




# Inteligencia Artificial con Python y scikit-learn


[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#)

## scikit-learn

Machine Learning in Python

[Getting Started](#) [Release Highlights for 1.6](#)

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

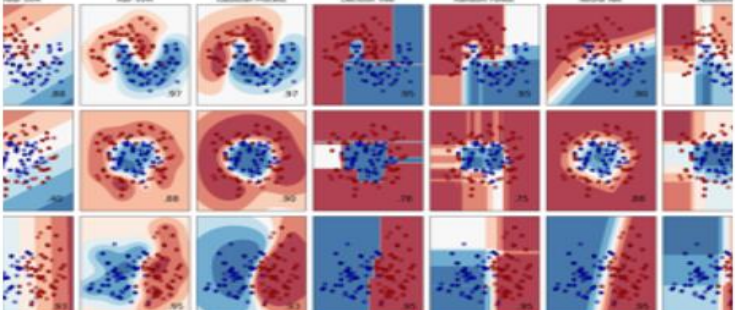


### Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [logistic regression](#), and [more...](#)

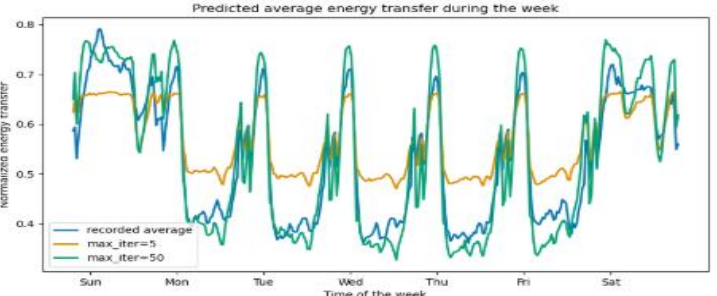


### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, stock prices.

**Algorithms:** [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [ridge](#), and [more...](#)




### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, grouping experiment outcomes.

**Algorithms:** [k-Means](#), [HDBSCAN](#), [hierarchical clustering](#), and [more...](#)



# Inteligencia Artificial con Python y scikit-learn

## Temario



### 1. Introducción a la IA con Python

- Entorno de trabajo, Instalación y configuración de librerías

### 2. Preprocesamiento de Datos

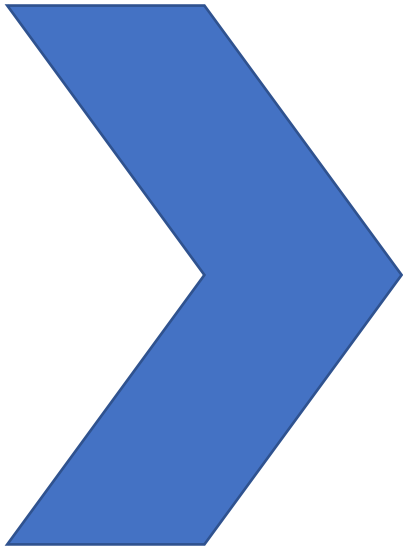
- Carga y exploración de datos
- Limpieza de datos
- Codificación de variables categóricas
- Normalización y escalado

### 3. Modelos de Clasificación

- Problemas de clasificación y ejemplos reales.
- **Regresión Logística**
- **K-Nearest Neighbors (KNN)**
- **Árboles de Decisión y Random Forest**
- Visualización con gráficos 2D y 3D.
- Evaluación y selección de modelos

# Inteligencia Artificial con Python y scikit-learn

## Temario



### 4. Regresión

- Implementación de Modelos de Regresión
- Regresión Lineal.
- Regresión Lineal Múltiple.

### 5. Evaluación de Modelos de Regresión

- Métricas: MSE, RMSE,  $R^2$ .
- Validación cruzada para regresión.

### 6. Optimización de Modelos

### 7. Reducción de Dimensionalidad

- Importancia de la reducción de dimensiones en IA.
- **PCA (Análisis de Componentes Principales):** teoría y aplicación.
- Interpretación de los resultados y visualización.

# Google Colab for Python

- Google Colab es un entorno gratuito basado en la nube de Jupyter Notebook que utilizan desarrolladores de Python, científicos de datos y desarrolladores del aprendizaje automático, inteligencia artificial y redes neuronales en todo el mundo.
- Permite escribir y ejecutar código Python directamente en el navegador.
- Es una herramienta popular debido a su facilidad de uso y al acceso a potentes recursos informáticos.





