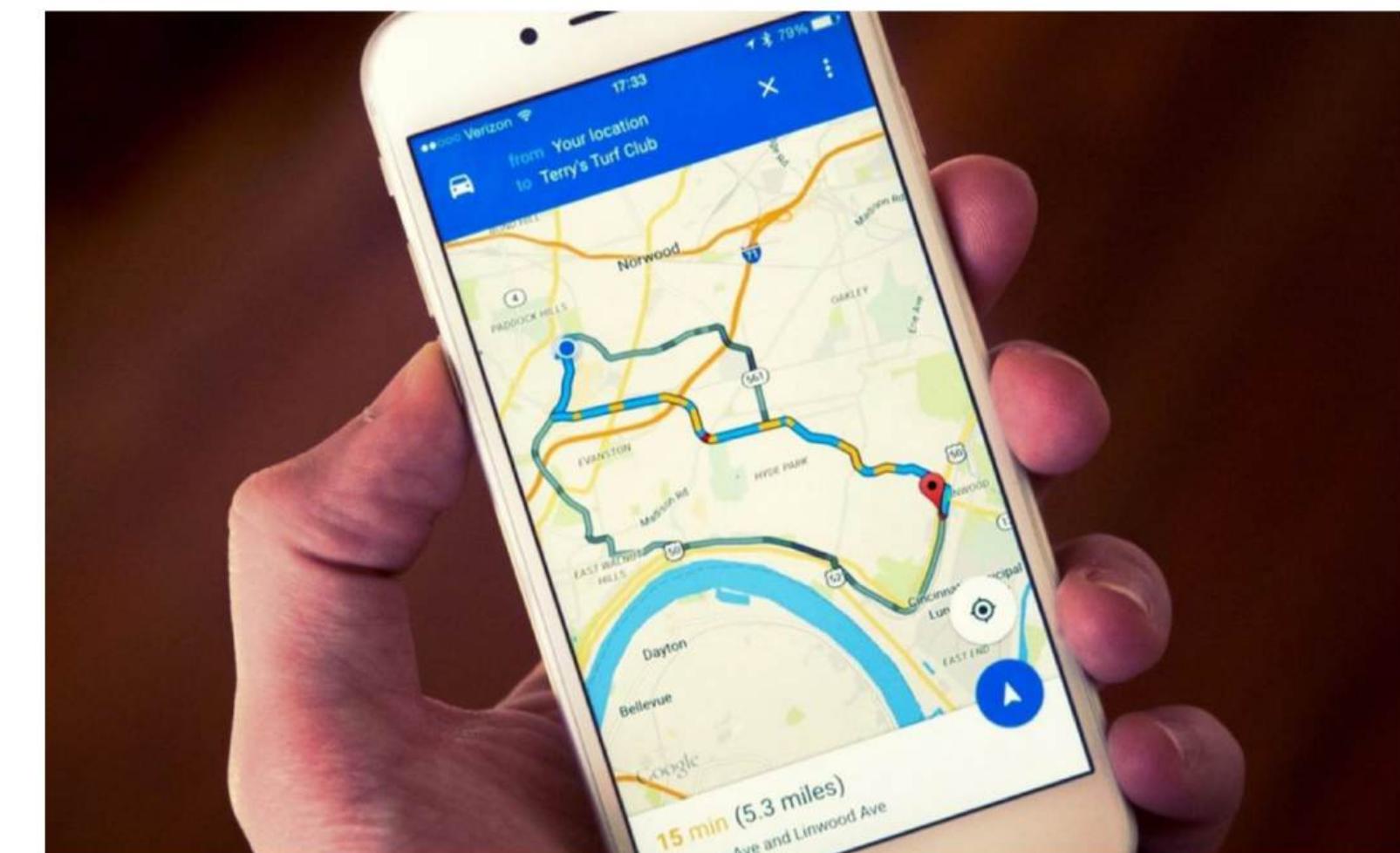
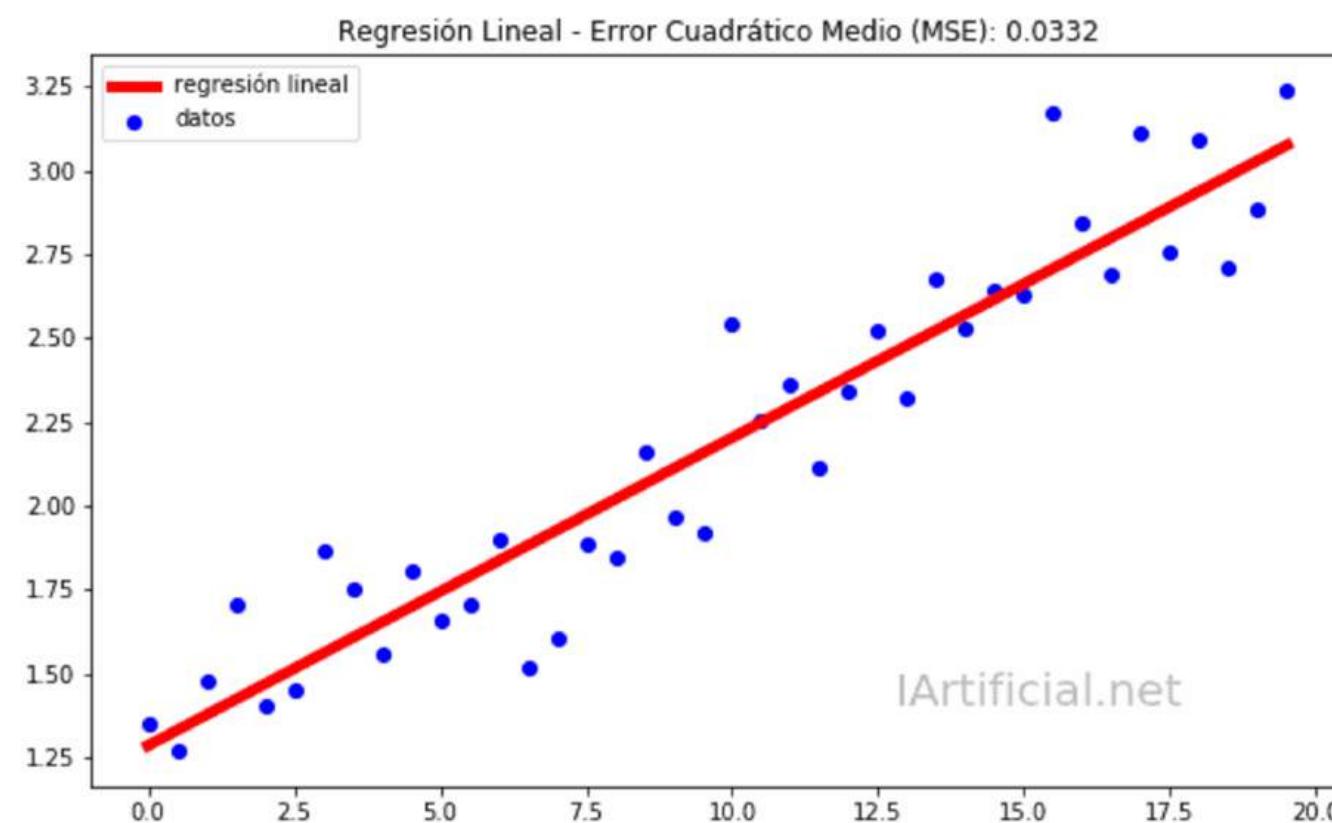
¿QUÉ ES LA REGRESIÓN?

Es cuando usamos regresión lineal, el resultado es un número. Es decir, el resultado de la técnica de machine learning que estemos usando será un valor numérico, dentro de un conjunto infinito de posibles resultados



REGRESIÓN

En el caso de los algoritmos de regresión, podemos decir que se trata de un subcampo del aprendizaje automático supervisado que tiene el fin de crear una metodología para relacionar un cierto número de características y una variable objetivo-continua



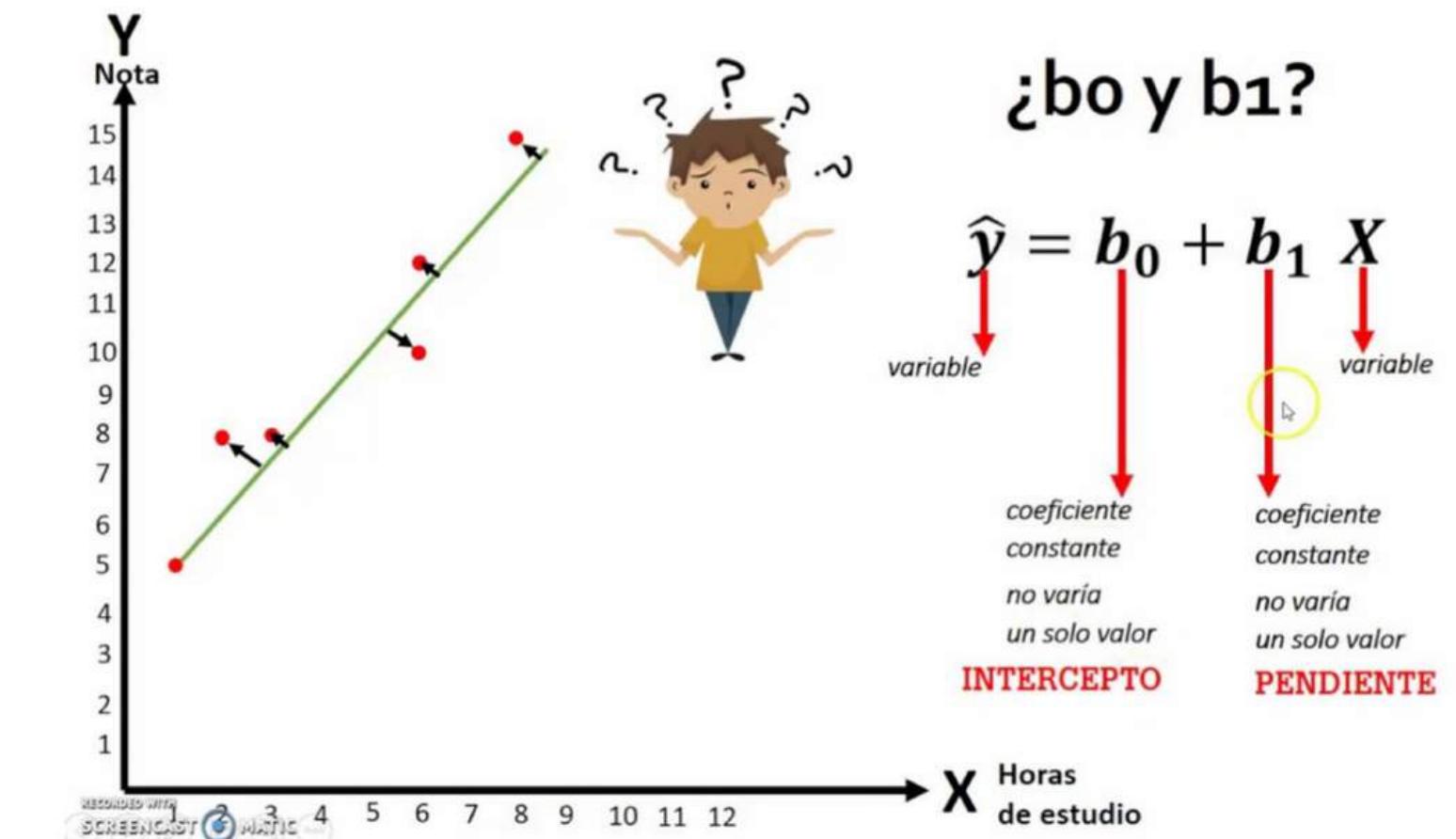
Ejemplo:

Es la estimación de cuánto tardará una persona en llegar a un destino (tiempo de trayecto)

TÉCNICAS DE REGRESIÓN

Existen distintas técnicas usadas en los algoritmos de regresión. Entre ellas, podemos encontrar:

- Regresión lineal (utiliza datos continuos)
- Regresión no lineal
- Regresión logística (utiliza datos discretos)
- Árboles de decisión
- Bosques aleatorios
- Máquinas de vectores de soporte
- Redes neuronales y aprendizaje profundo



EJEMPLO DE REGRESIÓN

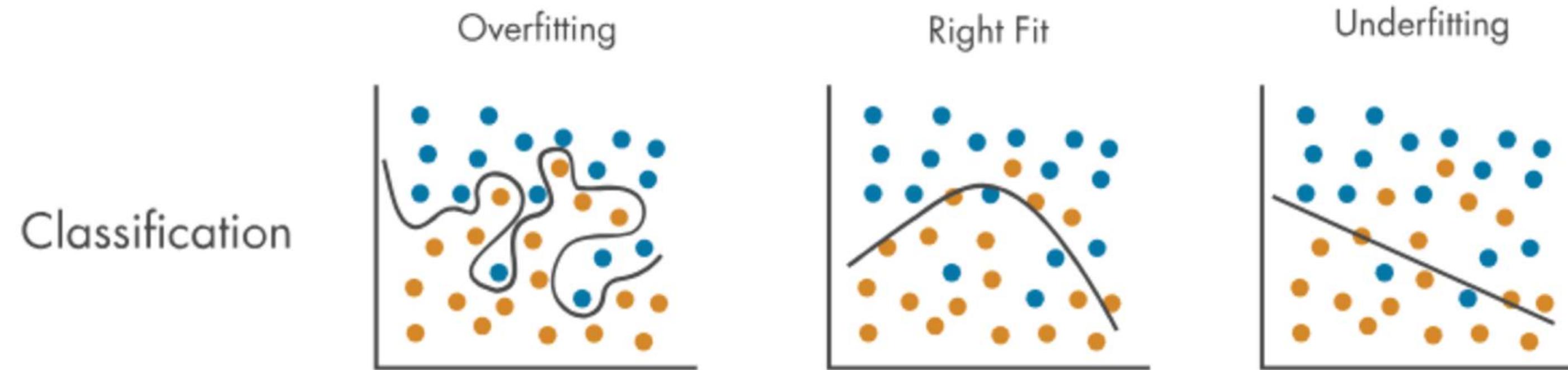
- Predecir por cuánto se va a vender una propiedad inmobiliaria
- Predecir cuánto tiempo va a permanecer un empleado en una empresa o compañía
- Estimar cuánto tiempo va a tardar un vehículo en llegar a su destino
- Estimar cuántos productos se van a vender
- Etc.



