ACÁMICA

¡Bienvenidos/as a Data Science!

Agenda

Proyecto 2 y Hasta Ahora

Explicación: Machine Learning

Break

Explicación: Aprendizaje Supervisado, Árbol de Decisión

Hands-on training

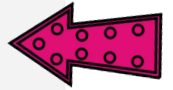
Cierre

¿Dónde estamos?

Cronograma

tiempo	ADQUISICIÓN Y EXPLORACIÓN		MODELADO				DEPLOY
entrega	Exploración de datos	Feature Engineering	Regresión	Optimización de parámetros	Procesam. del lenguaje natural	Sistema de recomendación	Publicación de modelos
	SEM 1	SEM 5	SEM 8	SEM 12	SEM 13	SEM 18	SEM 23
	SEM 2	SEM 6	SEM 9		SEM 14	SEM 19	SEM 24
	SEM 3	SEM 7	SEM 10		SEM 15	SEM 20	
	SEM 4		SEM 11		SEM 16	SEM 21	
					SEM 17	SEM 22	
bloque			APRENDIZAJE SUPERVISADO			APRENDIZAJE NO SUPERVISADO	

BLOQUE 2 (Parte 1)	**Regresión**	Semana 8	Machine Learning Clasificación, Árboles de decisión, Train test split
		Semana 9	KNN, métricas para la clasificación Conceptos generales Machine Learning Práctica integradora
		Semana 10	Regresión (Regresión Lineal, Árboles de decisión, KNN, métricas) Validación Cruzada y selección de modelos
		Semana 11	Datasets Desbalanceados + Teorema de Bayes Curva ROC Trabajo en el proyecto
	Optimización de parámetros	Semana 12	Optimización de parámetros - Validación cruzada y Gridsearch + lanzamiento entrega 4 Trabajo en el proyecto



BLOQUE 2 (Parte 2)

Procesamiento del lenguaje natural	Semana 13	Modelos avanzados - SVM Sesgo y Varianza
	Semana 14	Ensambles, Bagging, Random forest Ensambles, Boosting
	Semana 15	Redes Neuronales: Descenso por gradiente Redes Neuronales: Perceptrón
	Semana 16	Redes Neuronales: Perceptrón Multicapa Redes Neuronales: Repaso
	Semana 17	Procesamiento del lenguaje natural (NLP)
	Semana 18	Trabajo sobre el proyecto Intro aprendizaje no supervisado + Clustering
Sistema de recomendación	Semana 19	Métricas de evaluación para clustering Reducción de dimensionalidad: SVD
	Semana 20	PCA Sistemas de recomendación
	Semana 21	Sistemas de recomendación Ecosistema digital Trabajo sobre el proyecto
	Semana 22	Ecosistema digital Puesta en producción

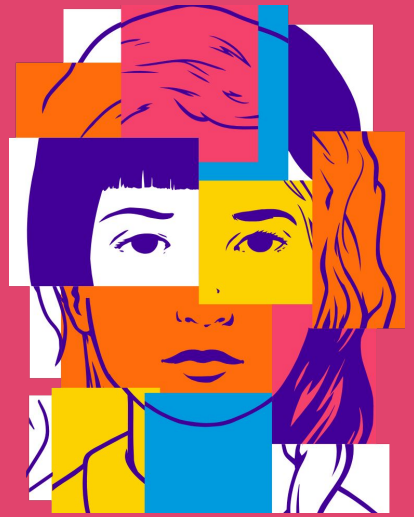
Hasta ahora...

Hasta ahora

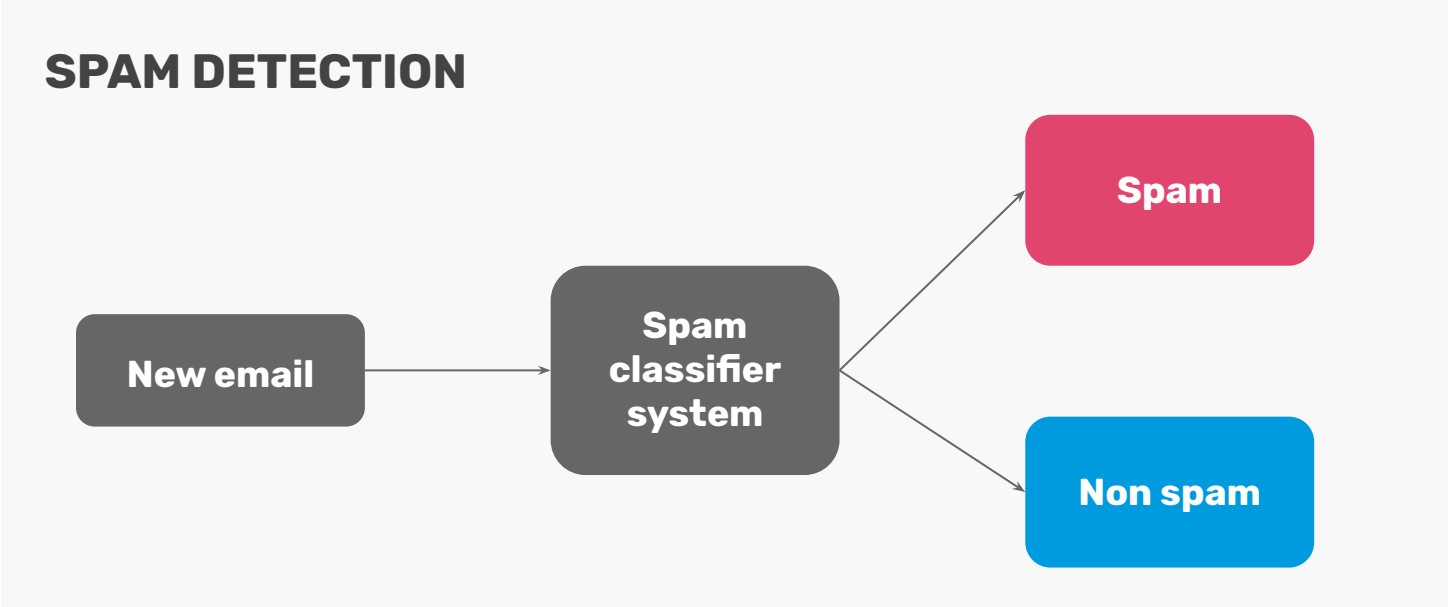
- Vimos un poco de programación con Python en un entorno particular, Jupyter, y aprendimos a utilizar muchas de las librerías típicas del ambiente de Data Science (Numpy, Pandas, Matplotlib, Seaborn, etc.)
- Repasamos varios conceptos de estadística: variables aleatorias, distribuciones, correlación, outliers, etc.
- Aprendimos algunas técnicas de preprocesamiento de datos con Pandas y con Scikit-Learn
- Aplicamos estas herramientas al **Análisis Exploratorio de Datos**

Vamos a seguir profundizando en herramientas (programación y librerías) y en estadística a lo largo de las clases. **¡Pero ahora vamos a ver cómo hace la computadora para aprender de los datos!**

Machine Learning



Machine Learning - Ejemplo clásico



¿Cual de estos mails parece ser spam?

Hola Juan,

Soy Pedro, el socio del proyecto inmobiliario.
Quería avisarte que la reunión del jueves se pasó para el viernes.

Saludos,
Pedro.

Hola juan_86,

Soy Namubi, príncipe de Nigeria.
Preciso que mande su numero de cuenta bancaria y contraseña para transferir herencia millonaria.

Caricias significativas,
Namubi

¿Cual de estos mails parece ser spam?

Hola Juan,

Soy Pedro, el socio del proyecto inmobiliario.
Quería avisarte que la reunión del jueves se pasó para el viernes.

Saludos,
Pedro.

Hola juan_86,

Soy Namubi, príncipe de Nigeria.
Preciso que mande su numero de cuenta bancaria y contraseña para transferir herencia millonaria.

Caricias significativas,
Namubi

¿Cómo distinguieron Spam de no Spam?

¿Como distinguieron Spam de no Spam?

No es una tarea sencilla de realizar, de hecho hoy en día mucha gente es víctima de publicidad engañosa o estafas por medio de mails.