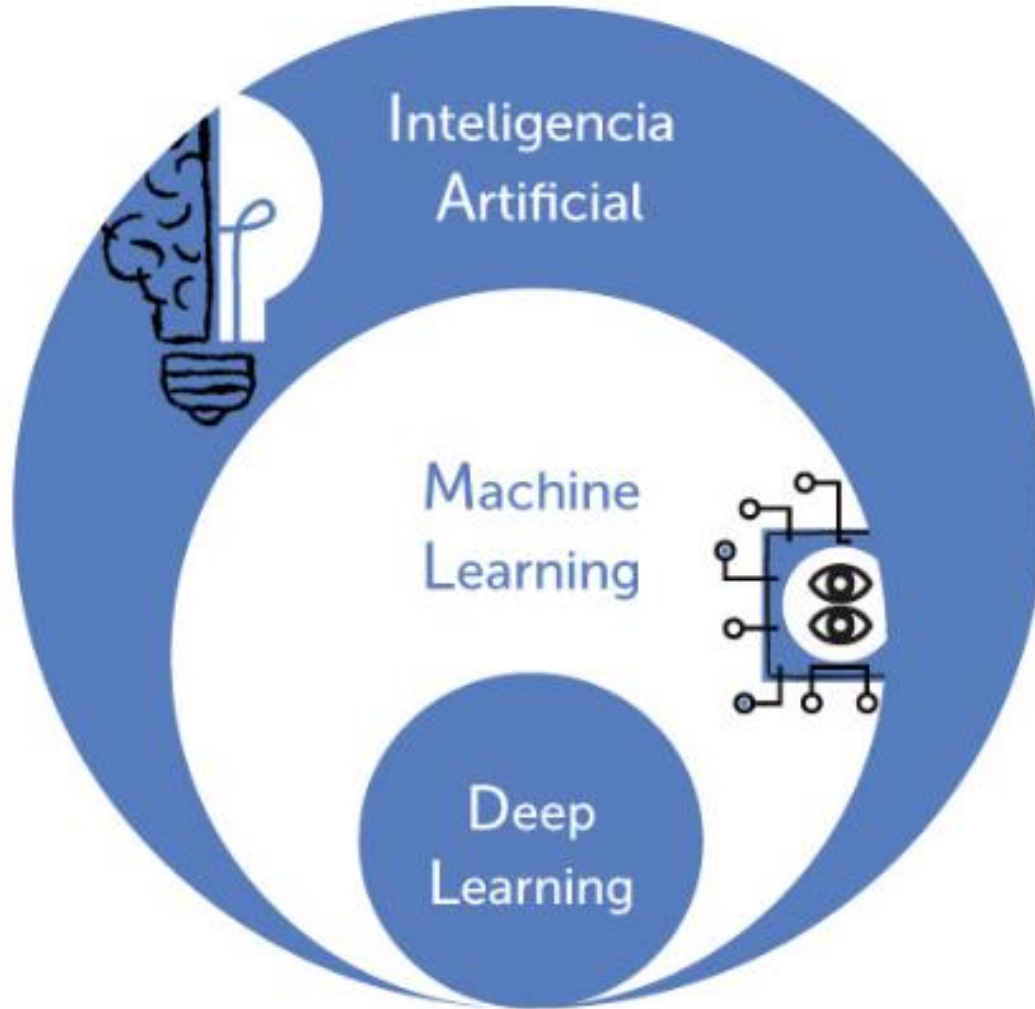
# Google Colab for Python

- **No requiere configuración** : no es necesario instalar Python ni ninguna biblioteca adicional en su equipo.
- **Acceso gratuito a GPU/TPU** : Google Colab ofrece acceso gratuito a hardware potente como GPU y TPU, lo que acelera significativamente tareas como el aprendizaje automático y el entrenamiento de aprendizaje profundo.
- **Colaboración** : Colab permite la colaboración en tiempo real, de forma similar a Google Docs. Varios usuarios pueden trabajar en el mismo bloc de notas y ver los cambios de los demás en tiempo real.
- **Integración de almacenamiento en la nube** : Colab se integra con Google Drive, lo que le permite guardar sus cuadernos y archivos en la nube y hacerlos accesibles desde cualquier lugar.
- **Admite bibliotecas populares** : las bibliotecas de Python preinstaladas como TensorFlow, Keras, PyTorch, OpenCV, NumPy, Pandas y muchas más están disponibles en Colab, lo que lo convierte en una plataforma ideal para proyectos de aprendizaje automático y ciencia de datos.
- **Compartir y publicar fácilmente** : puedes compartir fácilmente tus cuadernos generando un enlace para compartir. El código y los resultados se pueden insertar en sitios web o blogs directamente desde Colab.
- **Uso gratuito** : Colab es gratuito, con planes pagos opcionales para usuarios que necesitan más potencia computacional o tiempos de ejecución más prolongados.