¿QUÉ ES LA CLASIFICACIÓN?

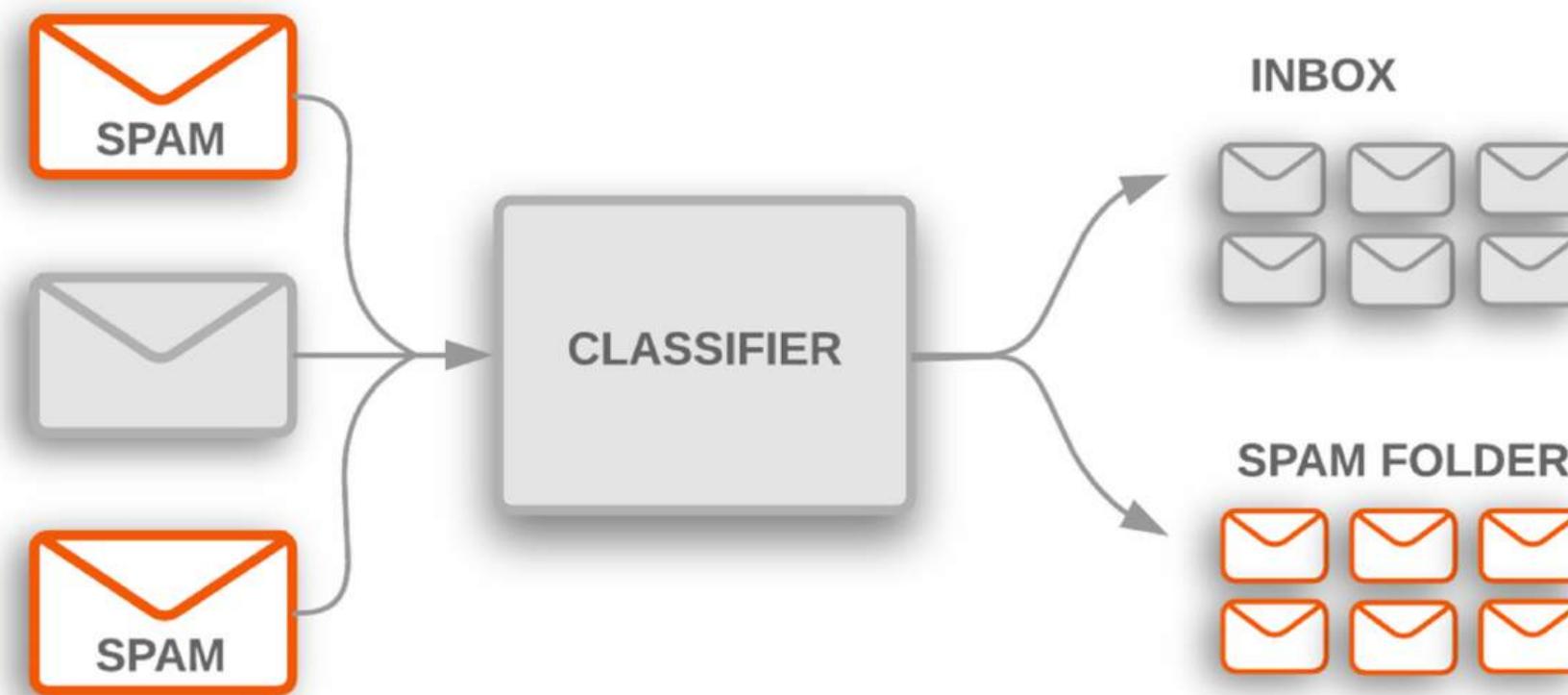
En la clasificación se tiene la tarea de asignar una clase, es decir predecir a qué clase pertenece un conjunto de datos, aquí es muy importante entender que en los problemas de clasificación los valores son discretos

Con clases nos referimos a categorías arbitrarias según el tipo de problema



CLASIFICACIÓN

A esta metodología también se la conoce por **clasificación binaria** o **regresión logística**. Los algoritmos de clasificación se utilizan en casos en los que el resultado es un conjunto infinito de resultados.



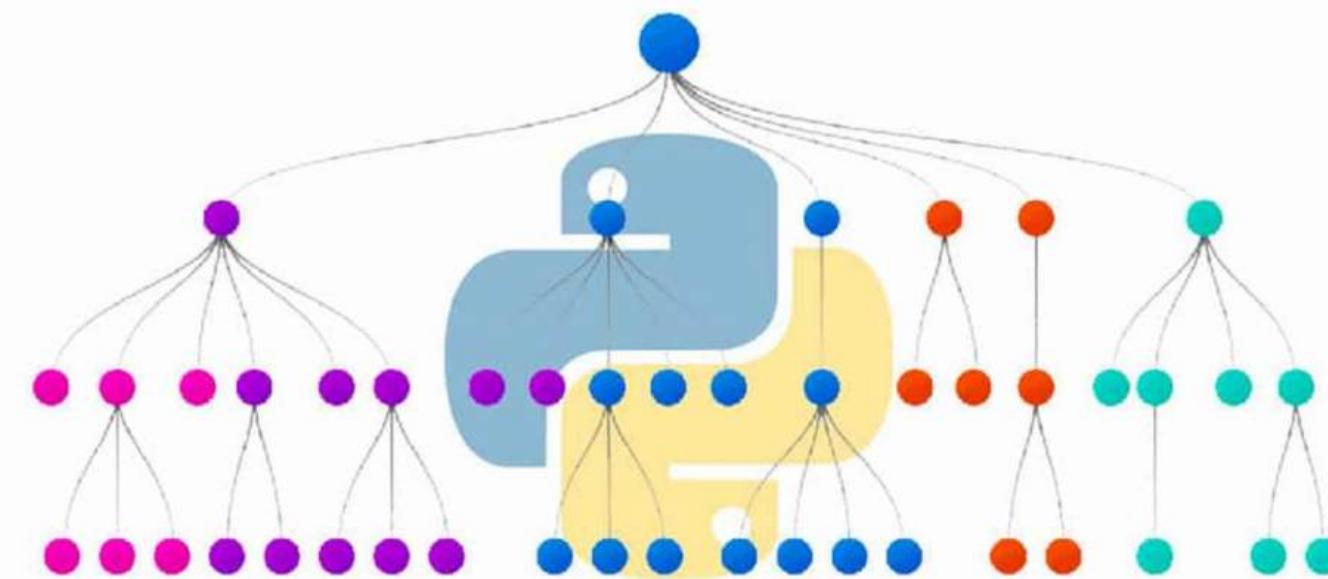
Ejemplo:

Un detector de spam o correo no deseado en el mail. Si buscamos saber si un correo es o no es spam, el algoritmo de clasificación decide a qué tipo pertenece.

TÉCNICAS USANDO CLASIFICACIÓN

En la actualidad se usan distintas técnicas de aprendizaje automático para problemas de clasificación. Entre ellos encontramos:

- La clasificación de Naïve Bayes
- Los bosques aleatorios
- La regresión logística
- Los árboles de decisión
- Máquinas de vectores de soporte
- Redes neuronales y aprendizaje profundo



EJEMPLOS USANDO CLASIFICACIÓN

- ¿Comprará el cliente este producto? [sí, no]
- ¿Qué tipo de tumor es? [maligno, benigno]
- ¿Subirá el dólar mañana? [sí, no]
- ¿Es este comportamiento una anomalía? [sí, no]
- ¿Nos devolverán el crédito? [sí, no]
- ¿Obtendrá una historia un número alto de visitas en cierto sitio Web? [sí, no]
- ¿Qué deporte estás haciendo? tal y como lo detectan los relojes inteligentes [caminar, correr, bicicleta, nadar]
- Concurso de clasificación de imágenes ImageNet
- Etc



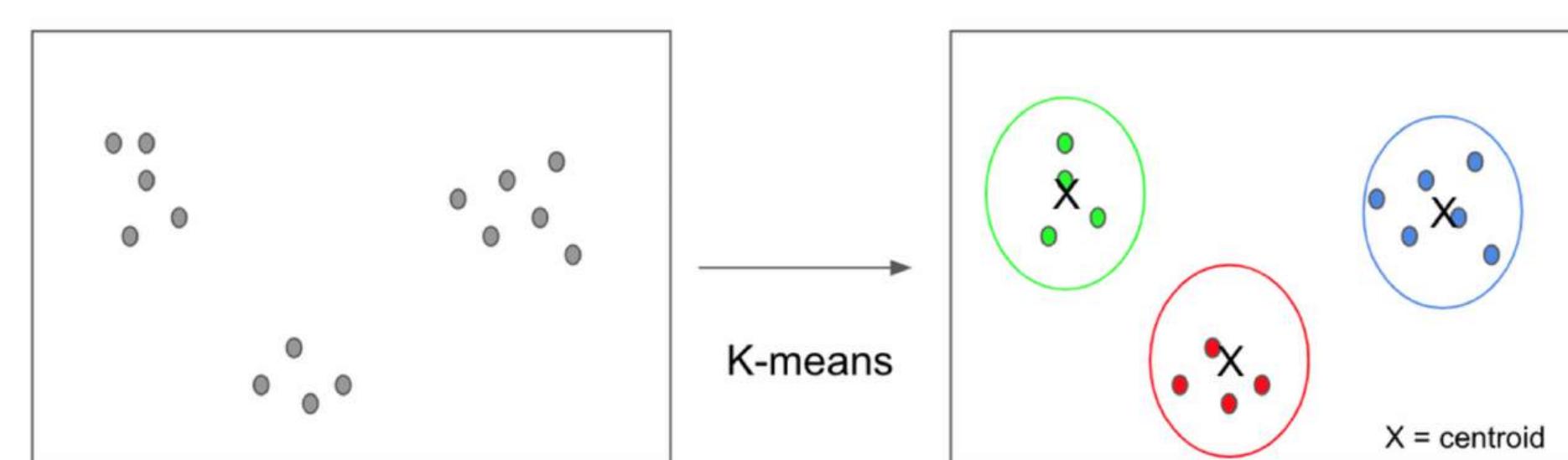
¿CUÁNDO USAR EL ALGORITMO DE AGRUPACIÓN?

Cuando tienes un conjunto de datos sin etiquetar, es muy probable que utilices algún tipo de algoritmo de aprendizaje sin supervisión

Puedes utilizar agrupamiento cuando intentas detectar anomalías para encontrar valores atípicos en tus datos. El agrupamiento ayuda a encontrar esos grupos y muestra los límites que determinarían si un punto de datos es un valor atípico o no

Si no estás seguro de qué características usar para tu modelo de aprendizaje automático, el agrupamiento descubre patrones que puedes usar para descubrir qué se destaca en los datos

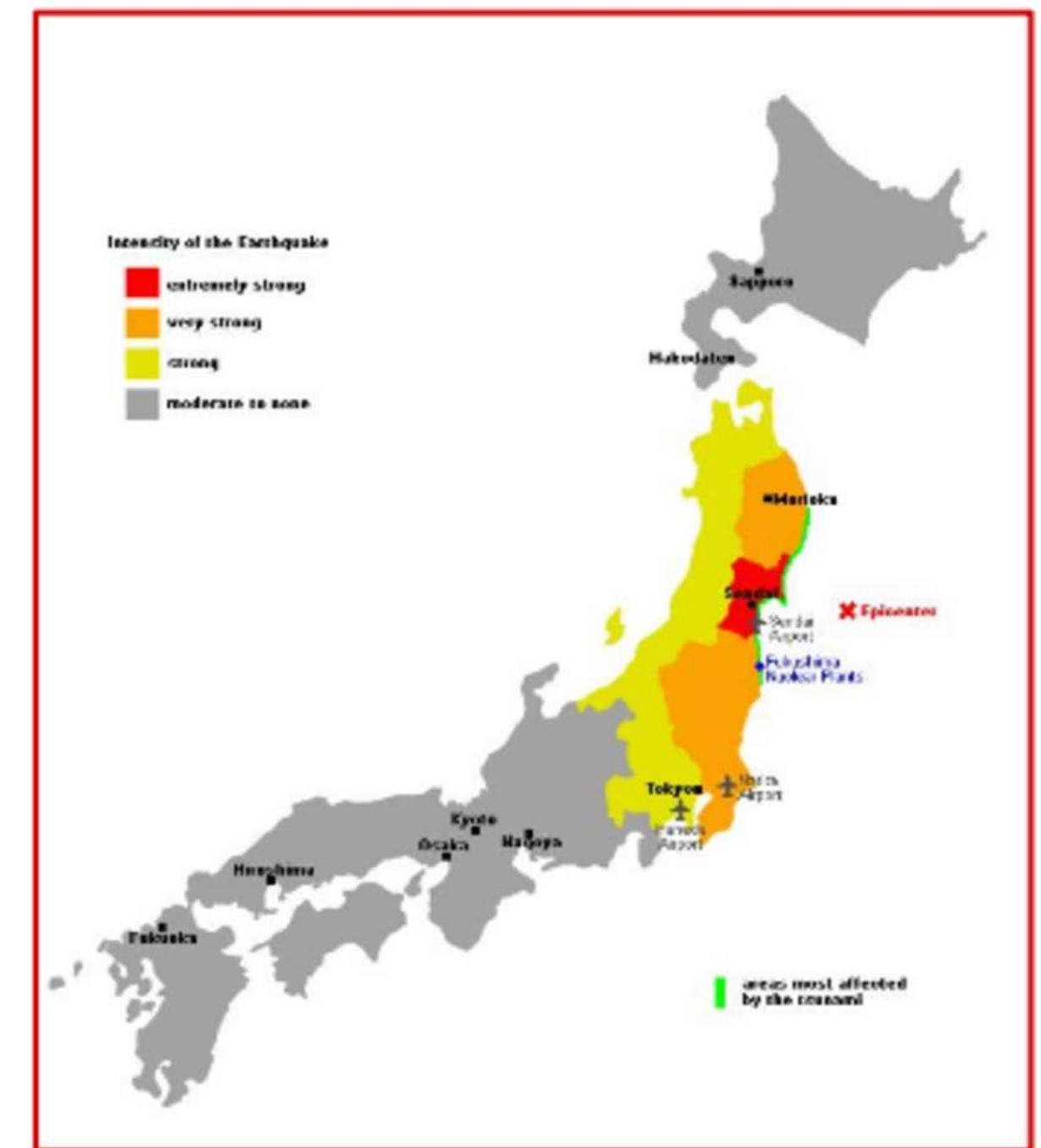
El agrupamiento es especialmente útil para explorar datos de los que no sabes nada. Puede llevar algún tiempo averiguar qué tipo de algoritmo de agrupamiento funciona mejor, pero cuando lo hagas, obtendrás información invaluable sobre tus datos