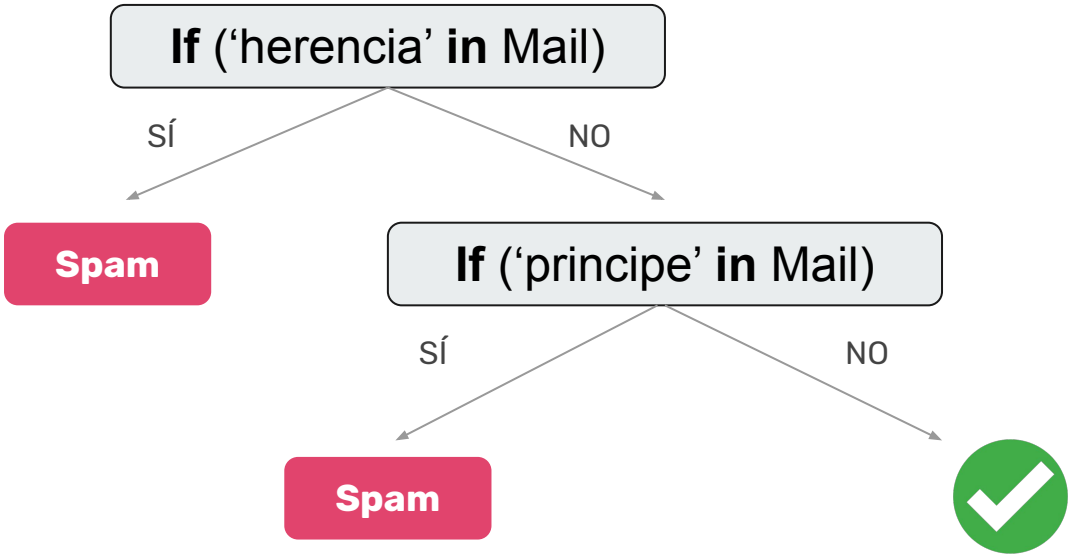
¿Como distinguieron Spam de no Spam?

No es una tarea sencilla de realizar, de hecho hoy en día mucha gente es víctima de publicidad engañosa o estafas por medio de mails.

La tarea implica un procesamiento de alto nivel de abstracción (saber leer, relacionar conceptos, etc...), por lo cual resulta difícil (casi imposible) programar explícitamente un algoritmo que la realice.

¿Cómo se imaginan un algoritmo (programa) que realice esta tarea?

Algoritmo de detección de spam

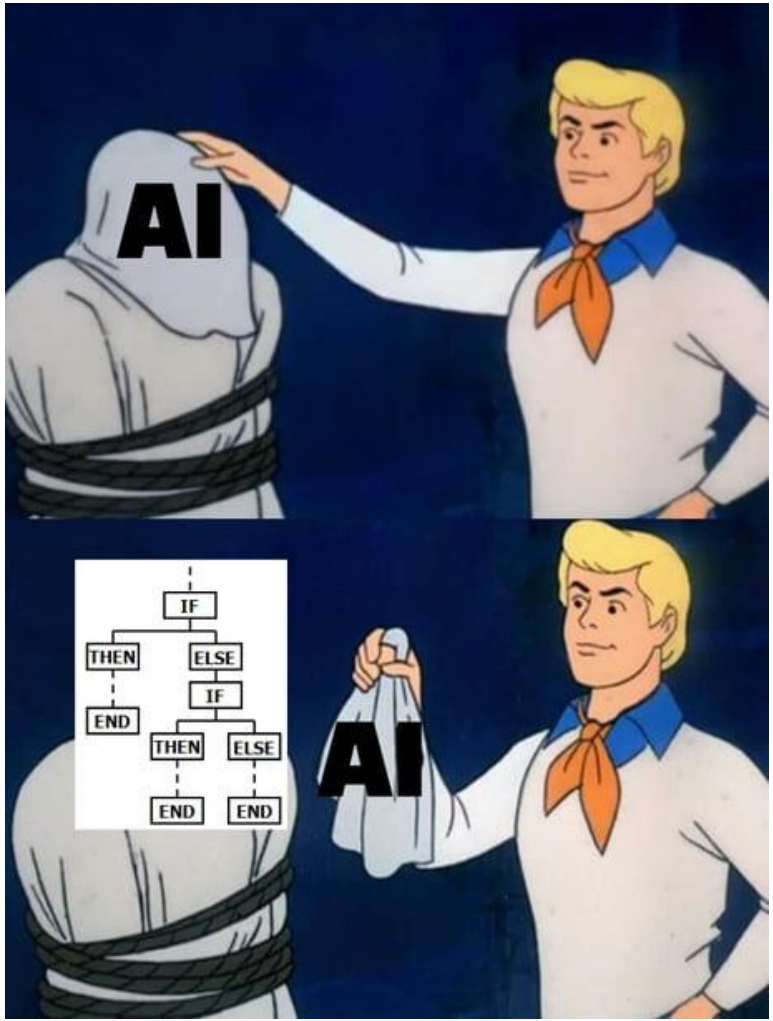


Algoritmo de detección de spam

Un algoritmo construido de esta forma ‘manual’ NO es lo que entendemos por Machine Learning.

Un algoritmo construido de esta forma ‘manual’ NO es lo que entendemos por Machine Learning.

¿POR QUÉ?



Un algoritmo construido de esta forma ‘manual’ NO es lo que entendemos por Machine Learning.

¿POR QUÉ?

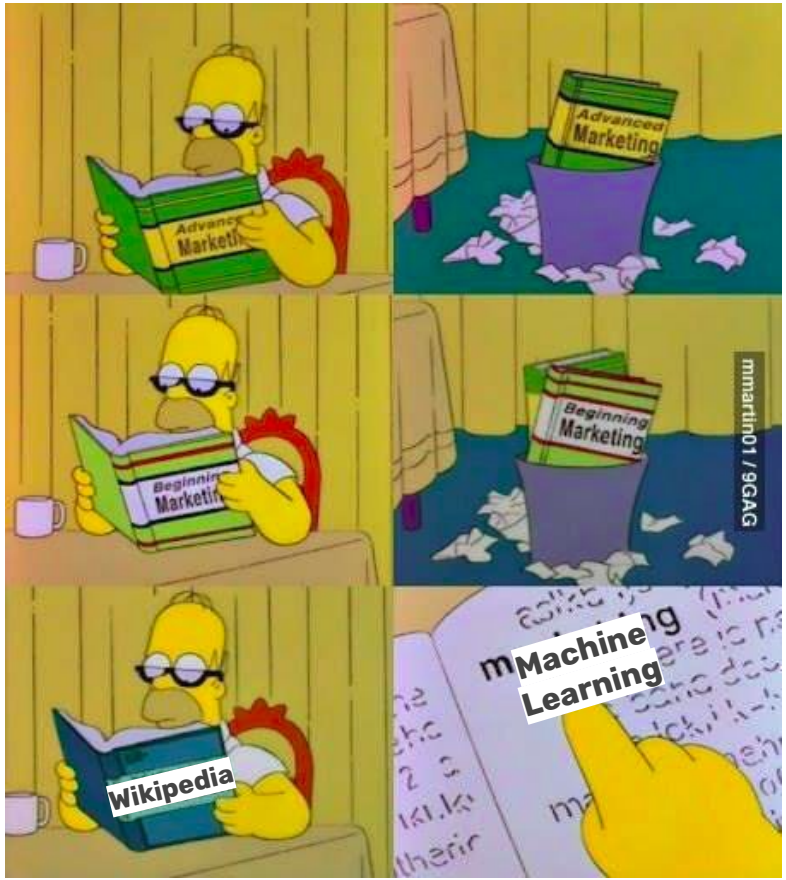
En Machine Learning el modelo (algoritmo) debe aprender a predecir a partir de los datos.

¿Cuál es la definición de MACHINE LEARNING?