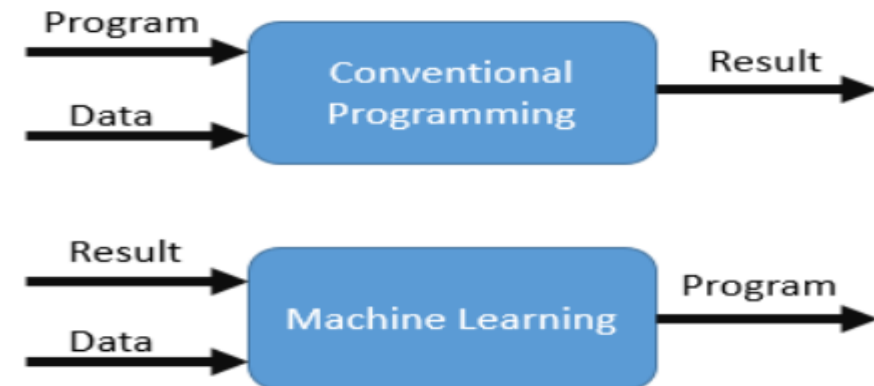
# Librería scikit-learn: Aprendizaje automático con Python

- **scikit-learn Homepage**  
<http://scikit-learn.org/>
- **scikit-learn User Guide**  
[http://scikit-learn.org/stable/user\\_guide.html](http://scikit-learn.org/stable/user_guide.html)
- **scikit-learn API reference**  
<http://scikit-learn.org/stable/modules/classes.html>





- ML es un subconjunto de AI. Se utiliza en escenarios en los que necesita que las máquinas **aprendan de grandes volúmenes de datos.**
- El conocimiento adquirido se aplica a un nuevo conjunto de datos.
- ML le da a una máquina la capacidad de aprender de (o acerca de) conjuntos de datos nuevos **sin dar instrucciones explícitas.**



- **El software tradicional suele estar compuesto por reglas lógicas sencillas y codificadas a mano.**
- Por ejemplo, SI la condición X es THEN, realice la acción Y.
- **El aprendizaje automático se basa en modelos estadísticos complejos para descubrir patrones** en grandes conjuntos de datos.
- Tomemos como ejemplo la aprobación de un préstamo.
  - Dados años de historial crediticio y otra información secundaria, un algoritmo de aprendizaje automático podría generar una probabilidad de que el solicitante incumpla.
  - La lógica detrás de esta evaluación no se codificaría a mano.
  - En cambio, el modelo se extrapolaría a partir de los registros de miles o millones de otros clientes.





## Aprendizaje automático

- Kevin Patrick Murphy define el aprendizaje automático como “...un **conjunto de métodos que pueden detectar automáticamente patrones en los datos, y luego utilizar los patrones detectados para predecir los datos futuros, o realizar otros tipos de toma de decisiones bajo incertidumbre**”.



**Aprendizaje automático supervisado:** Aprende a predecir los valores objetivo a partir de datos etiquetados.

- **Clasificación** (los valores objetivo son clases discretas)
- **Regresión** (los valores objetivo son valores continuos)

**Aprendizaje automático no supervisado:** encuentra la estructura en datos sin etiquetar

- Busca grupos de instancias similares en los datos (agrupación en clústeres)
- Búsqueda de patrones inusuales (detección de valores atípicos)



## Aprendizaje automático

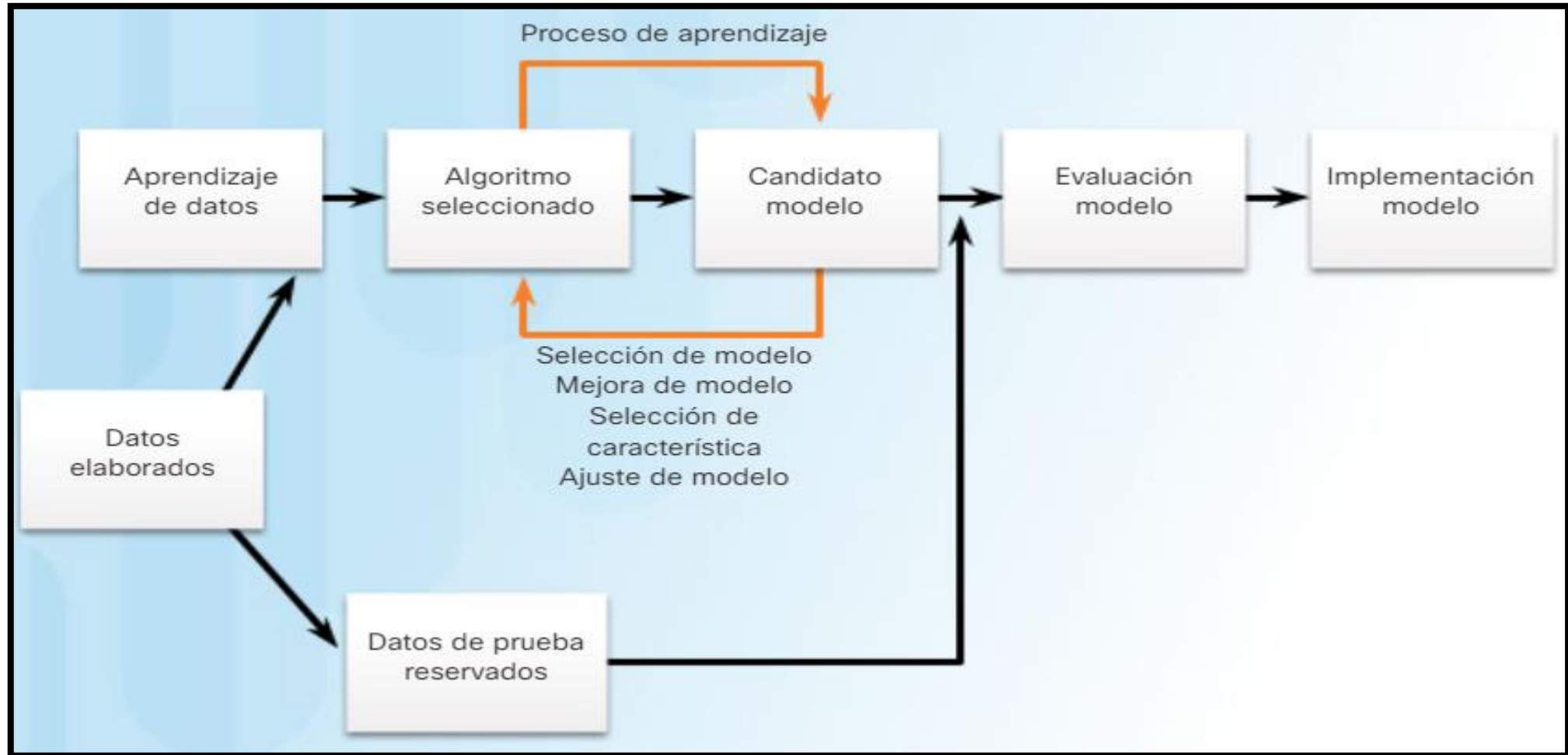
- Los **algoritmos de aprendizaje automáticos supervisados** son los algoritmos de aprendizaje automático más utilizados para el **análisis predictivo**.
- Estos **algoritmos dependen de conjuntos de datos que fueron procesados por los expertos humanos** (por lo tanto, se usa la palabra “supervisión”).
- Los algoritmos luego aprenden cómo realizar las mismas tareas de procesamiento de forma independiente en los nuevos conjuntos de datos.
- En particular, los métodos supervisados se utilizan para resolver problemas de **regresión y clasificación**:

## Proceso de Aprendizaje automático

- **Paso 1:** Este es el paso de **preparación de los datos**. En este paso, incluimos los procedimientos de limpieza de datos (es decir, la transformación a un formato estructurado, la eliminación de datos faltantes y las observaciones de ruido/error).
- **Paso 2:** Cree un **conjunto de aprendizaje** que se use realmente para entrenar el modelo.
- **Paso 3:** Cree un **conjunto de prueba** que se use para evaluar el rendimiento del modelo. El paso de prueba se realiza solo en caso de aprendizaje supervisado.
- **Paso 4:** Cree un bucle. **Se elige un algoritmo**, según el problema necesario, y sus rendimientos se evalúan en los datos de aprendizaje. Según el algoritmo elegido, podrían ser necesarios **pasos de preprocesamiento**, como la extracción de **características** del conjunto de datos que sean relevantes para el problema.
- **Paso 5:** La prueba de la solución en datos de prueba se denomina paso de **evaluación del modelo**.
- **Paso 6:** Cuando el modelo logra rendimientos satisfactorios en datos de prueba, el modelo puede **implementarse**.



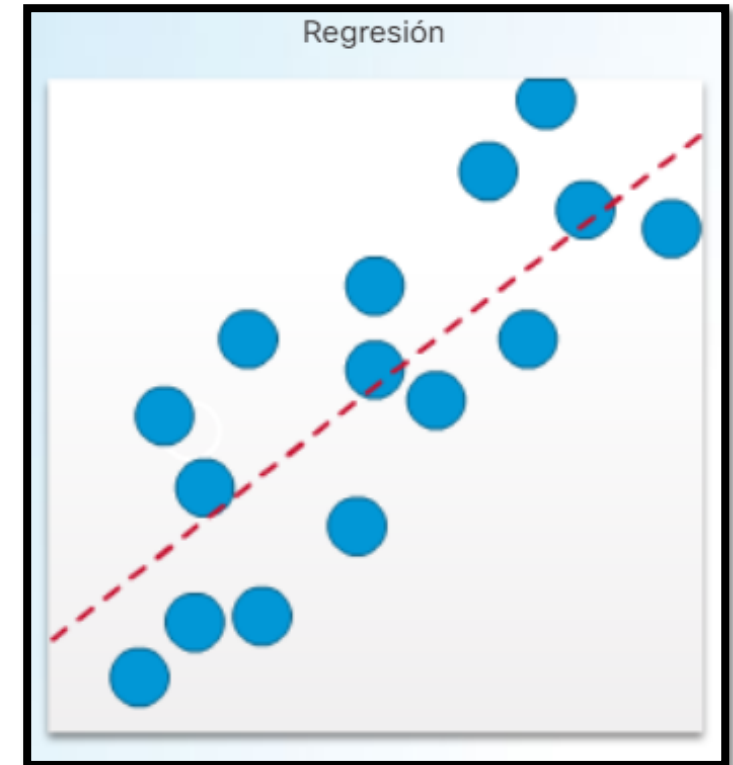
## Proceso de Aprendizaje automático



## Aprendizaje automático

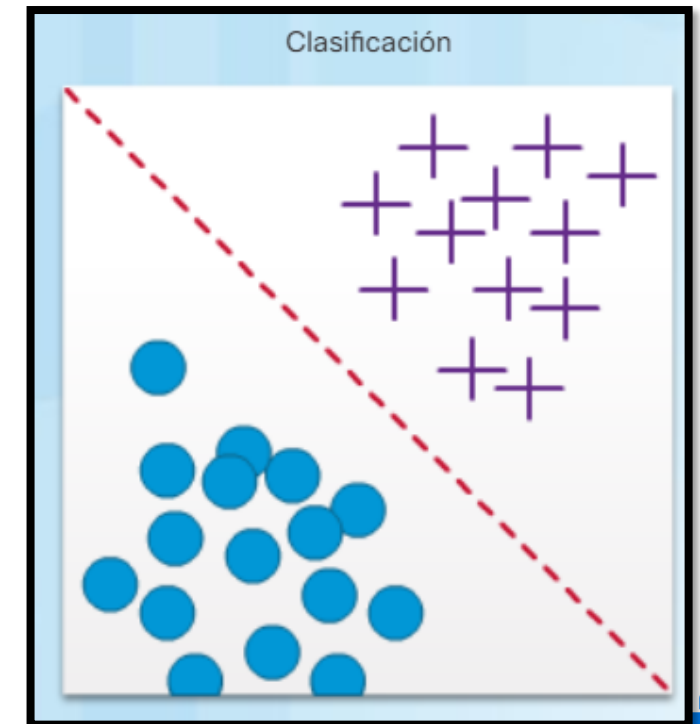
En particular, los métodos supervisados se utilizan para resolver problemas de:

- Regresión y
- Clasificación