EJEMPLOS DE AGRUPACIÓN

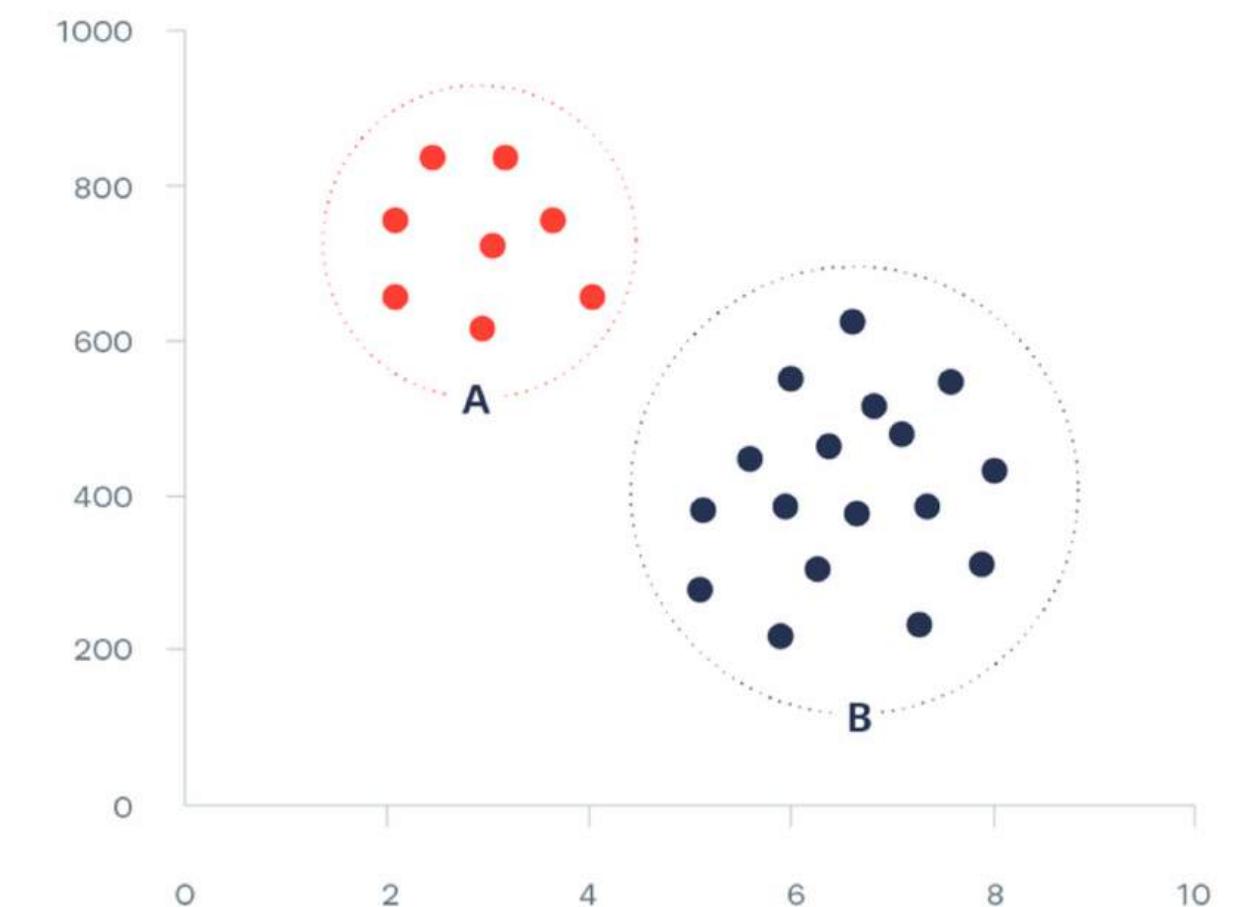
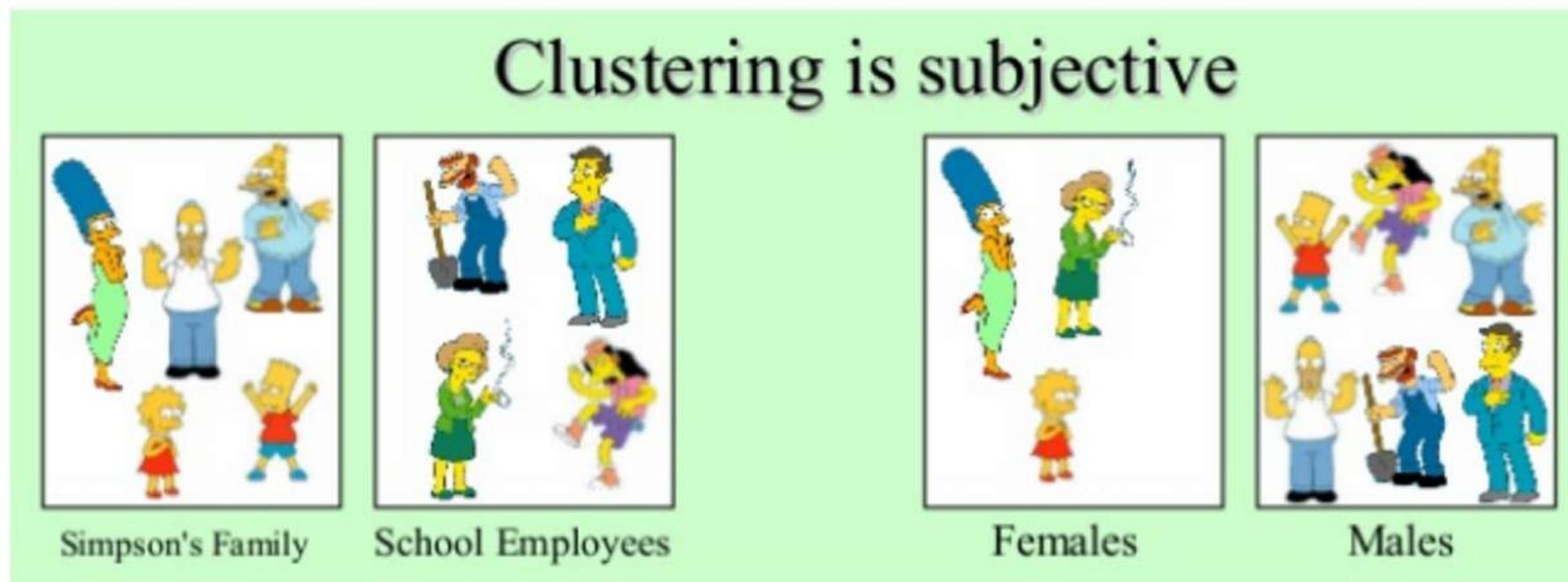
Algunas aplicaciones del mundo real del agrupamiento incluyen:

- La detección de fraudes en seguros
- La categorización de libros en una biblioteca
- La segmentación de clientes en sus compras
- Análisis de terremotos
- Planificación urbana
- Etc



EJEMPLO DE AGRUPACIÓN

What is a natural grouping among these objects?



TEACHABLE MACHINE

<https://teachablemachine.withgoogle.com/>

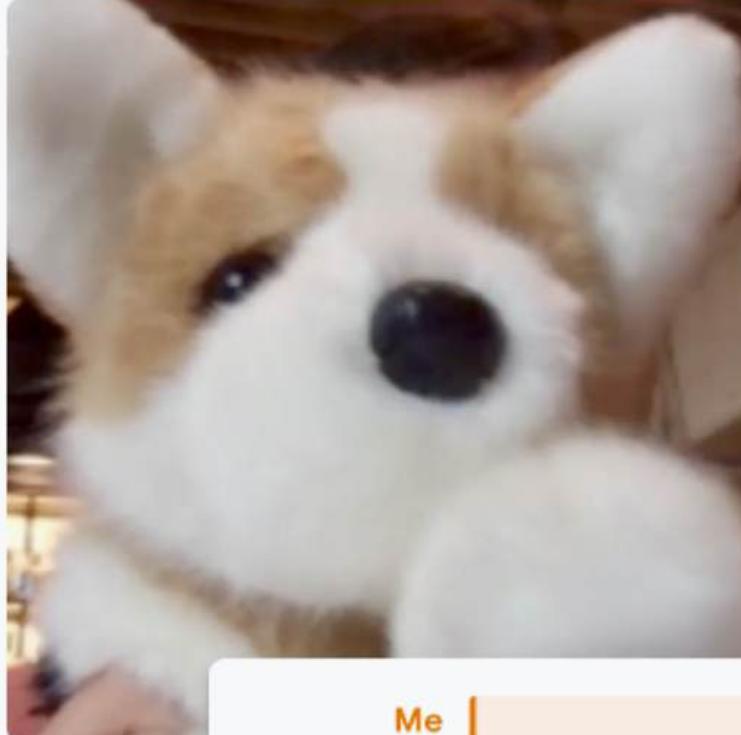
Teachable Machine

Train a computer to recognize your own images, sounds, & poses.

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

[Get Started](#)