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2

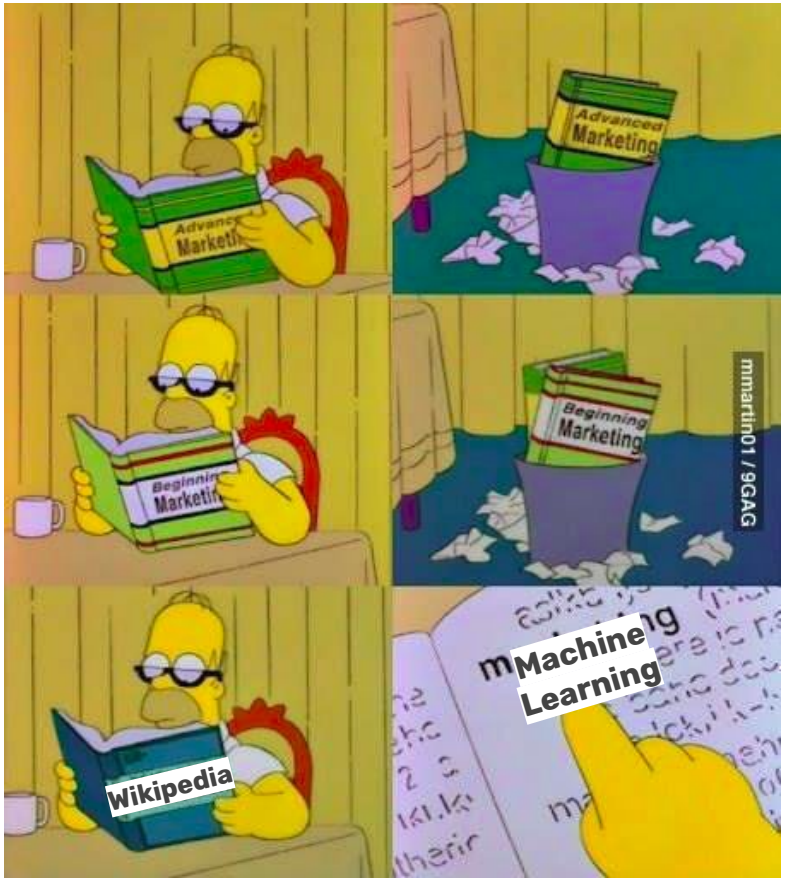
Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.



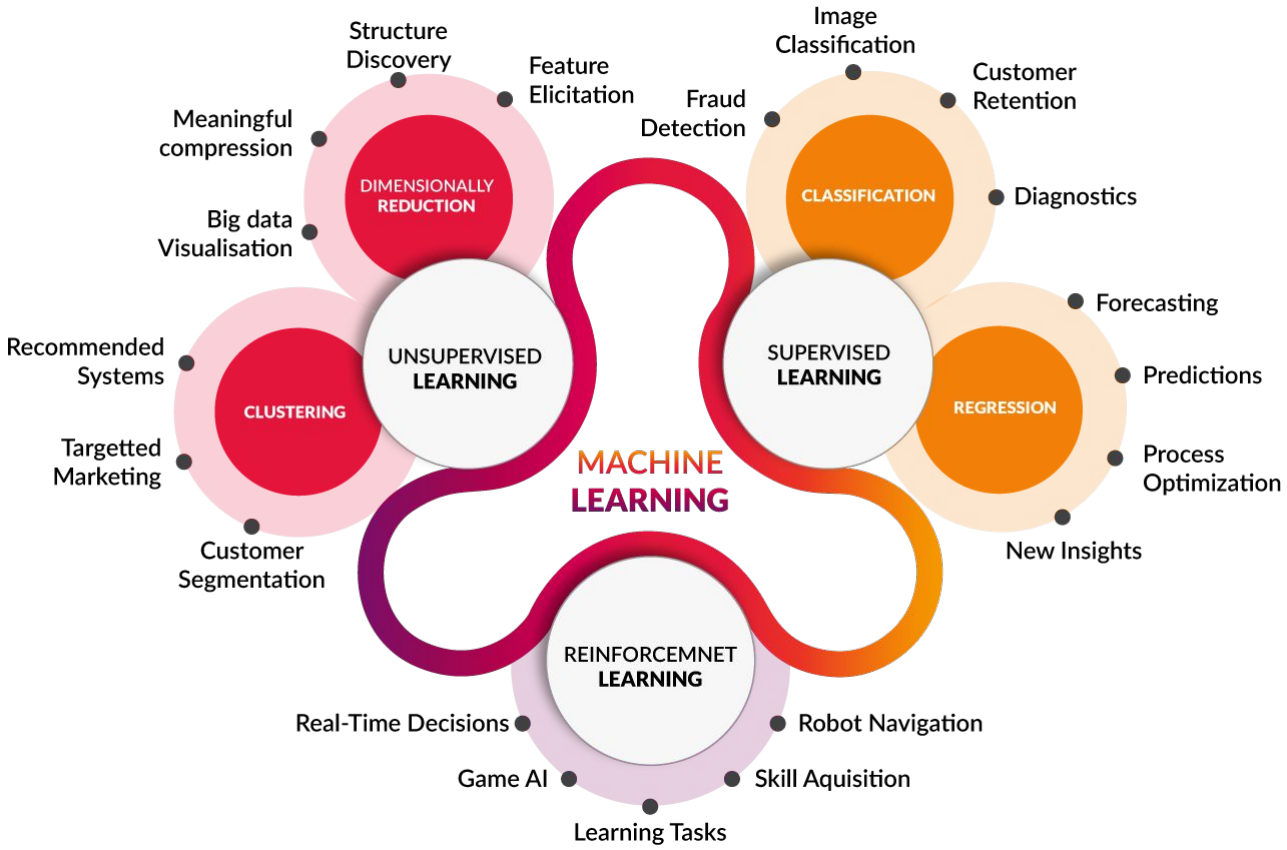
Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2

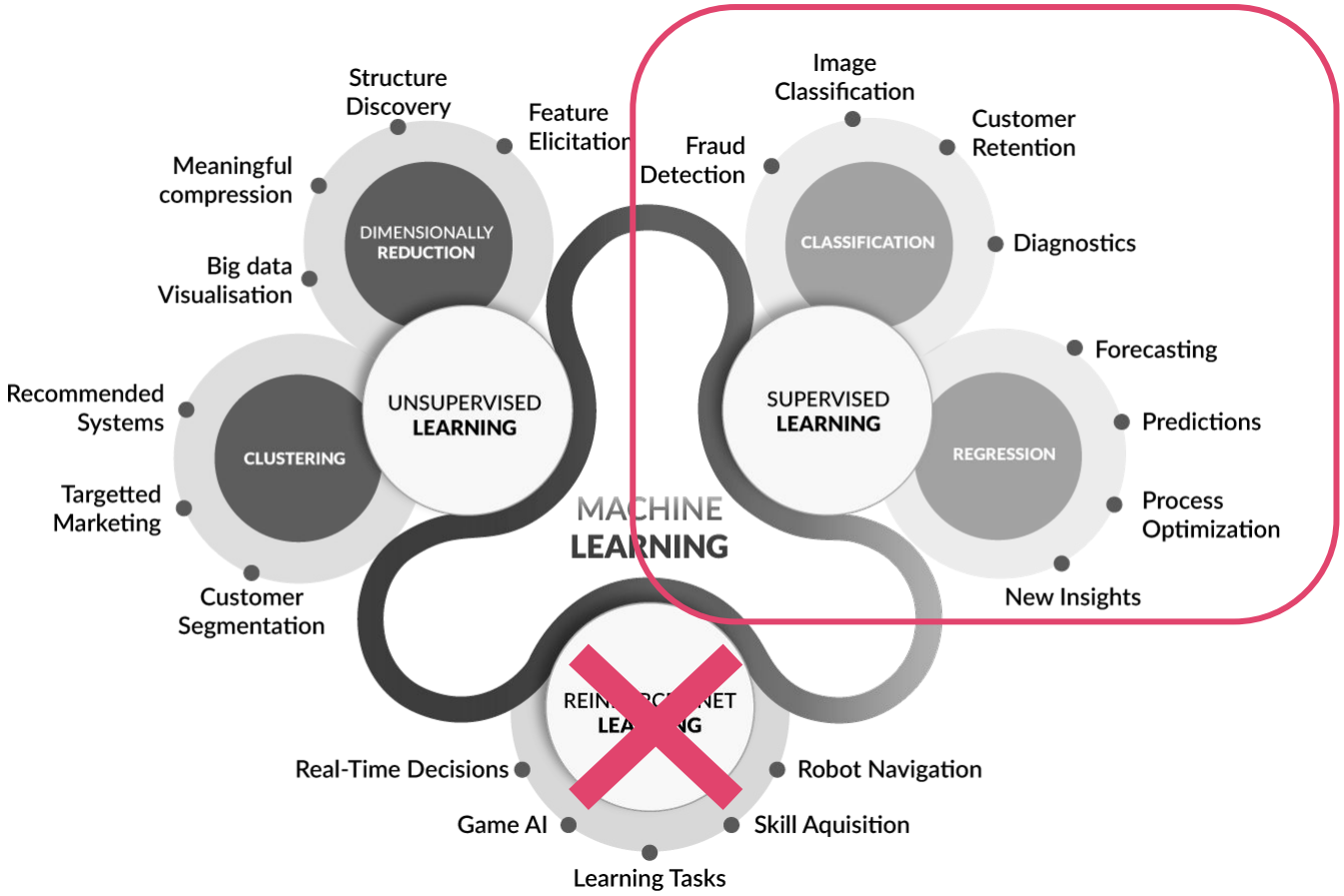
Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.