- **Problemas de regresión:** son el cálculo de las relaciones matemáticas entre una variable **continua** y una o más variables. Esta relación matemática luego puede utilizarse para calcular los valores de una variable desconocida dados los valores conocidos de las demás.
- La **regresión significa encontrar la línea que interpola mejor los valores**. La línea puede tomar varias formas y se expresa como **función de regresión**.
- Una función de regresión le permite estimar el valor de una variable dado el valor de la otra, para los valores que no se han obtenido antes



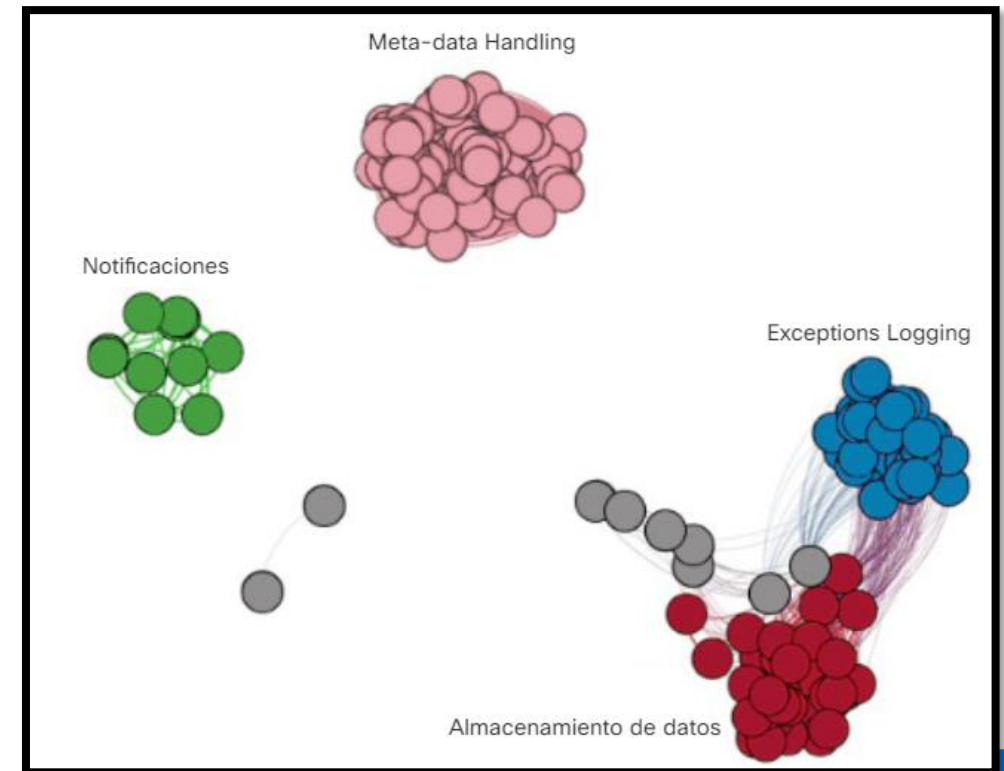
## Aprendizaje automático

- **Problemas de clasificación:** se utilizan cuando la variable desconocida es **discreta**. Por lo general, el problema comprende el cálculo al cual, de un conjunto de clases predefinidas, pertenece un ejemplo específico.
- Los ejemplos típicos de clasificación son reconocimiento de la imagen, o diagnóstico de las patologías de exámenes médicos, o identificación de rostros en una imagen.
- Un problema de clasificación se puede considerar en dos dimensiones, donde los puntos que pertenecen a diferentes clases se marcan. El algoritmo “aprende” ejemplos de la ubicación y la forma de la línea fronteriza entre las clases. Esta línea fronteriza luego puede utilizarse para clasificar nuevos ejemplos.