↑ ml5 p5.js Coral ↴ node TensorFlow ARDUINO



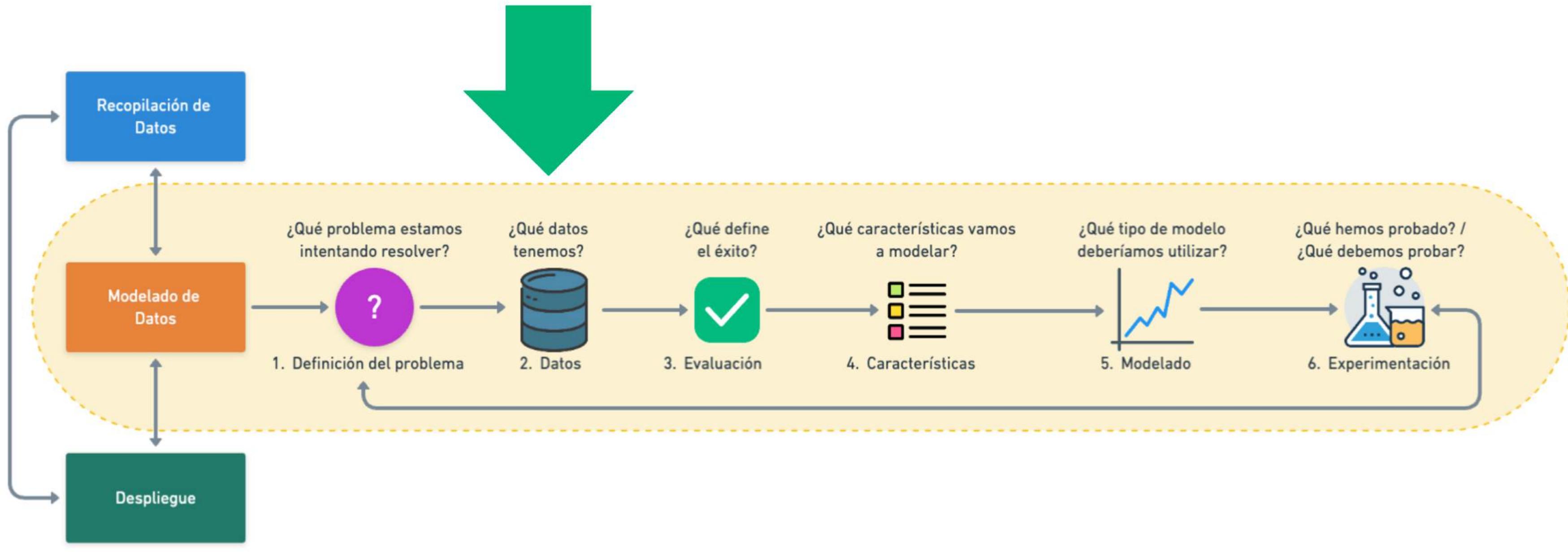
Me

Me + Dog <3

DATOS



FRAMEWORK



TIPOS DE DATOS

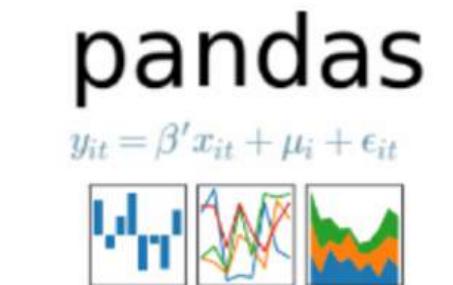
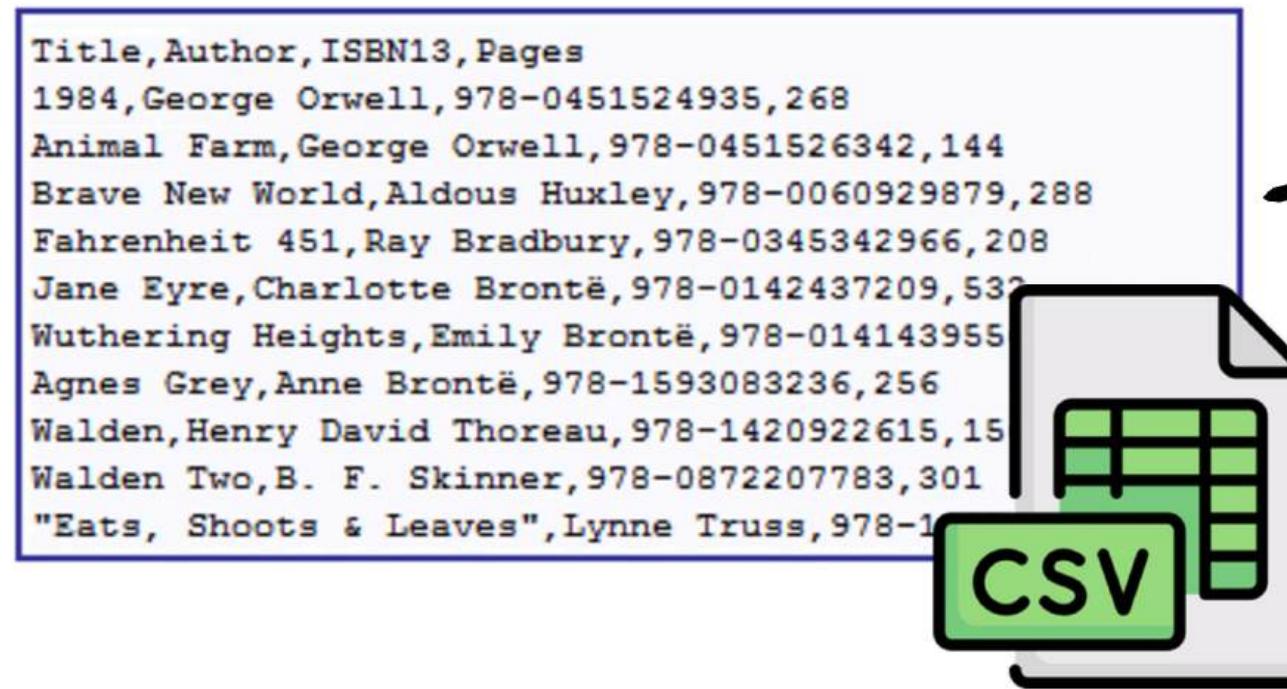
Data Table						
Name	Position	Office	Age	Start date	Salary	
Airi Satou	Accountant	Tokyo	33	2008/11/28	\$162,700	
Angelica Ramos	Chief Executive Officer (CEO)	London	47	2009/10/09	\$1,200,000	
Ashton Cox	Junior Technical Author	San Francisco	66	2009/01/12	\$86,000	
Bradley Greer	Software Engineer	London	41	2012/10/13	\$132,000	
Brenden Wagner	Software Engineer	San Francisco	28	2011/06/07	\$206,850	
Brielle Williamson	Integration Specialist	New York	61	2012/12/02	\$372,000	
Bruno Nash	Software Engineer	London	38	2011/05/03	\$163,500	
Caesar Vance	Pre-Sales Support	New York	21	2011/12/12	\$106,450	
Cara Stevens	Sales Assistant	New York	46	2011/12/06	\$145,600	
Cedric Kelly	Senior Javascript Developer	Edinburgh	22	2012/03/29	\$433,060	

Show 10 entries Search:

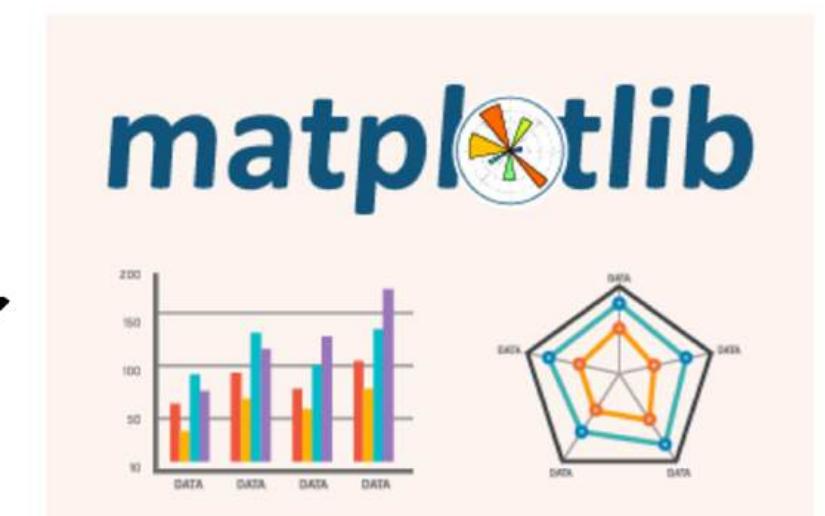
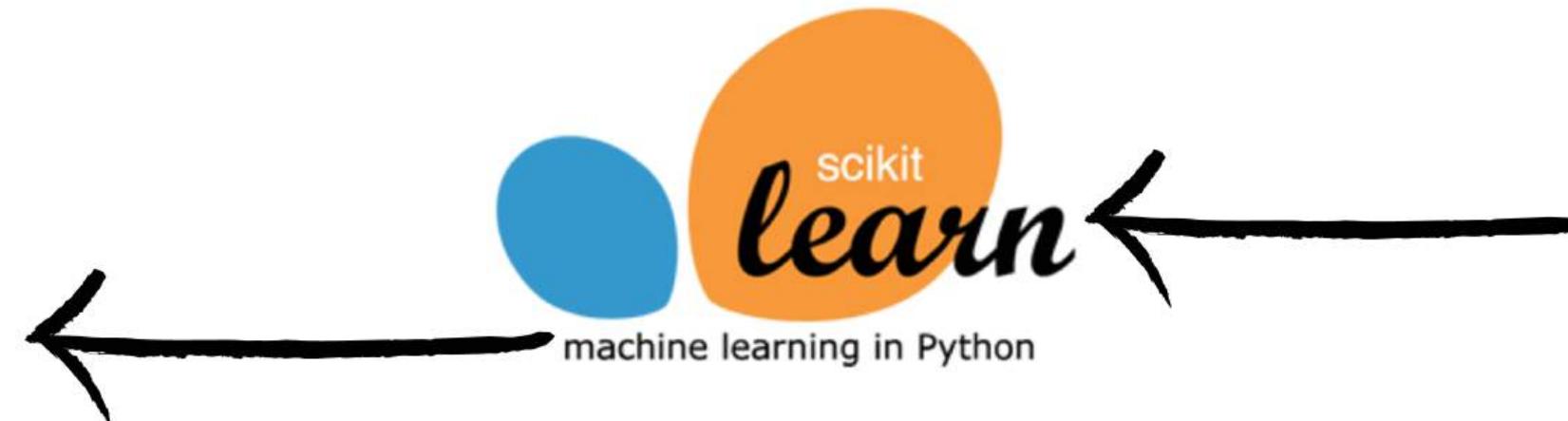
Showing 1 to 10 of 57 entries Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) Next



WORKFLOW DATA SCIENCE



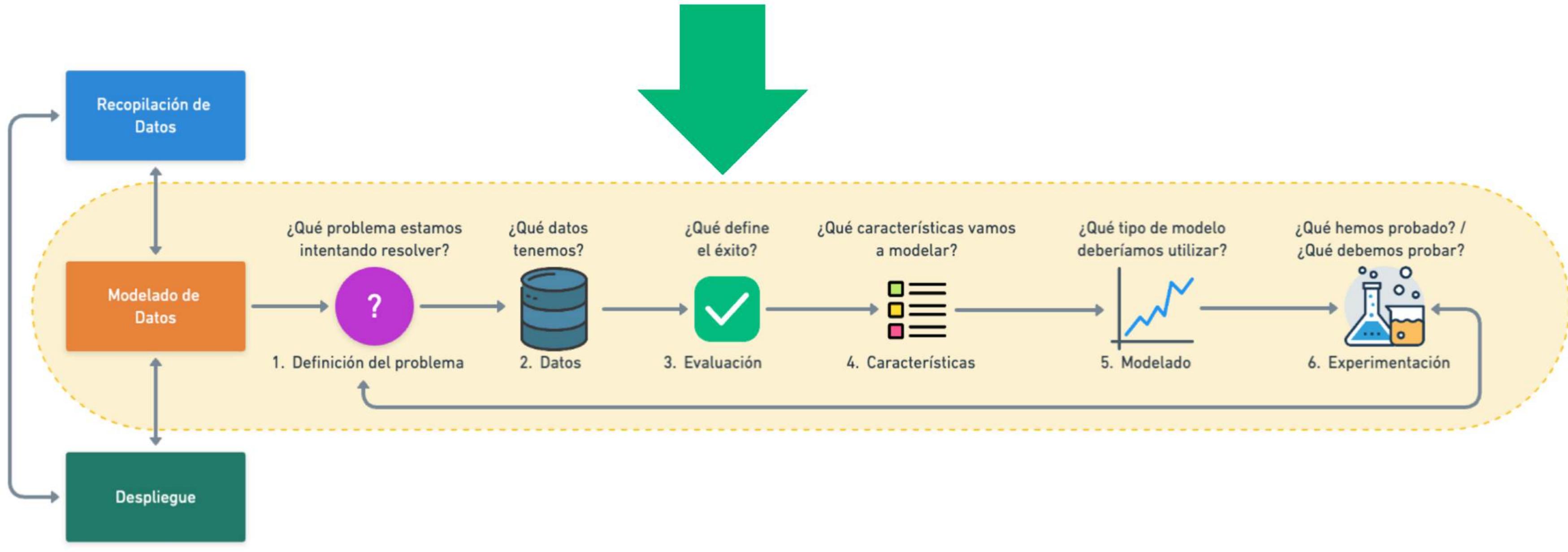
Heart Disease
Prediction
Model Building



EVALUACIÓN



FRAMEWORK



TIPOS DE EVALUACIÓN

Classification

Accuracy

Precision

Recall

Regression

Mean absolute error (MAE)

Mean squared error (MSE)

Recommendation

Precision at K

CLASIFICACIÓN

¿En qué consiste la clasificación?



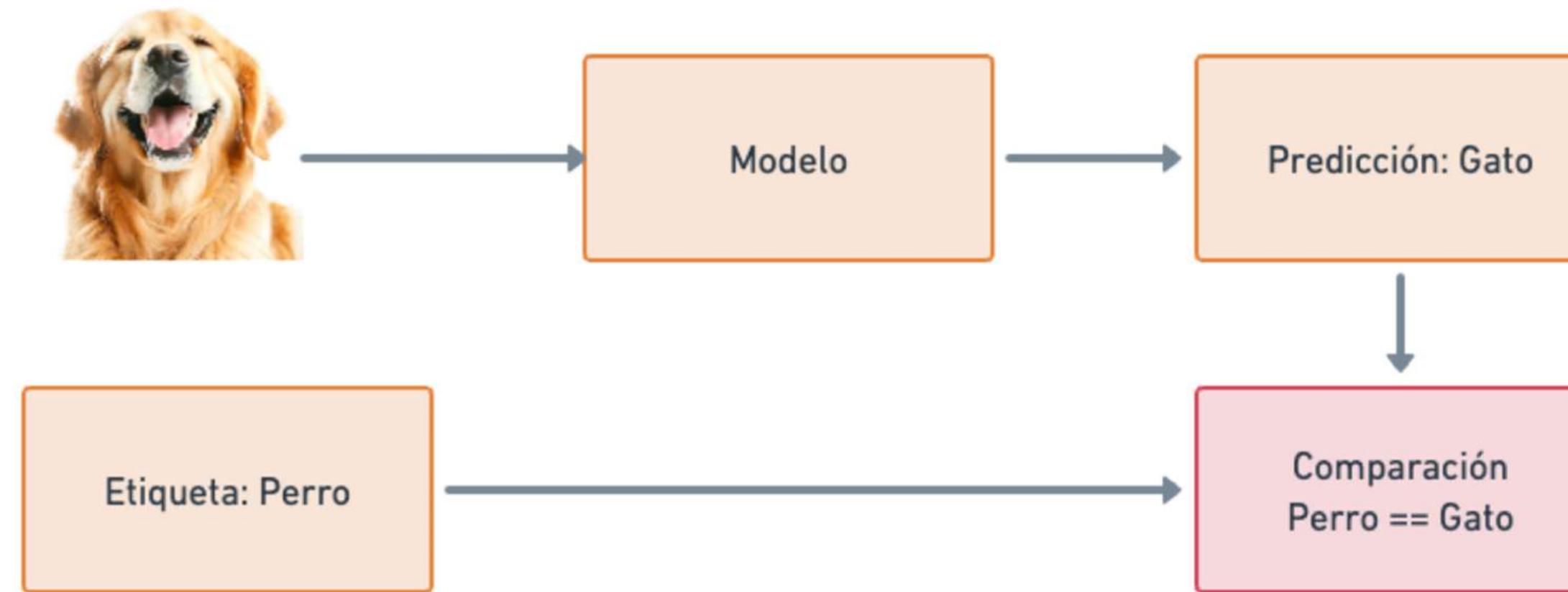
La predicción fue
Correcta



La predicción fue
Incorrecta

1. Crearemos un modelo usando los datos train
2. Utilizaremos los datos Test y el modelo para hacer una predicción
3. Compararemos la predicción contra el resultado real

EVALUACIÓN



Repetiremos esto para todas nuestras fotos y acabaremos con el número de casos correctos e incorrectos donde podremos crear una [matriz de confusión](#)

MATRÍZ DE CONFUSIÓN

		Predicción	
		Positivo	Negativo
Actual	Positivo	Verdadero Positivo	Falso Negativo
	Negativo	Falso Positivo	Verdadero Negativo

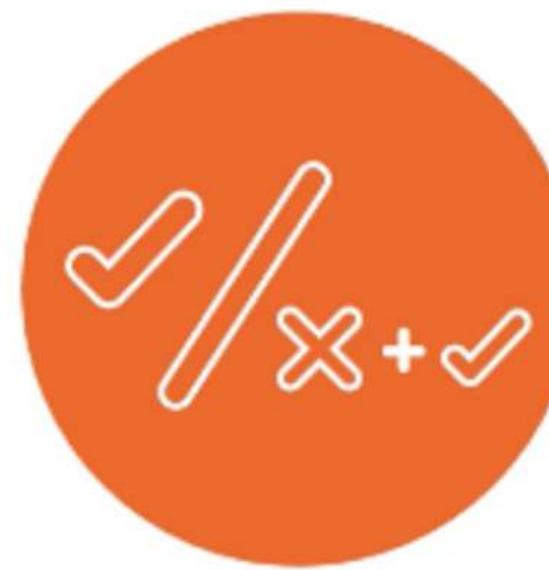
		Predicción	
		Positivo	Negativo
Actual	Positivo	True Positive (TP)	False Negative (FN)
	Negativo	False Positive (FP)	True Negative (TN)

EVALUACIÓN DE MODELOS DE CLASIFICACIÓN

¿Qué métricas podemos usar?