Mapa


MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REINFORCEMNET LEARNING
DIMENSIONALLY REDUCTION
CLUSTERING
CLASSIFICATION
REGRESSION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
Image Classification
Customer Retention
Fraud Detection
Diagnostics
Recommended Systems
Targetted Marketing
Customer Segmentation
Forecasting
Predictions
Process Optimization
New Insights
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Mapa


Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
DIMENSIONALLY REDUCTION
Recommended Systems
Targetted Marketing
Customer Segmentation
CLUSTERING
UNSUPERVISED LEARNING
Image Classification
Customer Retention
Fraud Detection
Diagnostics
CLASSIFICATION
SUPERVISED LEARNING
Forecasting
Predictions
Process Optimization
New Insights
REGRESSION
MACHINE LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Aprendizaje Supervisado

$$f(X) = Y$$

tenemos datos Y

tenemos datos X

Aprendizaje Supervisado

$$f(X) = Y$$

¿Qué buscamos con “f”?

tenemos datos X

tenemos datos Y

Aprendizaje Supervisado

$$f(X) = Y$$

Un modelo **f** que permita determinar la salida a partir de la entrada

tenemos datos X

tenemos datos Y

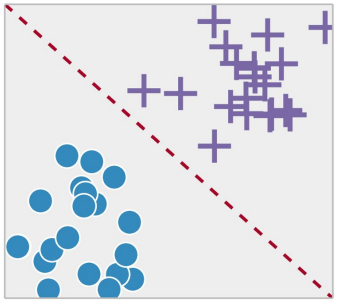
Aprendizaje Supervisado

$$f(X) = Y$$

Con este modelo podremos predecir **Y**, para nuevos datos **X** de los cuales no conozcamos la salida.

Aprendizaje Supervisado

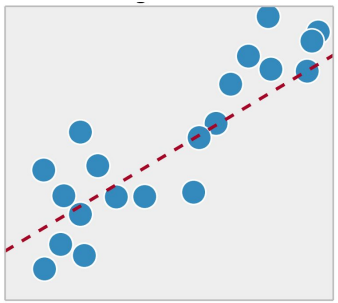
Clasificación



La variable de salida
es una categoría:

- Enfermo / Sano
- Gato / Perro / Pájaro
- **Spam / no Spam**

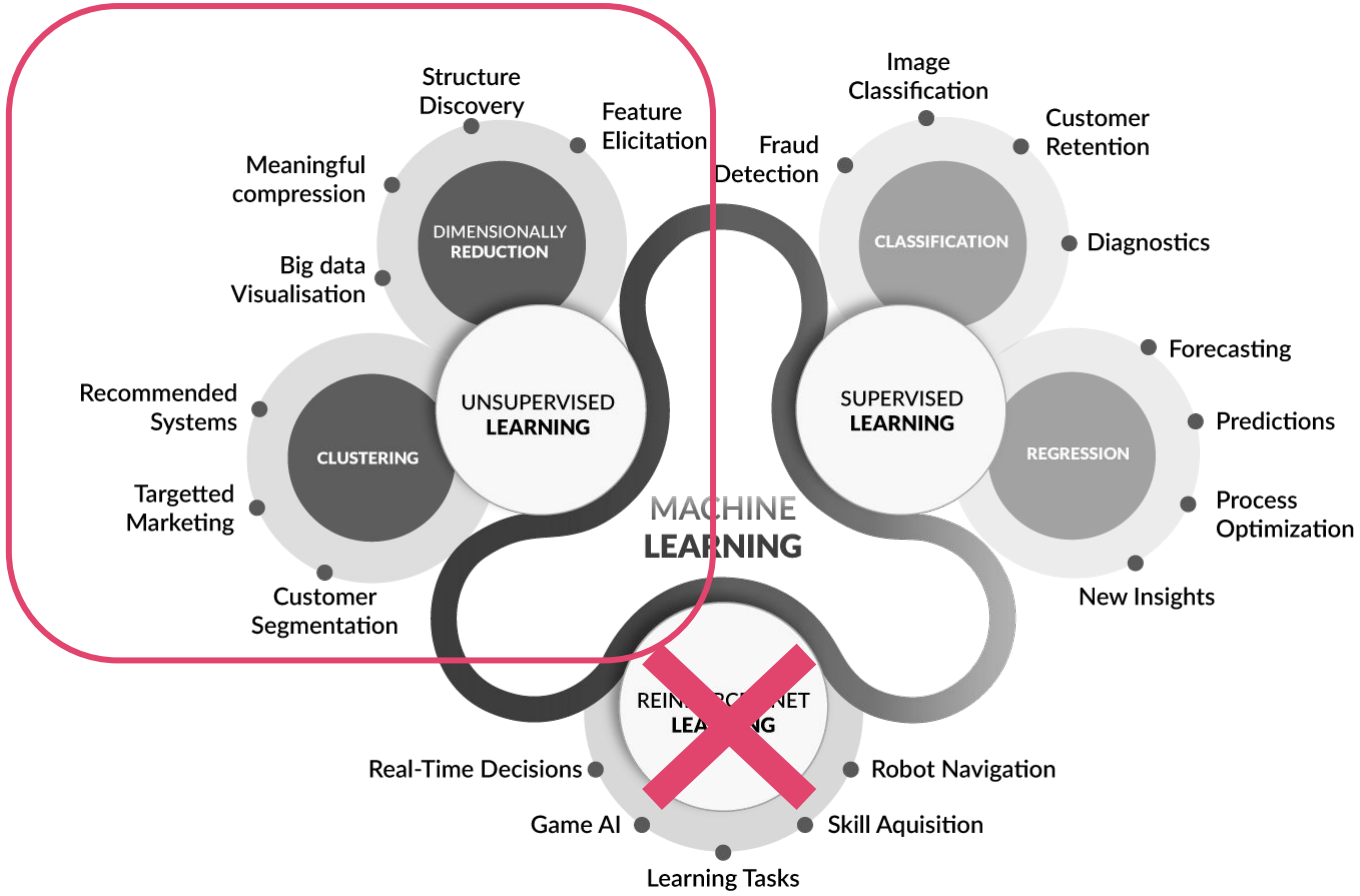
Regresión



La variable de salida
es un valor:

- Precio
- Cantidad

Mapa

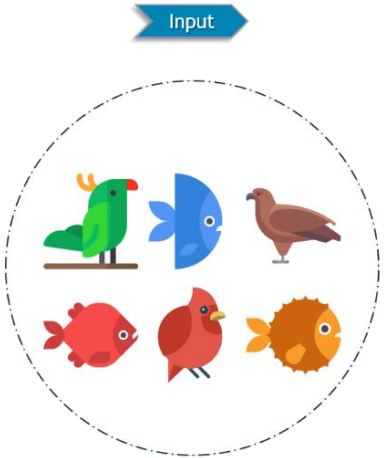

Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
DIMENSIONALLY REDUCTION
UNSUPERVISED LEARNING
Recommended Systems
Targetted Marketing
Customer Segmentation
CLUSTERING
Image Classification
Customer Retention
Fraud Detection
Diagnostics
CLASSIFICATION
SUPERVISED LEARNING
Forecasting
Predictions
Process Optimization
New Insights
REGRESSION
MACHINE LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Aprendizaje NO Supervisado

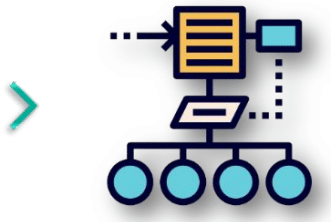
En este tipo de algoritmos solo tenemos los datos de entrada **X**, no hay una salida deseada **Y**.
Lo que se buscan son patrones de similaridad en los datos de entrada.