## Aprendizaje automático

- Los algoritmos de **aprendizaje automático no supervisados** no requieren expertos humanos de los que aprender, sino que descubren patrones en los datos de forma autónoma. Algunos ejemplos de problemas resueltos con métodos no supervisados son el **agrupamiento** y la **asociación**:
- **Métodos de agrupamiento**: estos se pueden ver como la **detección automática de grupos** de ejemplos que tienen **características similares**, que pueden indicar posiblemente el hecho de que un miembro del grupo **pertenece a una clase bien definida**.
- Por ejemplo, los algoritmos de agrupamiento se utilizan para identificar grupos de usuarios basados en su historial de compras en línea, y luego envían avisos dirigidos a cada miembro.
- En la Figura, el algoritmo de agrupamiento ha asignado automáticamente un color diferente al grupo de observaciones que son “estrechas” entre sí.







# EDUCACIÓN

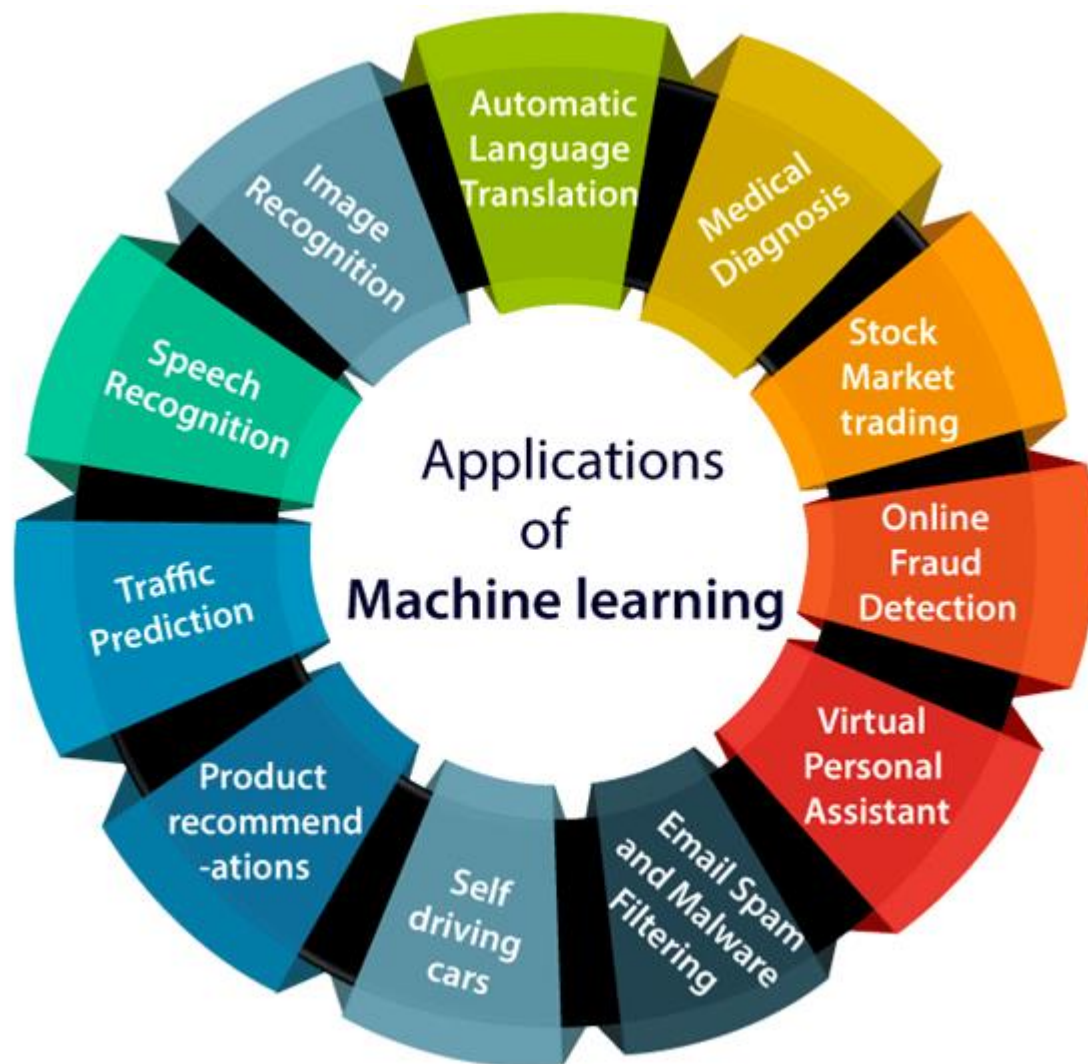
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO  
NACIONAL DE MÉXICO

## TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE DURANGO



## Aprendizaje automático

- Un programa informático es diseñado por un servicio de video por demanda para recomendar películas que podrían gustarles a los usuarios individuales.
- El algoritmo analiza las películas que los espectadores han visto ya y las películas que las personas con preferencias similares de visualización calificaron con buena puntuación.
- El objetivo es mejorar la satisfacción del cliente con el servicio de video



## Aprendizaje automático

### Motor de recomendación de videos

Los algoritmos de clasificación pueden permitir que los usuarios encuentren más fácilmente videos que les agraden. Según el comportamiento en alquiler de videos del cliente y el comportamiento de otros clientes, el algoritmo predice los videos que probablemente disfrute un cliente y hace recomendaciones.





# Aprendizaje automático

## Predicción de brotes de enfermedades en plantas

Los granjeros usan teléfonos móviles para enviar imágenes sobre enfermedades de plantas a los investigadores. Estas imágenes se utilizan en el sistema de reconocimiento de imagen para diagnosticar enfermedades de plantas. Combinadas con algoritmos de regresión de datos ambientales pueden predecir brotes futuros de enfermedades.





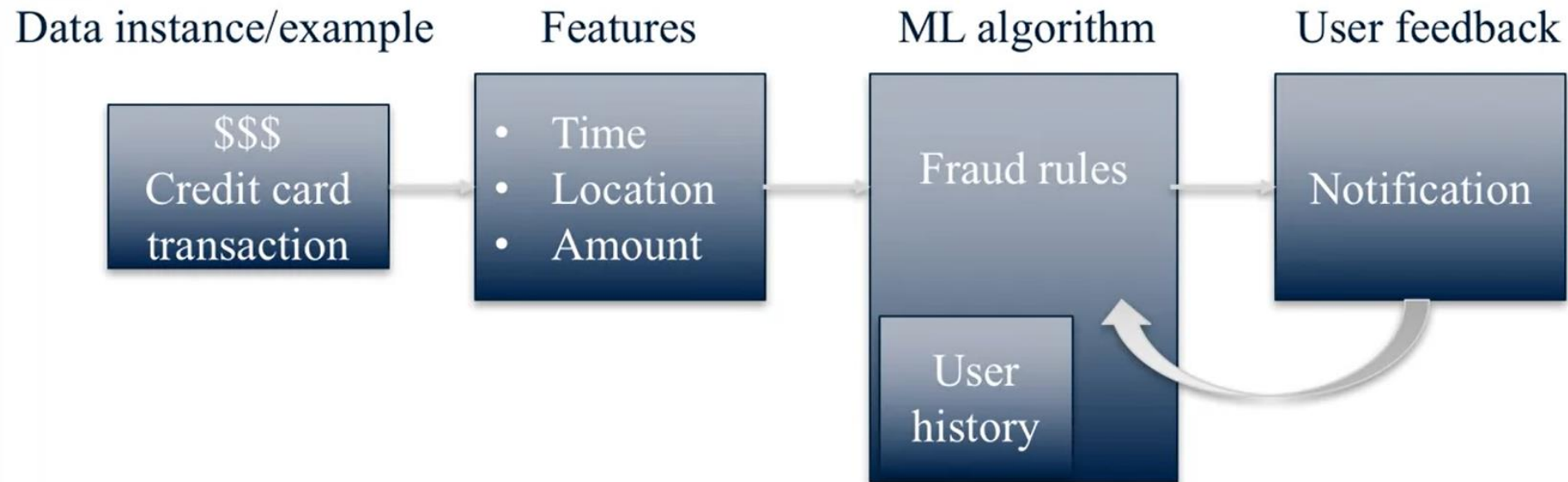
## Aprendizaje automático

**Predicción de cáncer**

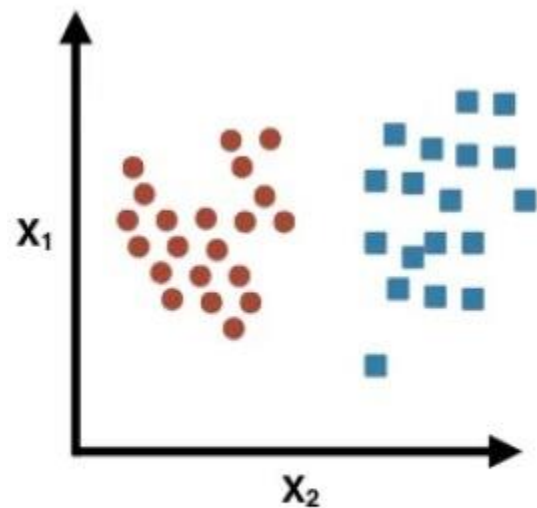


En un algoritmo de clasificación para aprendizaje automático se utilizan 20 variables de entrada para predecir la posibilidad de cáncer de mama. Este enfoque puede identificar con exactitud a pacientes que deben ser controladas cuidadosamente para detectar la enfermedad en forma temprana.