Exactitud
(Accuracy)



Sensibilidad
(Recall)



Precisión



Puntuación F1
(*F*₁ Score)

MÉTRICAS - EXACTITUD



- Predicciones correctas
- Dividido (/)
- Número total de predicciones

Es una buena métrica para clases balanceadas, osea que tengan un número parecido de perros y gatos

Por ejemplo, si tuviésemos 98 imágenes de perros y 2 de gatos, si nuestro modelo fuese una línea que solo predice siempre el perro, tendríamos una exactitud de 98%

MÉTRICAS - SENSIBILIDAD (RECALL)



Permite encontrar todos los casos relevantes:

- Número de verdadero positivo
- Dividido
- Número de verdadero positivo + falsos negativos

$$48 / (48 + 2) = 0.96$$

MÉTRICAS - PRECISIÓN



Permite encontrar solo los casos relevantes:

- Número de verdadero positivo
- Dividido
- Número de verdadero positivos + falsos positivos

$$48 / (48 + 5) = 0.905$$

COMPENSACIÓN

Normalmente habrá una compensación entre Sensibilidad (Recall) y Precisión

Sensibilidad te encontrará los casos relevantes, mientras Precisión encontrará la proporción de esos casos relevantes que realmente son relevantes

PUNTUACIÓN F1 (F1 SCORE)



Es una combinación entre **Sensibilidad (Recall)** y **Precisión**

Utilizando la media harmónica

$$F1 = \frac{2 * (\text{sensibilidad} * \text{precisión})}{(\text{sensibilidad} + \text{precisión})}$$

$$2 * (0.96 * 0.905) / (0.96 + 0.905) = 0.93$$

¿POR QUÉ USAMOS LA MEDIA HARMÓNICA?

Castigamos valores extremos:

Sensibilidad = 0

Precisión = 1

$$F1 = 2 * (\text{sensibilidad} * \text{precisión}) / (\text{sensibilidad} + \text{precisión})$$

$$F1 = 0$$

¿Qué métrica debo usar?

Depende de cada situación

¿QUÉ MÉTRICA DEBO UTILIZAR?

Por ejemplo – una enfermedad

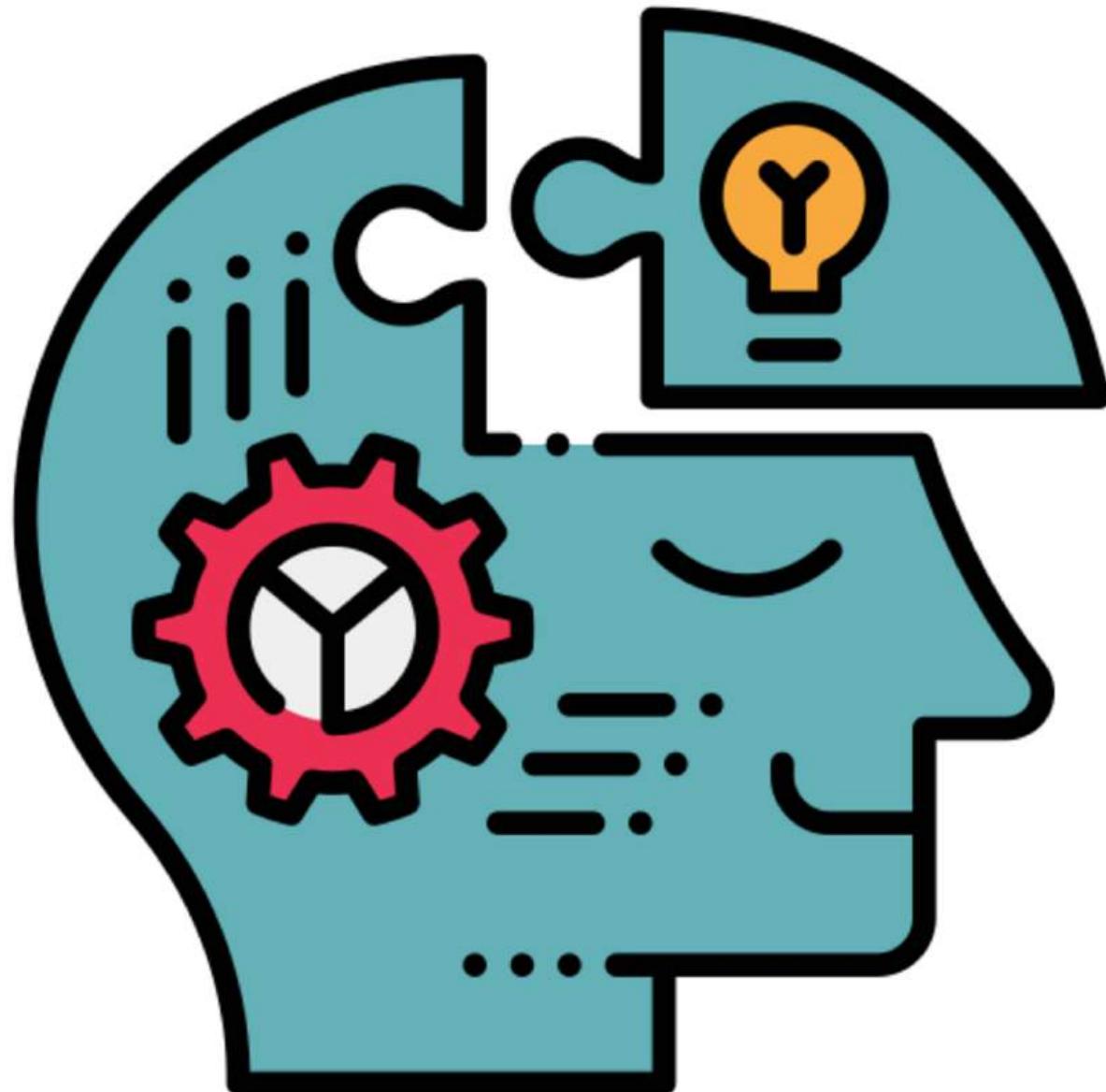
Primero de todo, ¿son las clases balanceadas? Probablemente no, tendremos una compensación entre sensibilidad y precisión.

Y para muchos de estos modelos son simplemente para tener una idea rápida, por ejemplo para cáncer de próstata, podemos hacer una prueba rápida antes de cualquier prueba invasiva.

¿El modelo debería reducir falsos positivos o falsos negativos?

En una enfermedad es mejor reducir falsos negativos e incrementar los falsos positivos, porque queremos poder incluir más de los positivos para incluir cuántos más positivos posibles. Esto tendrá un coste de incrementar nuestros falsos positivos.

APRENDIZAJE AUTOMÁTICO



Aprendizaje automático no es un proceso tan sencillo, siempre se necesita contexto e información de los expertos y conocimiento del dominio (domain knowledge)

MÉTRICAS - REGRESIÓN

- Regresión - sirve para predecir variables continuas



MAE

Mean Absolute
Error

Error Absoluto
Medio



MSE

Mean Squared
Error

Error Cuadrado
Medio
o
Mínimos
Cuadrados



RMSE

Root Mean Square
Error

Error Cuadrático
Medio

ERROR ABSOLUTO MEDIO (MAE)

Es básicamente la media del error absoluto de cada predicción. Muy fácil de entender.

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Básicamente haces una predicción, por ejemplo, del precio de una casa y tomas la diferencia con el precio real.

Si es negativo, lo coges como positivo, por eso el absoluto.

Tomas todos los errores que van a ser positivos, (porque los negativos los has convertido a positivos) y tomas la diferencia de cada predicción y haces la media.

ERROR CUADRADO MEDIO (MAE)

Es básicamente la media del error absoluto de cada predicción pero cuadrado

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

Es básicamente lo mismo la media del error absoluto de cada predicción, pero al cuadrado. Y esto por qué es al cuadrado? Es porque es para penalizar los errores más grandes.

Si tú tienes un número grande y lo haces al cuadrado, pues el error será aún más grande.

Entonces tú tienes un error muy grande, muchos errores muy grandes

ERROR CUADRÁTICO MEDIO (RMSE)

Es básicamente la media del error absoluto de cada predicción pero cuadrado y después cogiendo la raíz cuadrada

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

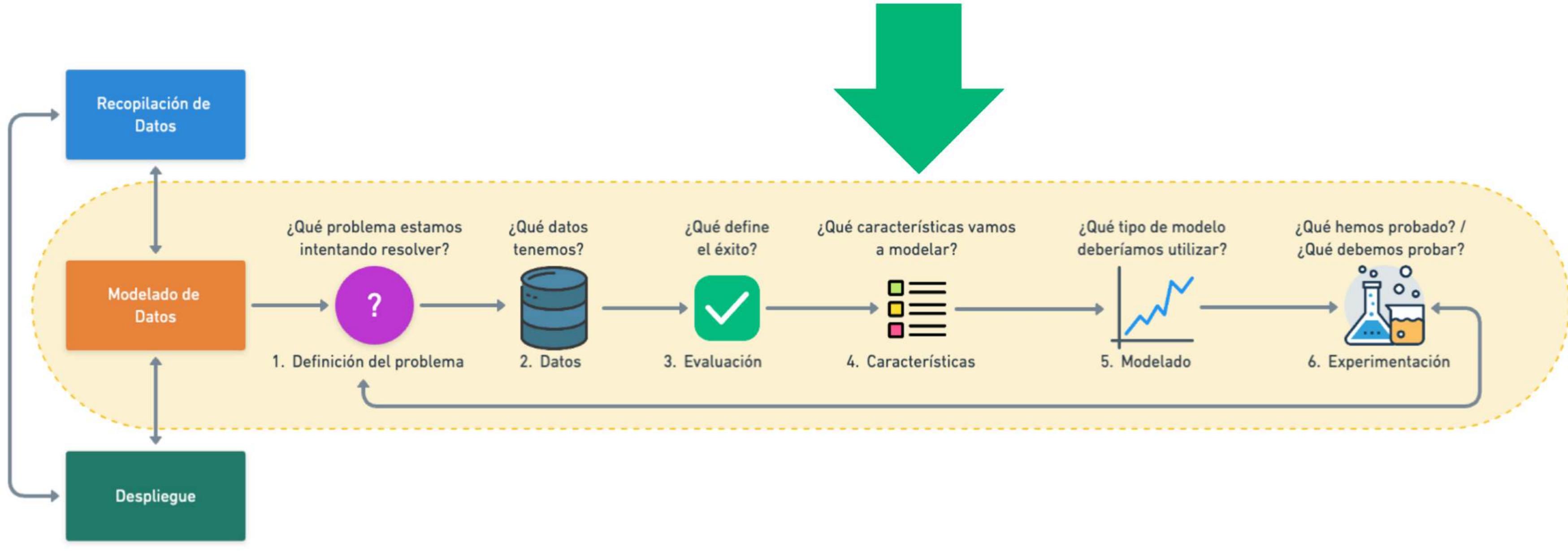
A diferencia del anterior, este si te permite incluir los errores grandes, por eso se obtiene la raíz.

Por ende, es el más usado...

CARACTERÍSTICAS



FRAMEWORK



CARACTERÍSTICAS

ID	Feature variables				Target variable
	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110Kg	M	81	4	Yes
5681	64Kg	F	61	1	No
7911	81Kg	M	57	0	No

Table 1.0: Patient records

CARACTERÍSTICAS

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?	Derived feature
4326	110Kg	M	81	4	Yes	visit in last year?
5681	64Kg	F	61	1	No	Yes
7911	81Kg	M	57	0	No	No

Table 1.0: Patient records

Numerical features

Categorical features

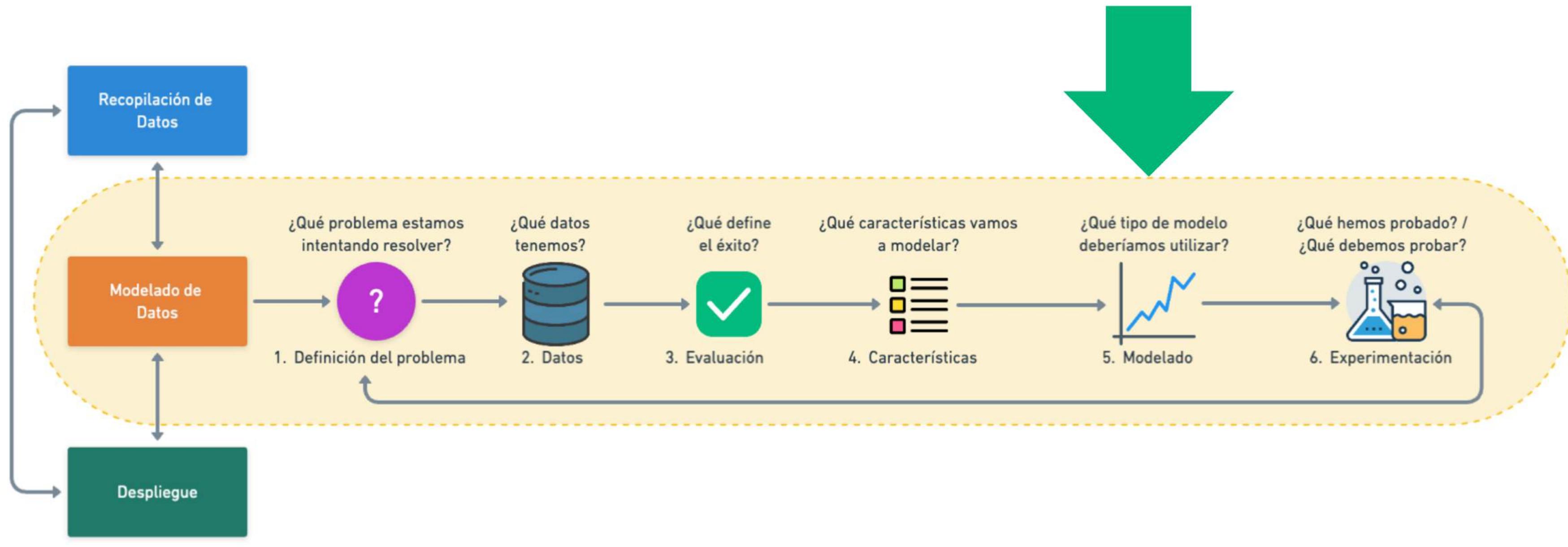
Feature engineering

Looking at different features of data and creating new ones/altering existing ones