Aprendizaje NO Supervisado

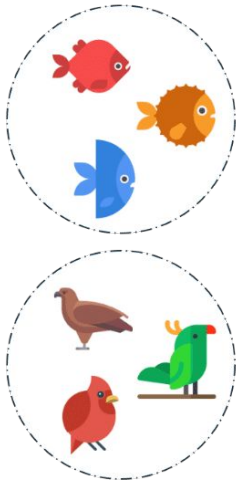
Datos de entrada **X**



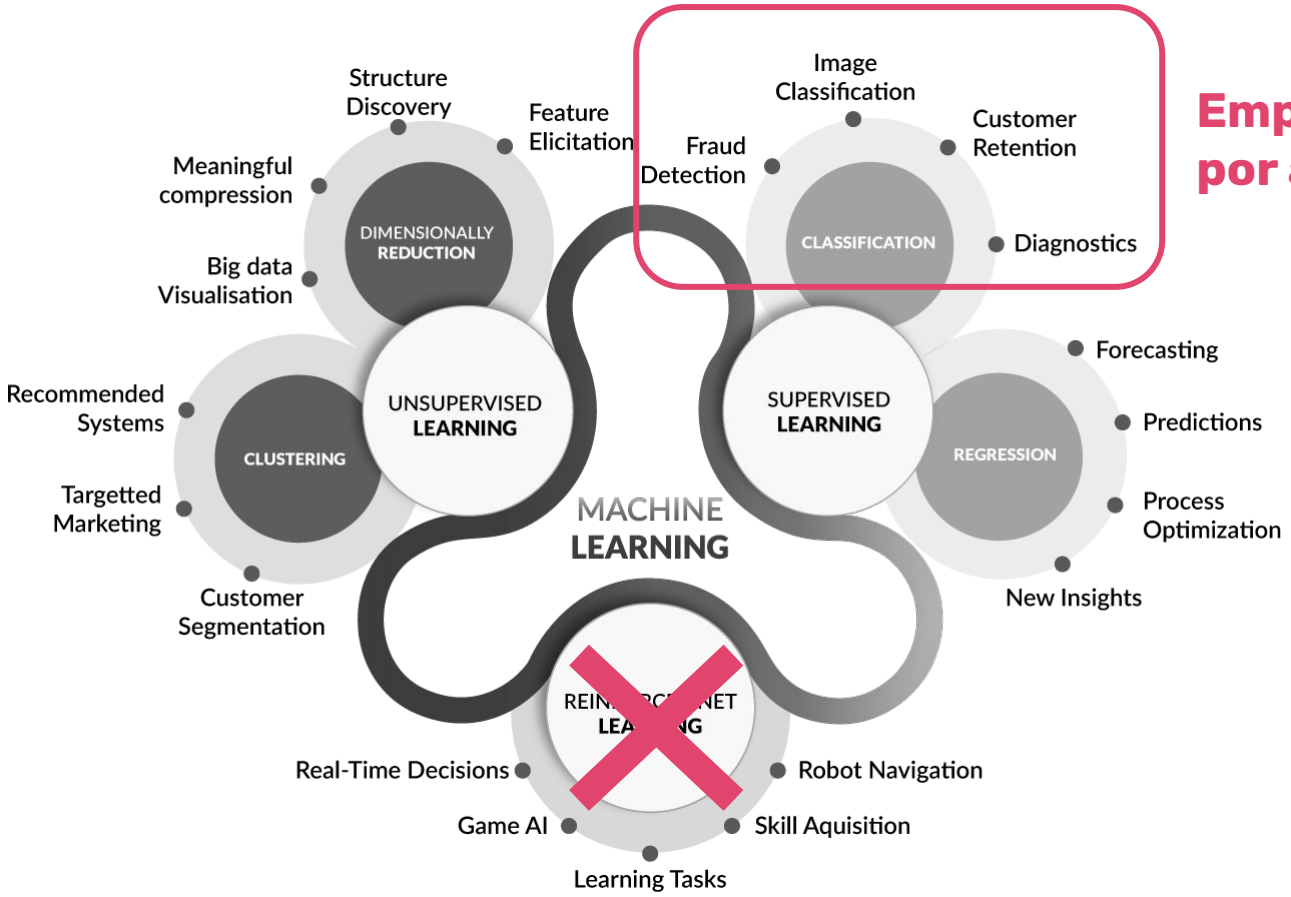
Algoritmo no Supervisado



Patrones encontrados (grupos)



Mapa



Empezamos por acá

¡BREAK!

Ph. Credit: Drew Coffmann

Aprendizaje Supervisado: Árbol de decisión

Machine Learning

↓

Aprendizaje Supervisado

↓

Clasificación

Modelos

- **Árbol de Decisión**
- Support Vector Machines
- k-nearest neighbors
- Random Forest
- Perceptrón
- etc…

Árbol de decisión - DataSet iris

```
pd.concat([pd.DataFrame(data=X, columns=iris.feature_names),
           pd.Series(y, name='target')], axis=1)
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

150 rows × 5 columns

Árbol de decisión

```
pd.concat([pd.DataFrame(data=X, columns=iris.feature_names),
           pd.Series(y, name='target')], axis=1)
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

150 rows × 5 columns

Features **X**

Target **Y**

Árbol de decisión - Train

Entrenamos un Modelo de **DecisionTree** Clasificador sobre el dataset de iris

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_iris
```

```
iris = load_iris()
```

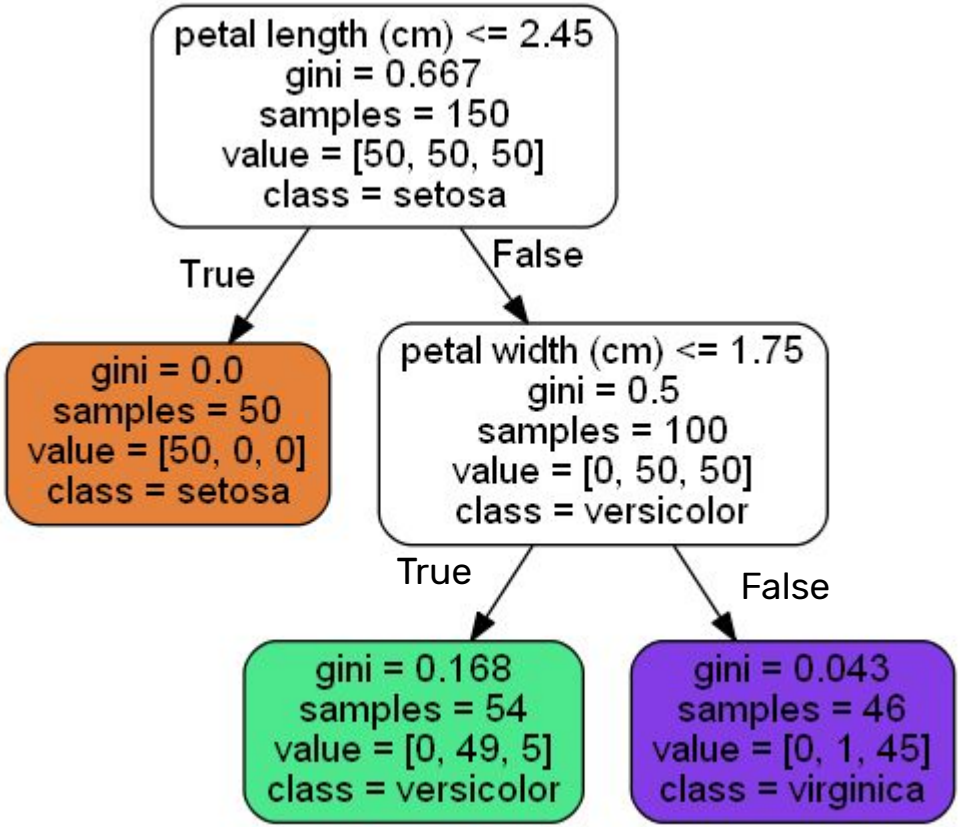
```
X = iris.data
y = iris.target

tree = DecisionTreeClassifier(max_depth=3).fit(X, y)

print(f'Features importance: \n{list(zip(iris.feature_names, tree.feature_importances_))}'.replace('), ', ')\n'))
```