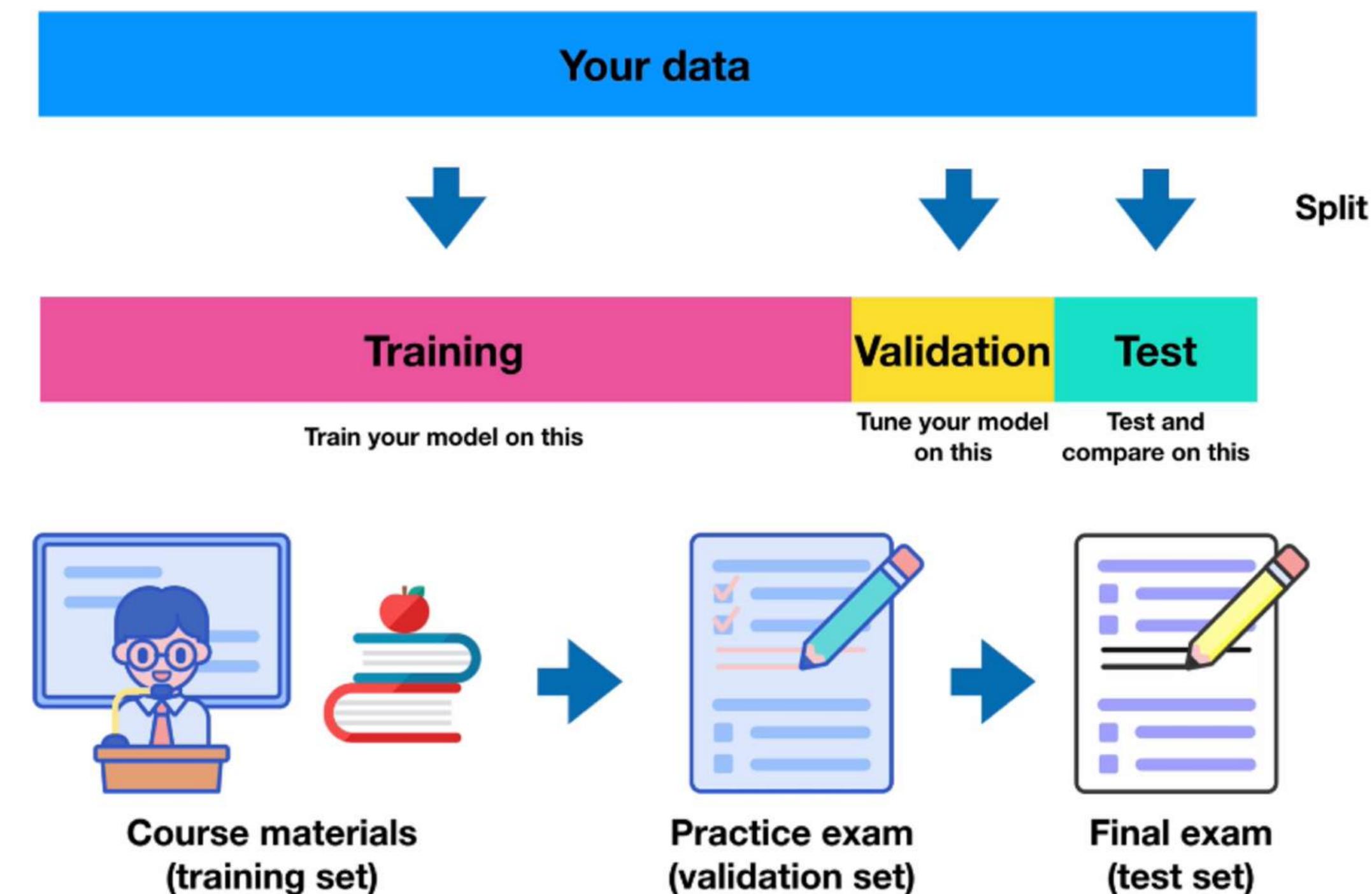
MODELADO



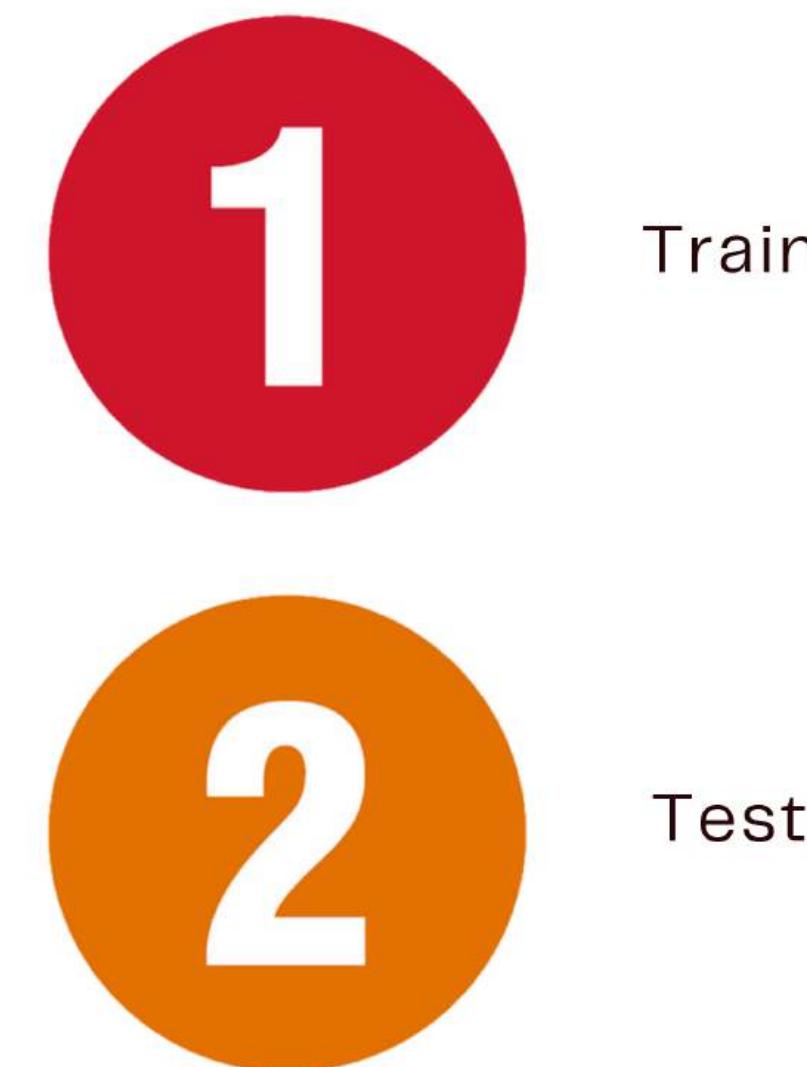
FRAMEWORK



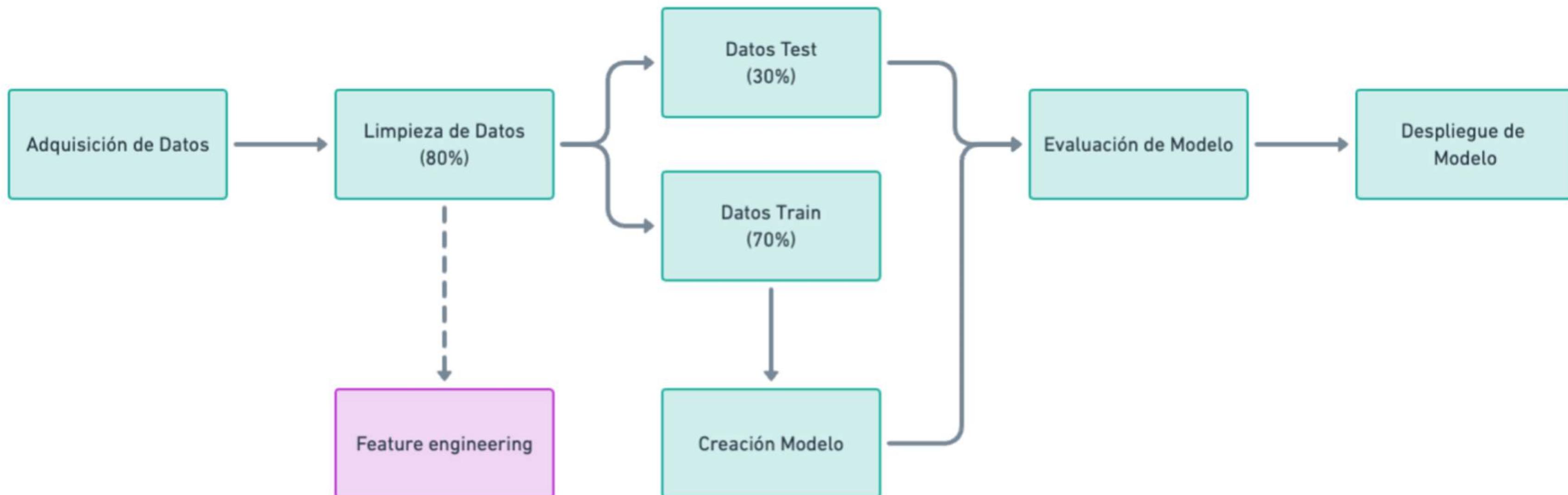
PROCESO DE APRENDIZAJE



PROCESO DE APRENDIZAJE



PROCESO DE APRENDIZAJE



PARTES DEL MODELADO

1. Choosing and training a model



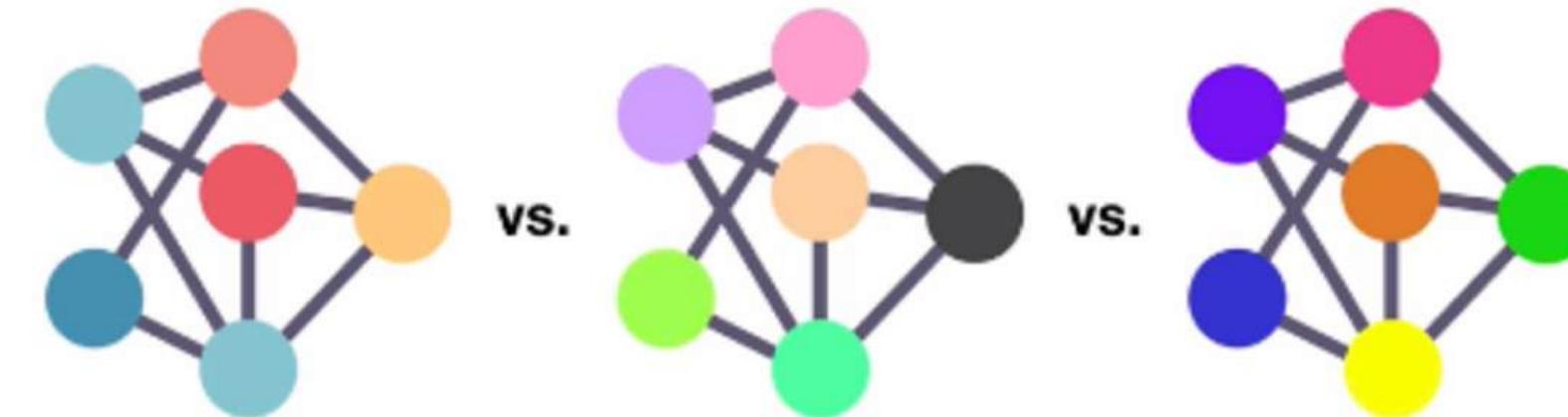
or



2. Tuning a model



3. Model comparison



ELIGIENDO EL MODELO



Problem 1

Model 1



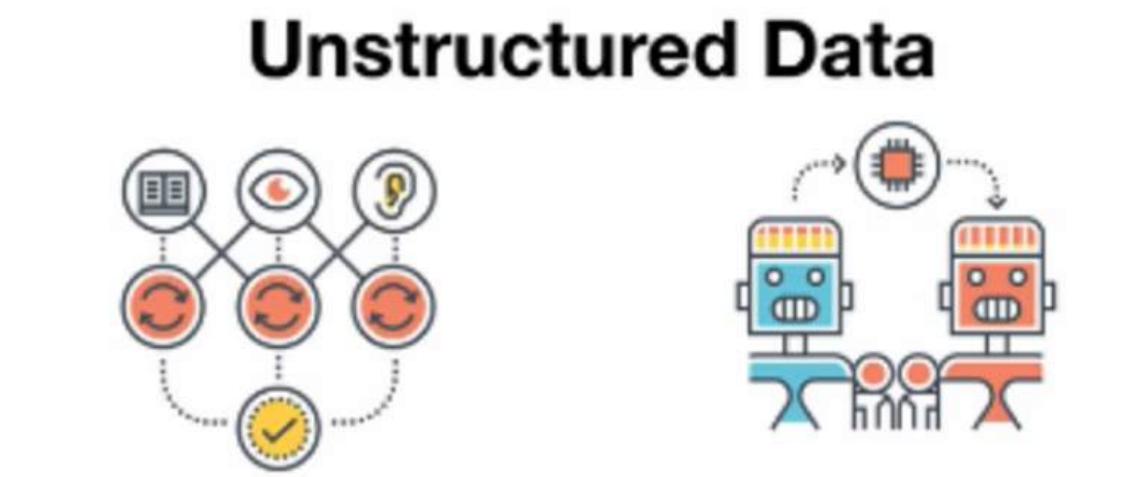
Problem 2

Model 2

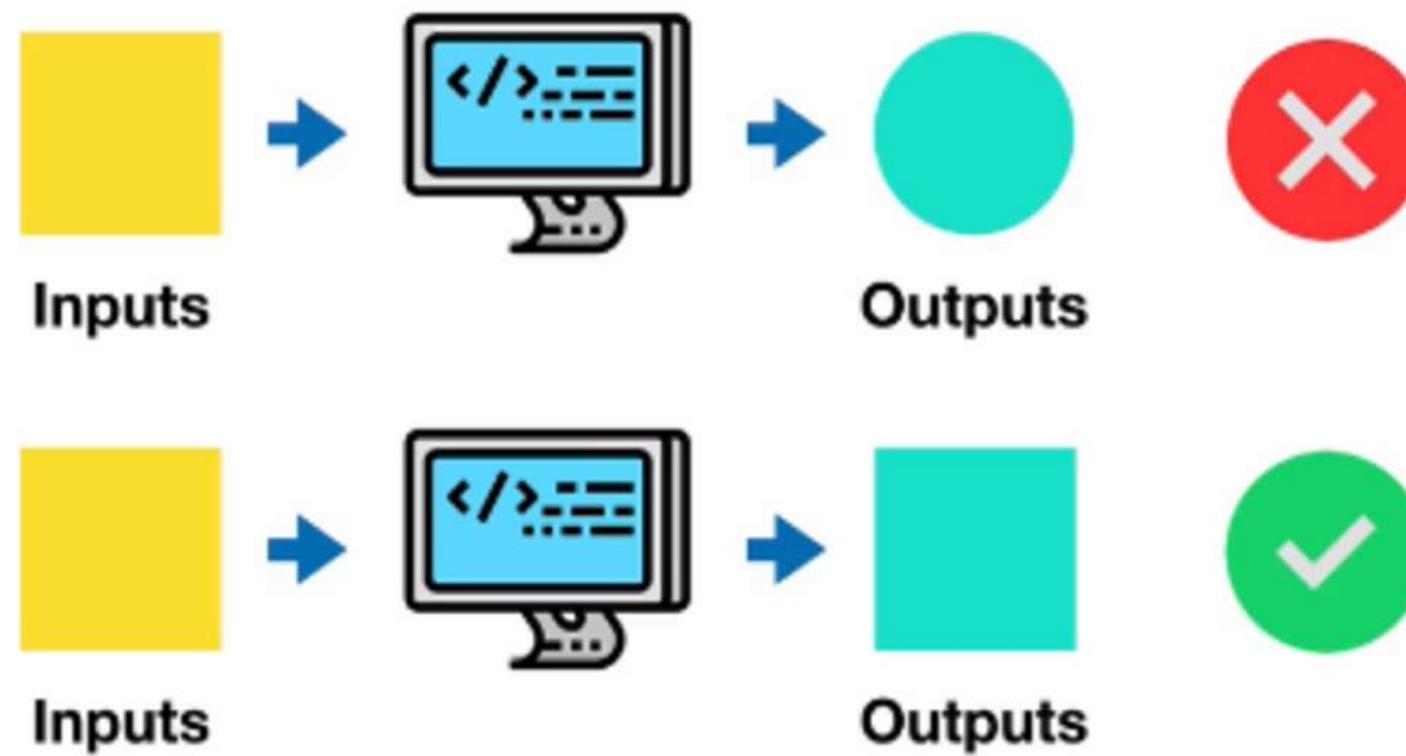


CatBoost

Random Forest



ENTRENANDO UN MODELO

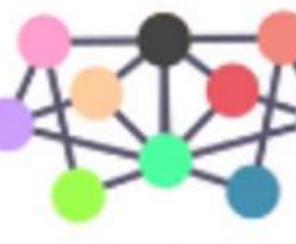


X (data)					y (label)
ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110kg	M	81	4	Yes
5681	64kg	F	61	1	No
7911	81kg	M	57	0	No

Table 1.0: Patient records

TIEMPOS EN EL EXPERIMENTO

Experiment

			Accuracy	Training time
1	 →  → 		87.5%	3 min
2	 →  → 		91.3%	92 min
3	 →  → 		94.7%	176 min

TUNING DEL MODELO



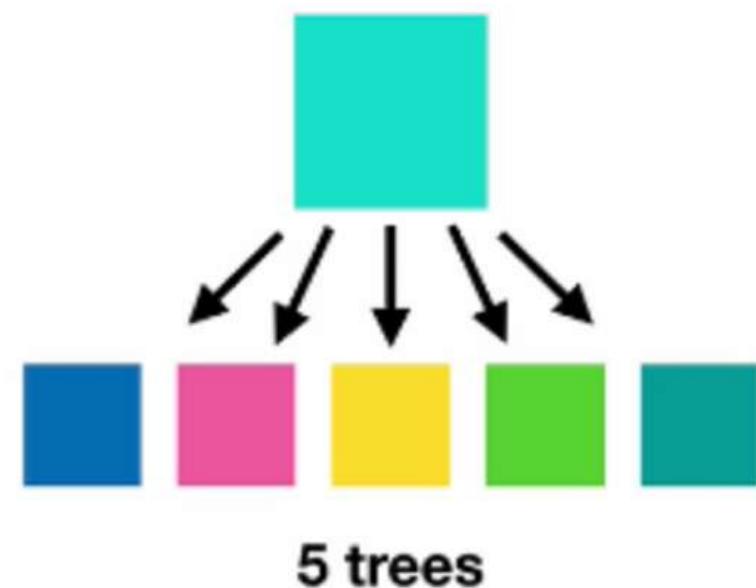
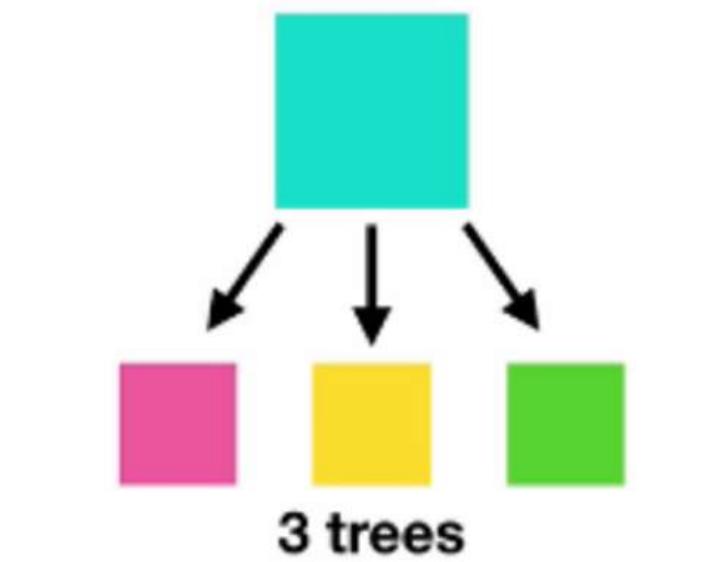
Cooking time: 1 hour
Temperature: 180°C



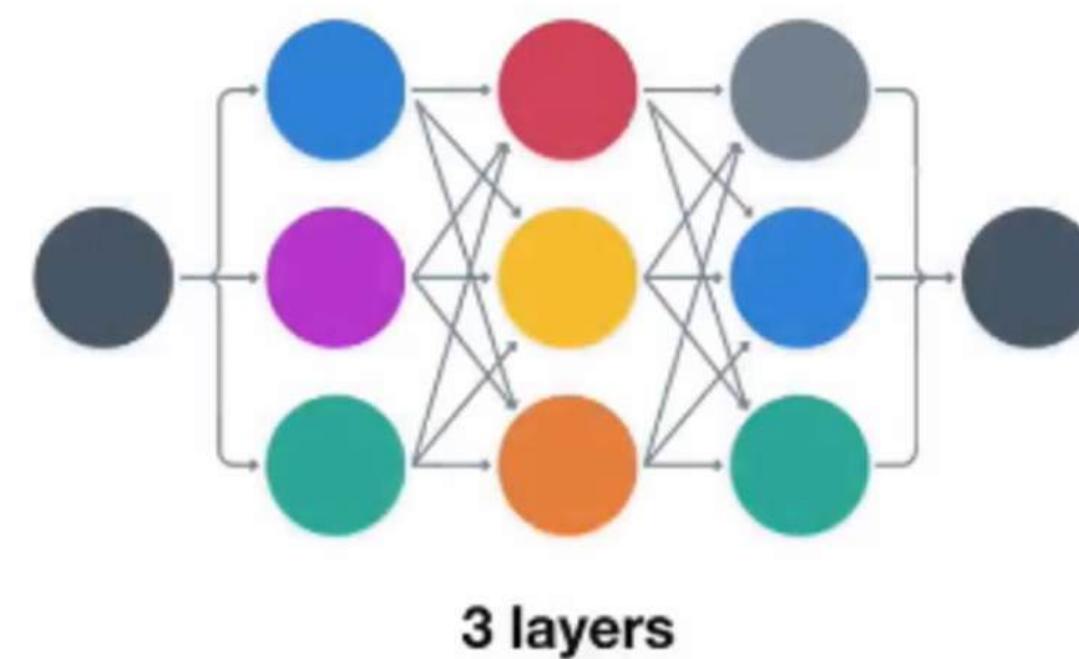
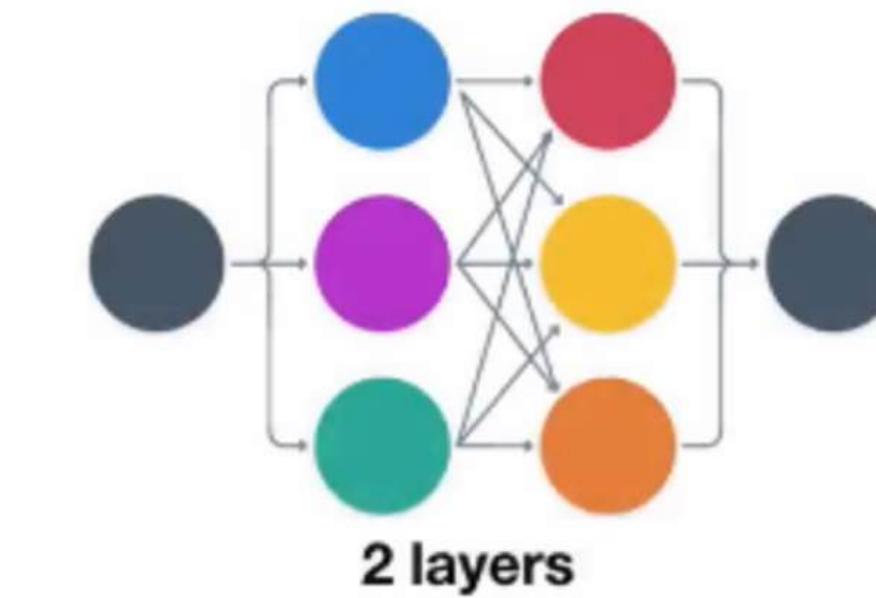
Cooking time: 1 hour
Temperature: 200°C

TUNING DEL MODELO

Random Forest



Neural Networks



TESTING DEL MODELO



Data Set

Performance

Training

98%

Test

96%



Data Set

Performance

Training

64%

Test

47%

Data Set

Performance

Training

93%

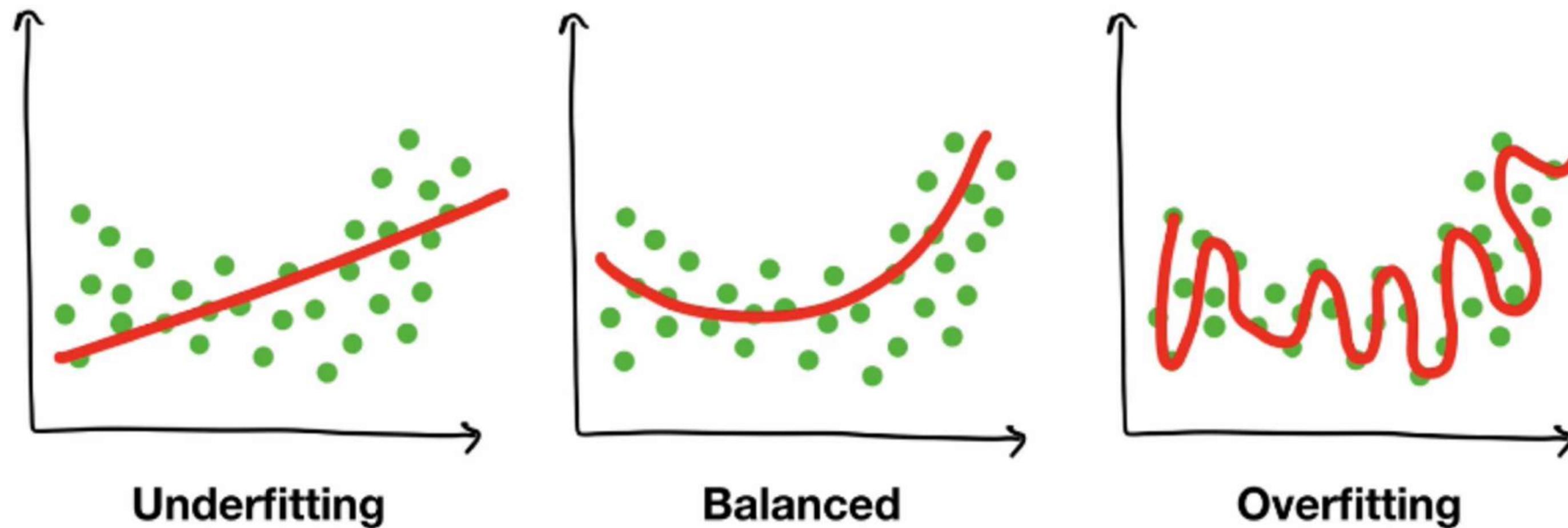
Test

99%

Underfitting
(potential)

Overfitting
(potential)