```
Features importance:
[('sepal length (cm)', 0.0)
('sepal width (cm)', 0.0)
('petal length (cm)', 0.5856155514031495)
('petal width (cm)', 0.4143844485968506)]
```

Árbol de decisión - Que paso?

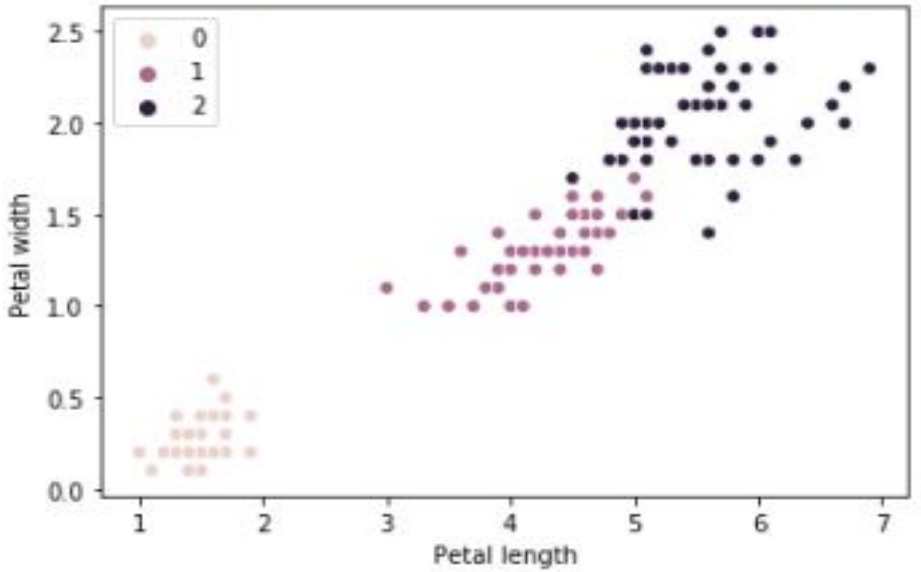


Árbol de decisión - Plot

```
import seaborn as sns
print(list(enumerate(iris.target_names)))

chart = sns.scatterplot(x=X[:,2], y=X[:, 3], hue=y);
chart.set_xlabel('Petal length')
chart.set_ylabel('Petal width');
```

```
[(0, 'setosa'), (1, 'versicolor'), (2, 'virginica')]
```

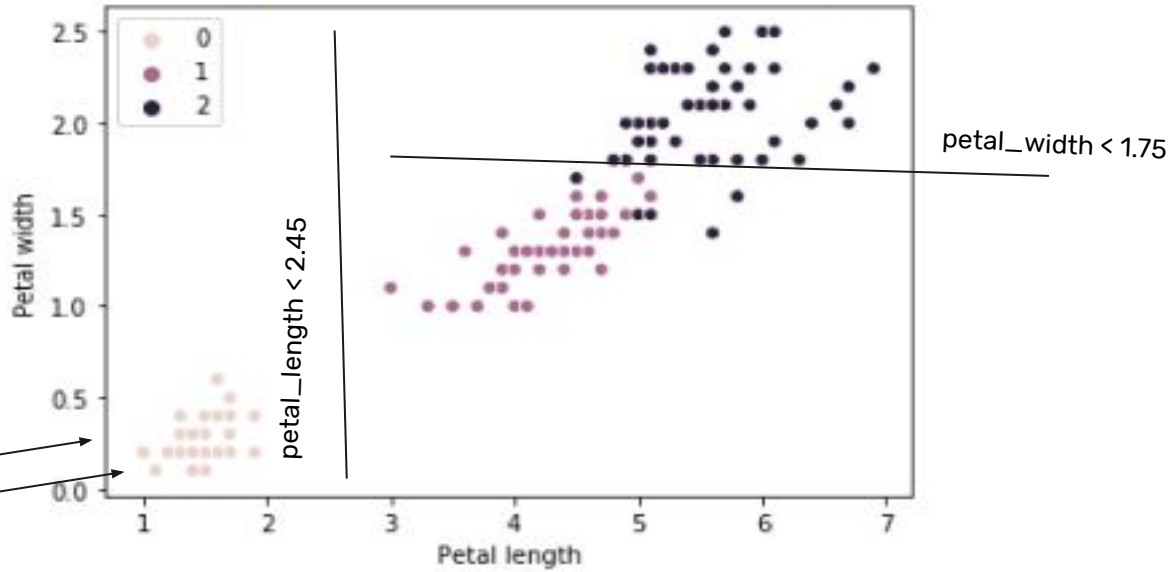


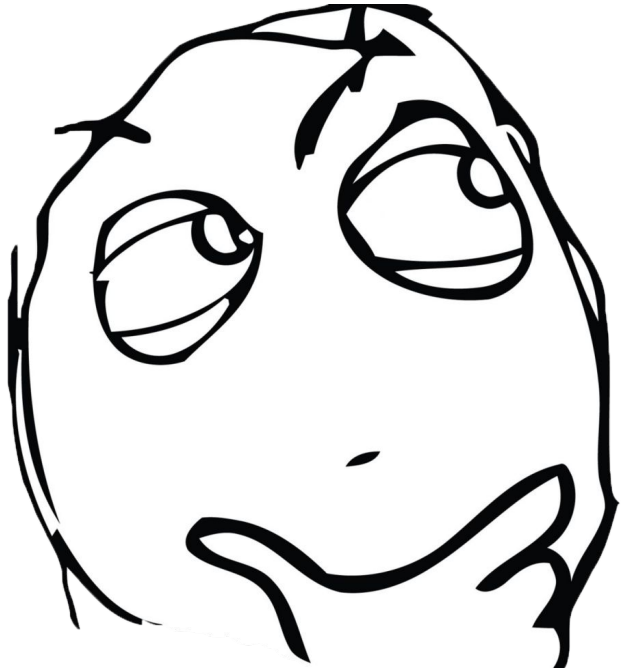
Árbol de decisión - Plot

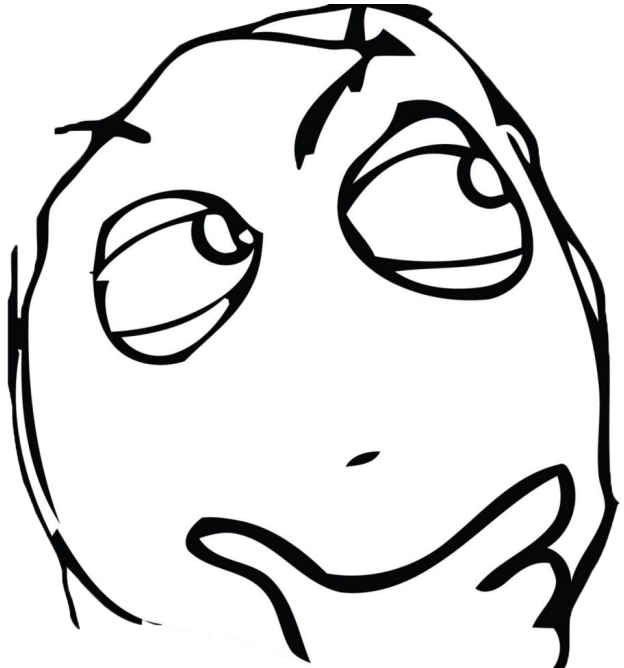
```
import seaborn as sns
print(list(enumerate(iris.target_names)))

chart = sns.scatterplot(x=X[:,2], y=X[:, 3], hue=y);
chart.set_xlabel('Petal length')
chart.set_ylabel('Petal width');
```

```
[(0, 'setosa'), (1, 'versicolor'), (2, 'virginica')]
```



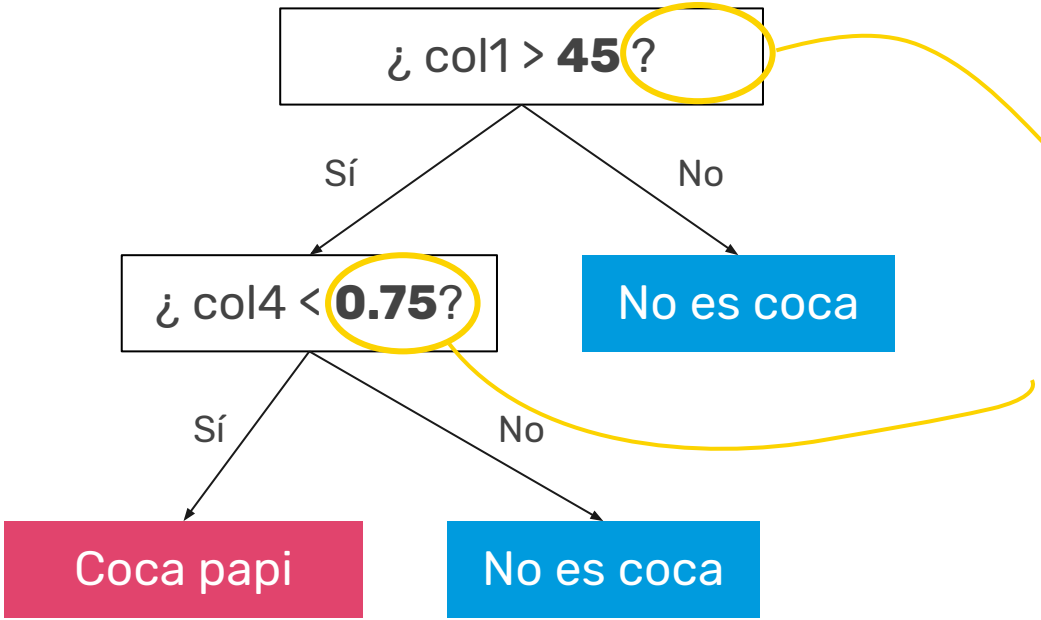




Es una serie de **IF**s...

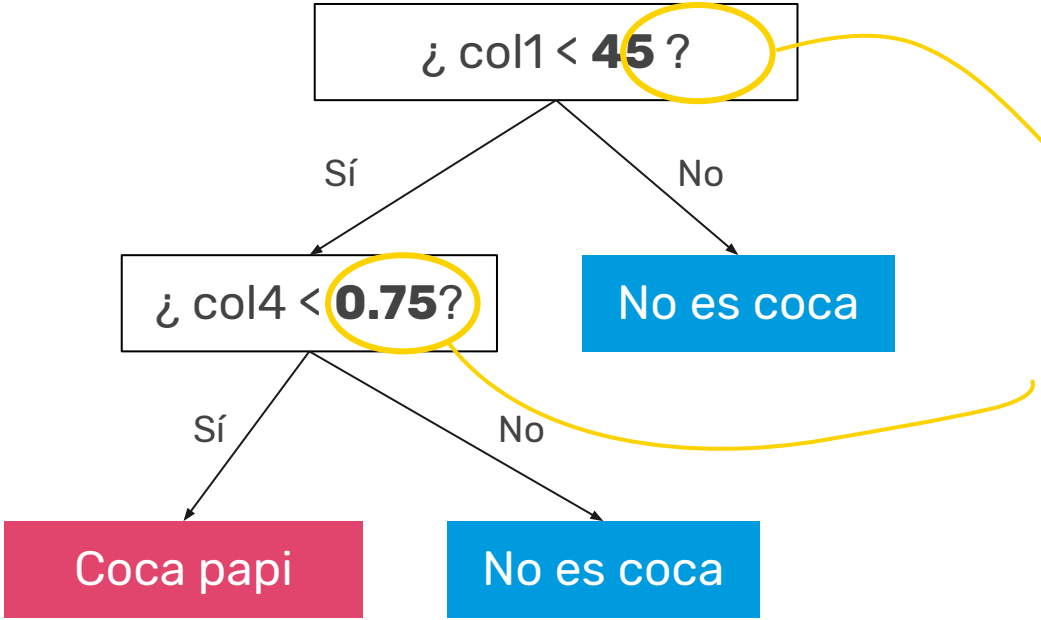
¿Por qué ahora decimos que es Machine Learning?

¿Por qué decimos que es Machine Learning?



Es **Machine Learning** porque estos valores se eligen automáticamente al entrenar el modelo, a partir de los datos **X** e **Y**.

¿Por qué decimos que es Machine Learning?

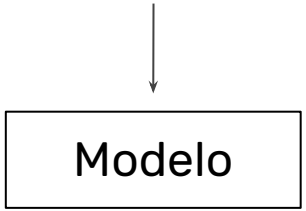


La próxima clase vamos a ver **cómo** es que se eligen estos valores a partir de un proceso matemático.

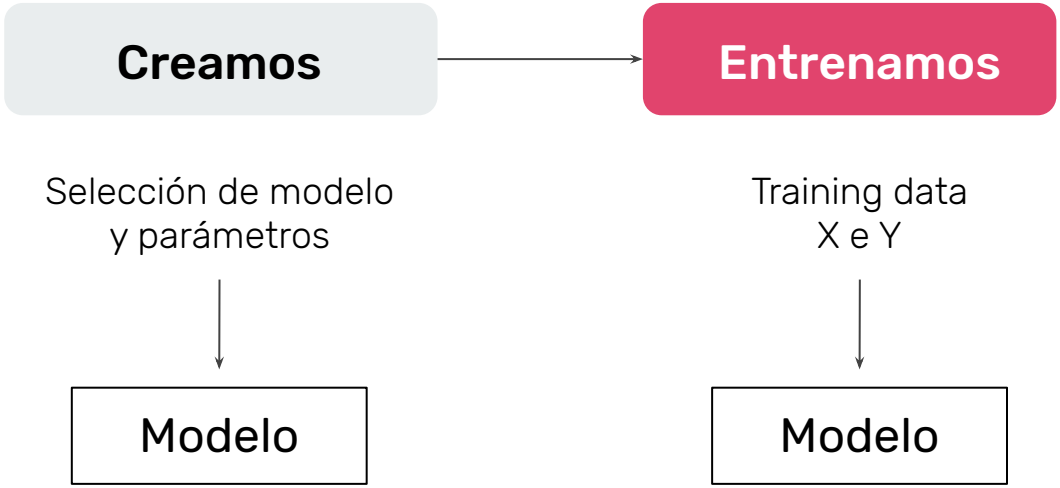
Flujo de trabajo Scikit Learn

Creamos

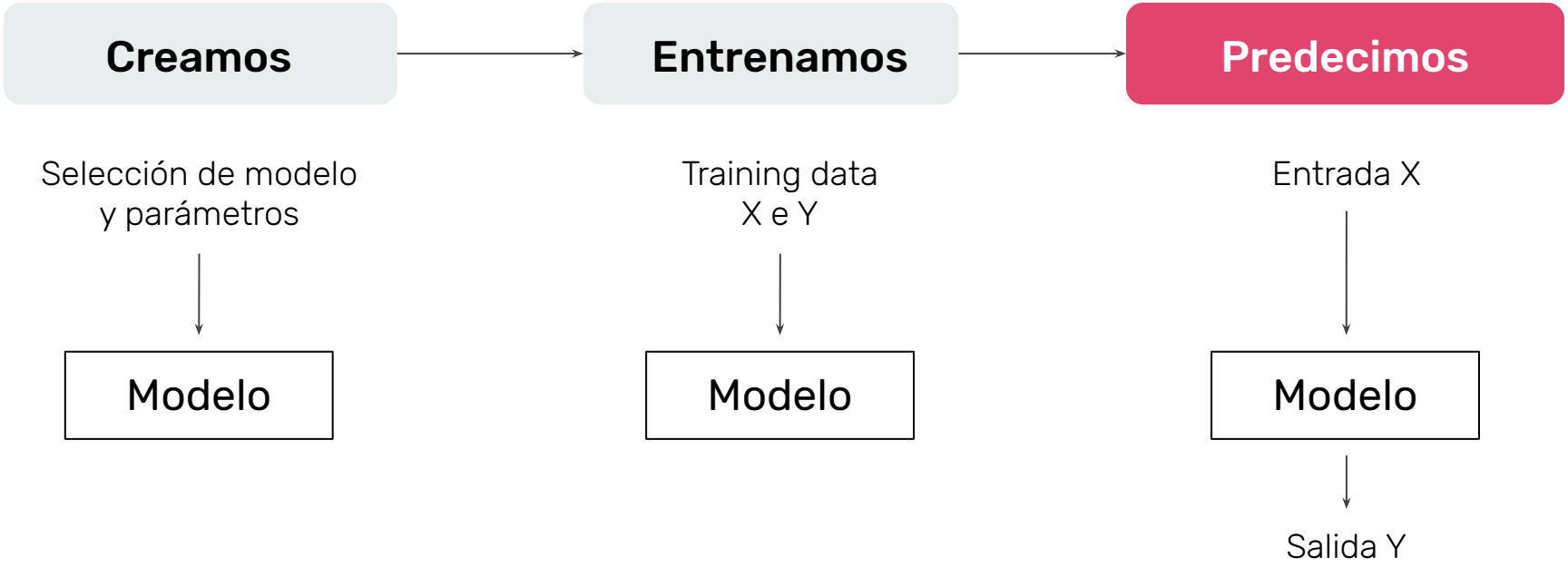
Selección de modelo
y parámetros



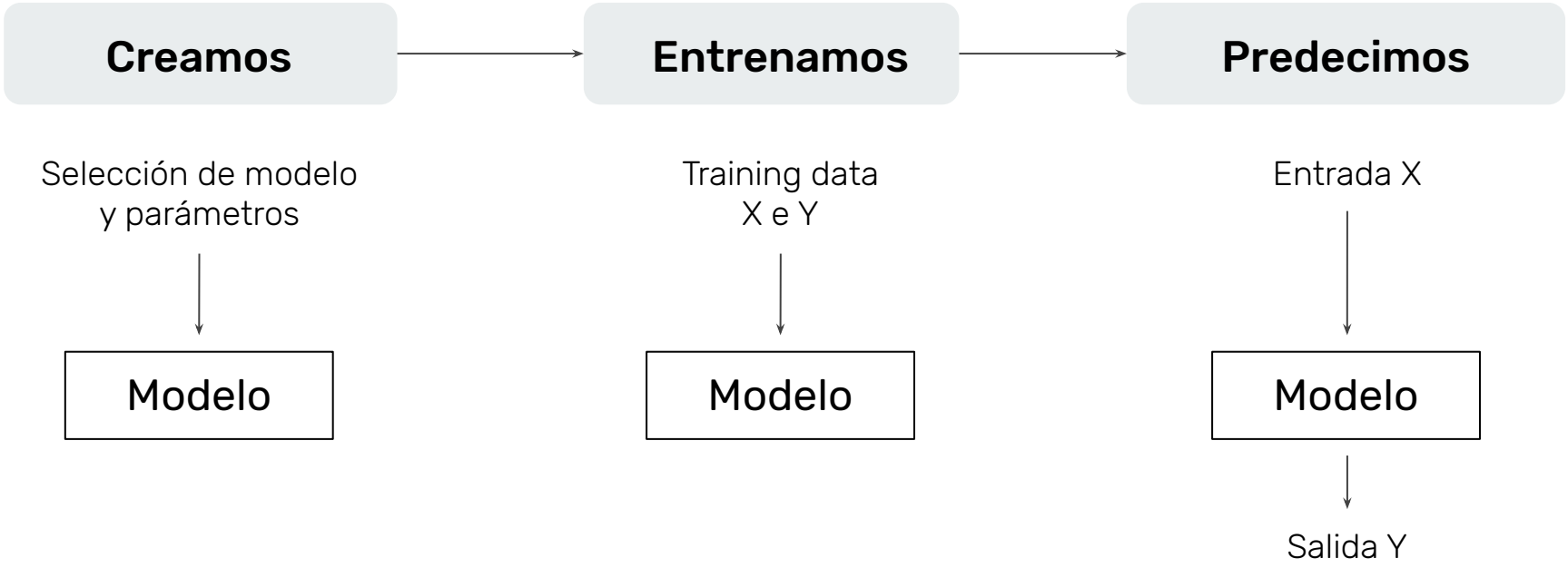
Flujo de trabajo Scikit Learn



Flujo de trabajo Scikit Learn



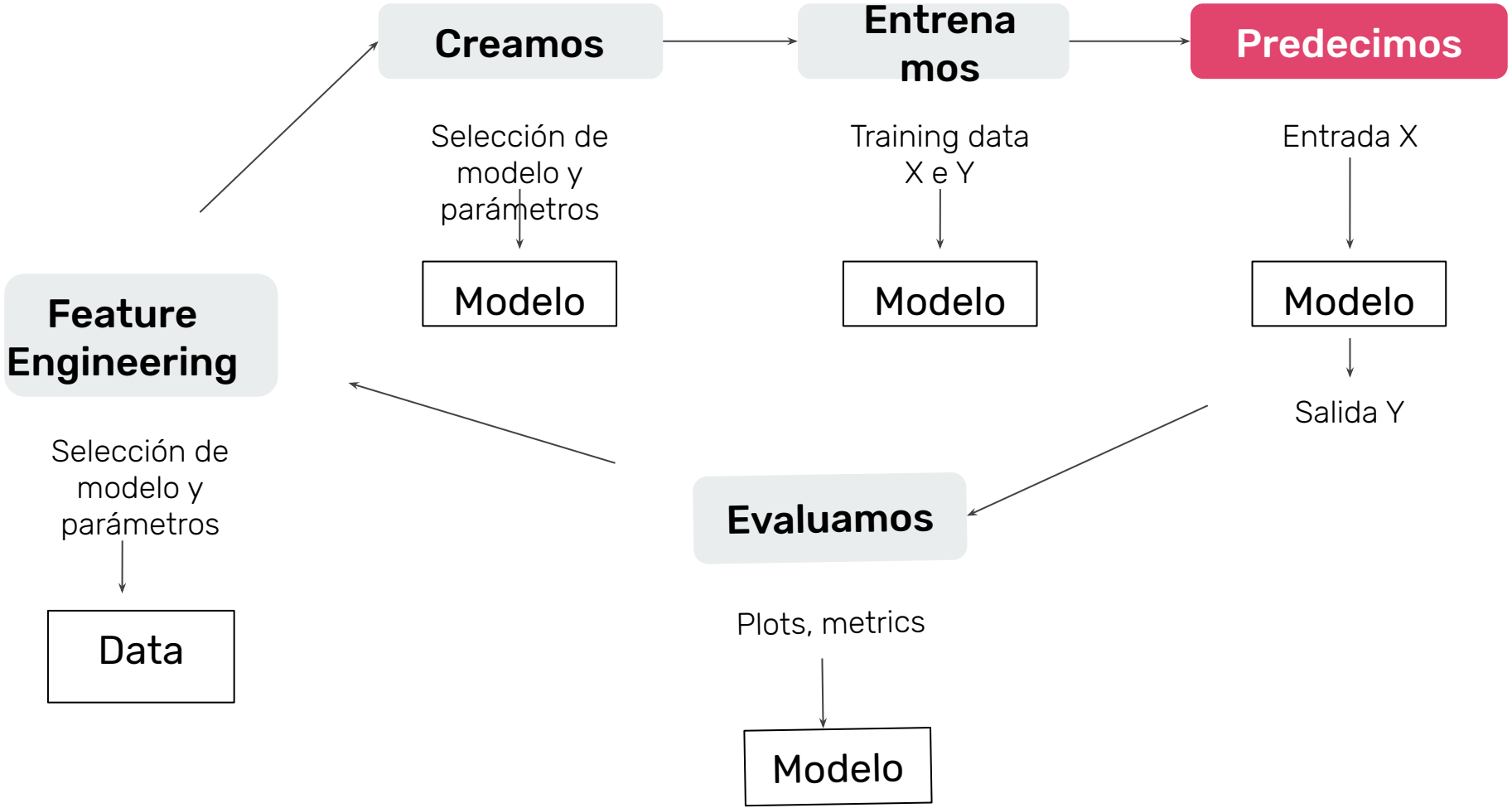
Flujo de trabajo Scikit Learn