TESTING DEL MODELO



DILEMA VARIANZA SESGO

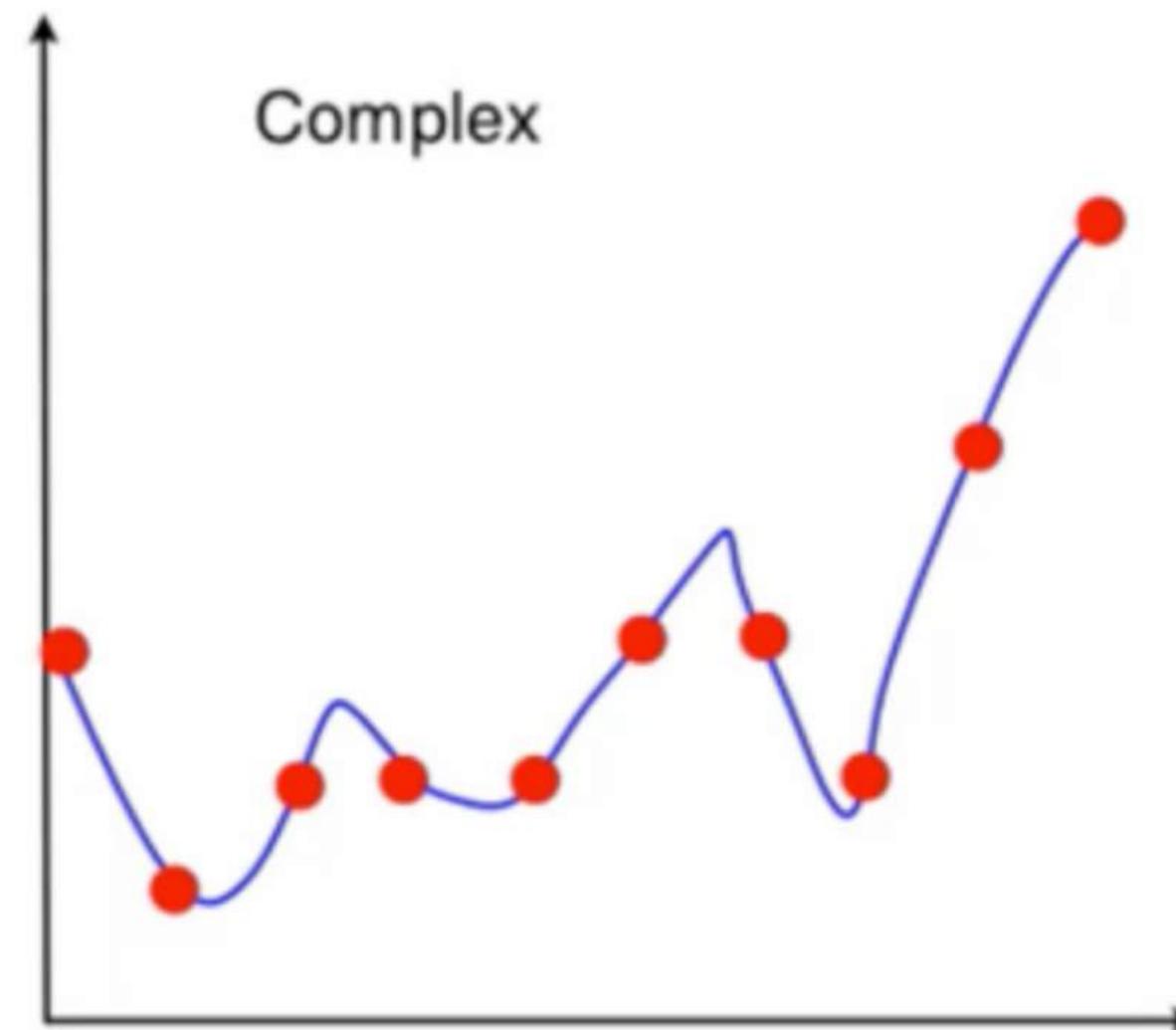
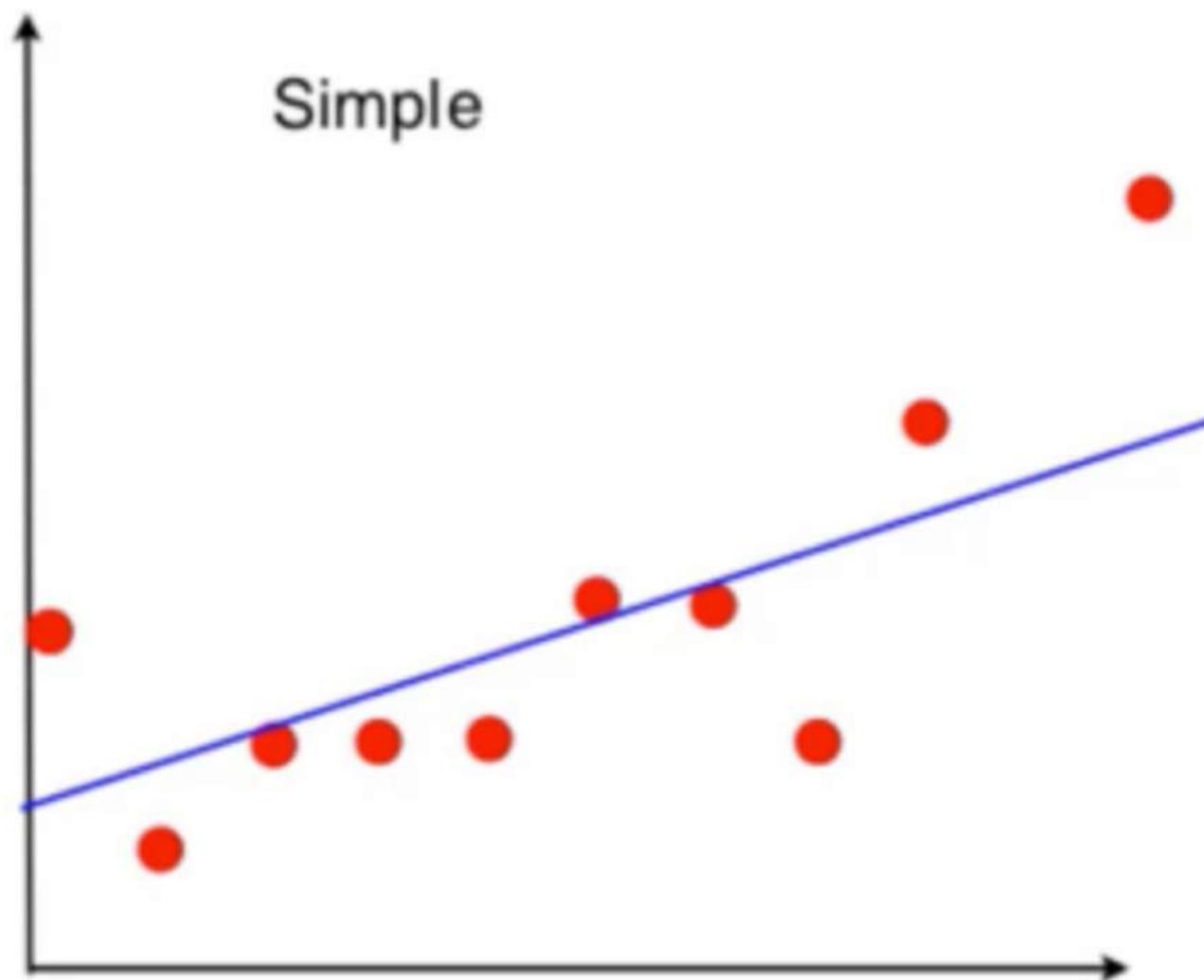
Es un tema muy importante para la evaluación de modelos

Realmente es el punto donde estamos dando complejidad al modelo

Las métricas de entrenamiento Train mejoran, pero las validación test empeoran.

Tal vez se pueda traducir overfitting como “sobreajuste” y underfitting como “subajuste” y hacen referencia al fallo de nuestro modelo al generalizar -encajar- el conocimiento que pretendemos que adquieran.

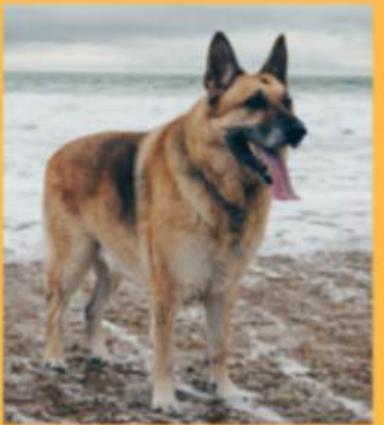
DILEMA VARIANZA SESGO



DIFERENCIAS

Underfitting

Entreno al modelo con
1 sola raza de perro



Muestra nueva:
¿Es perro?



NO
FALLO

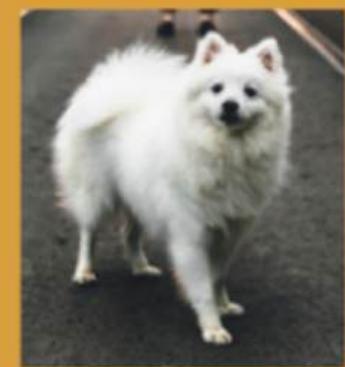
La máquina fallará en reconocer al perro por falta de
suficientes muestras. No puede generalizar el conocimiento.

Overfitting

Entreno al modelo con
10 razas de perro color marrón



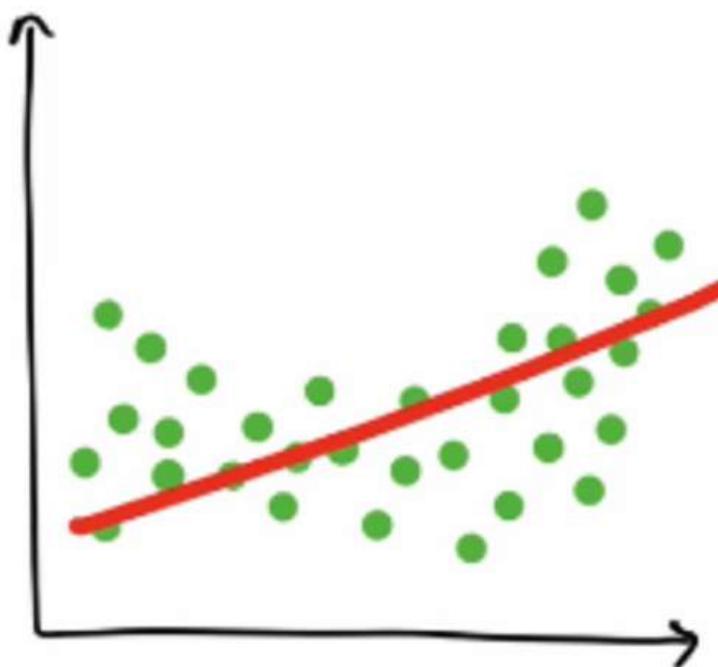
Muestra nueva:
¿Es perro?



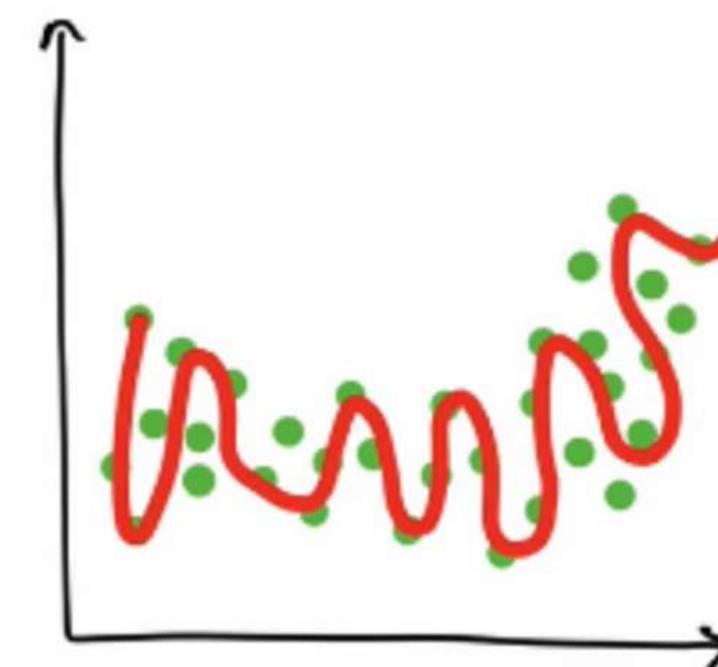
NO
FALLO

La máquina fallará en reconocer un perro nuevo porque no tiene
estrictamente los mismos valores de las muestras de
entrenamiento.

DIFERENCIAS



Underfitting



Overfitting

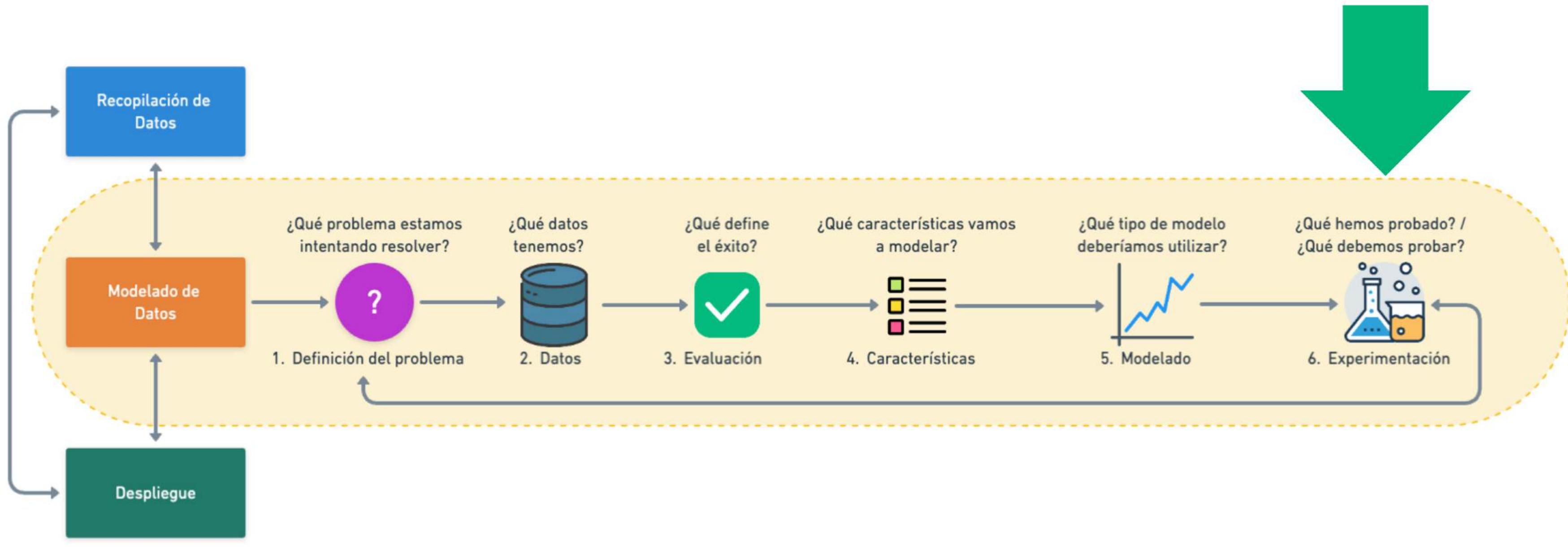
- Try a more advanced model
- Increase model hyperparameters
- Reduce amount of features
- Train longer

- Collect more data
- Try a less advanced model

EXPERIMENTACIÓN



FRAMEWORK



PREGUNTAS Y RESPUESTAS

Mtro. Alfonso Gregorio Rivero Duarte

Senior Data Manager - CBRE

(+52) 5528997069

devil861109@gmail.com

<https://www.linkedin.com/in/alfonso-gregorio-rivero-duarte-139a9225/>