```
from sklearn import tree
clf = tree.DecisionTreeClassifier()
```

```
clf = clf.fit(X,Y)
```

```
Y_pred = clf.predict(X)
```


Creamos
Entrenamos
Predecimos
Selección de modelo y parámetros
Training data X e Y
Entrada X
Modelo
Modelo
Modelo
Feature Engineering
Salida Y
Selección de modelo y parámetros
Evaluamos
Plots, metrics
Data
Modelo

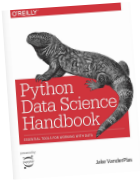
Hands-on training

Hands-on training

DS_Clase_15_ML.ipynb

Recursos

Recursos



1. **Scikit-Learn Decision Trees Explained**: Buena (y completa) introducción a Árboles de decisión con Scikit-Learn.

2. **Capítulo 5, "Machine Learning", de Python Data Science Handbook.** Acá van a encontrar una introducción general a ML.

3. **Capítulo 5.08, "In-Depth: Decision Trees and Random Forests", de Python Data Science Handbook.** Acá van a encontrar árboles de decisión explicado con código funcional para copiar y pegar en sus proyectos.

4. **Video muy interesante** sobre el impacto de Machine Learning (Automatización moderna) en la sociedad.

Para la próxima

1. Ver los videos de la plataforma “Machine Learning: Árboles de Decisión” y “Validación y testeo de modelos: Validación y testeo” (¡nos salteamos algunos videos!)

2. Completar el notebook de hoy

3. ¡Terminar la entrega 02 si aún no lo hicieron!

ACÁMICA