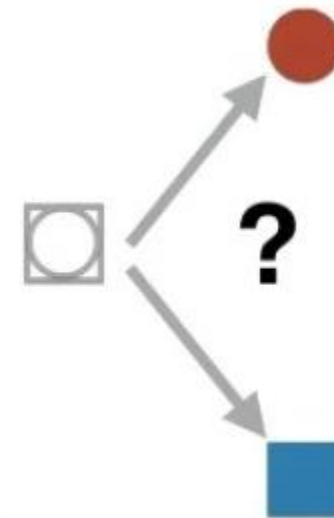
# Machine Learning for fraud detection and credit scoring



# Aprendizaje Supervisado (Clasificación)



1) Aprender de los  
datos de entrenamiento



2) Mapear nuevos  
datos (nunca vistos)

# Aprendizaje Supervisado (Clasificación)

## Set de entrenamiento

X Sample	y Target Value (Label)
$x_1$	Apple $y_1$
$x_2$	Lemon $y_2$
$x_3$	Apple $y_3$
$x_4$	Orange $y_4$



Classifier  
 $f : X \rightarrow Y$



En el momento del entrenamiento el clasificador utiliza ejemplos etiquetados para aprender las reglas para reconocer cada tipo de fruta.

Muestra







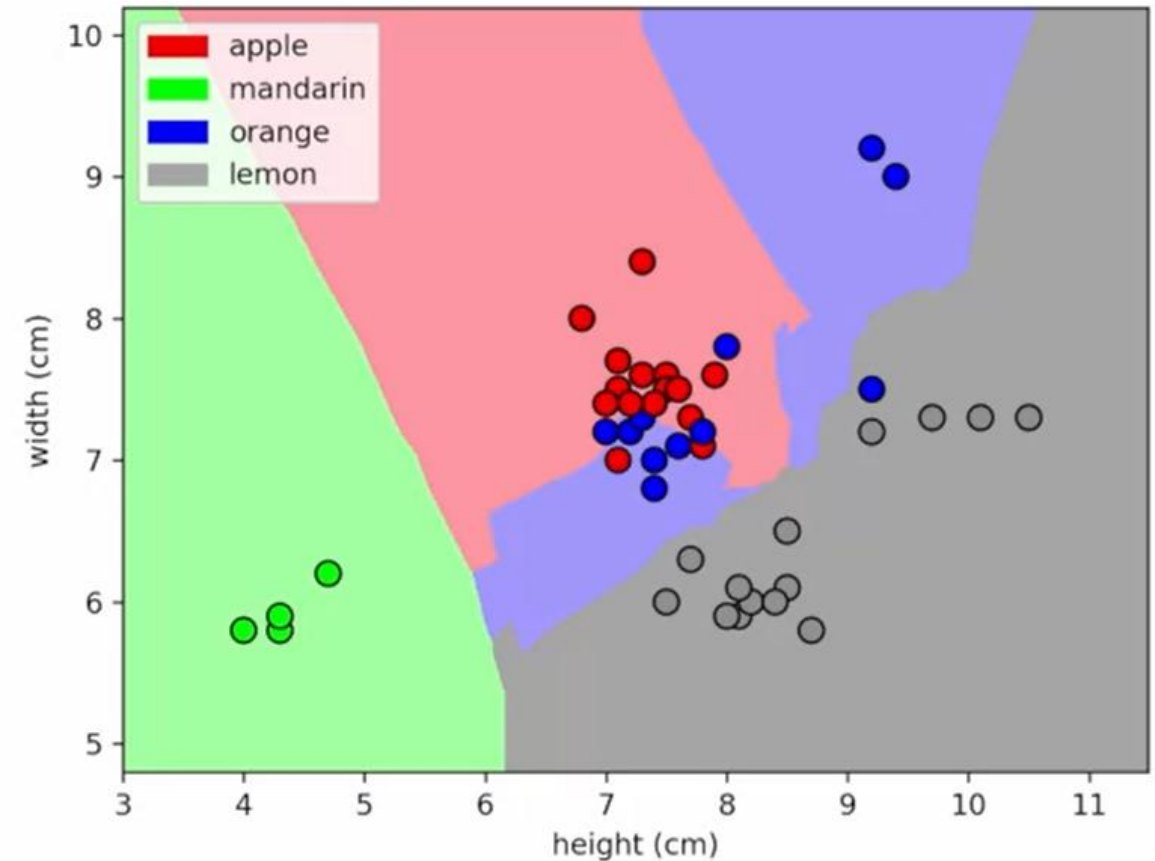
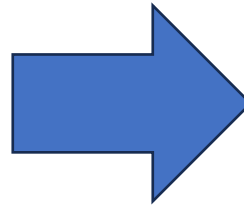
Etiqueta: Naranja

Después del entrenamiento, en el momento de la predicción, el modelo entrenado se usa para predecir el tipo de fruta para las nuevas instancias mediante las reglas aprendidas.



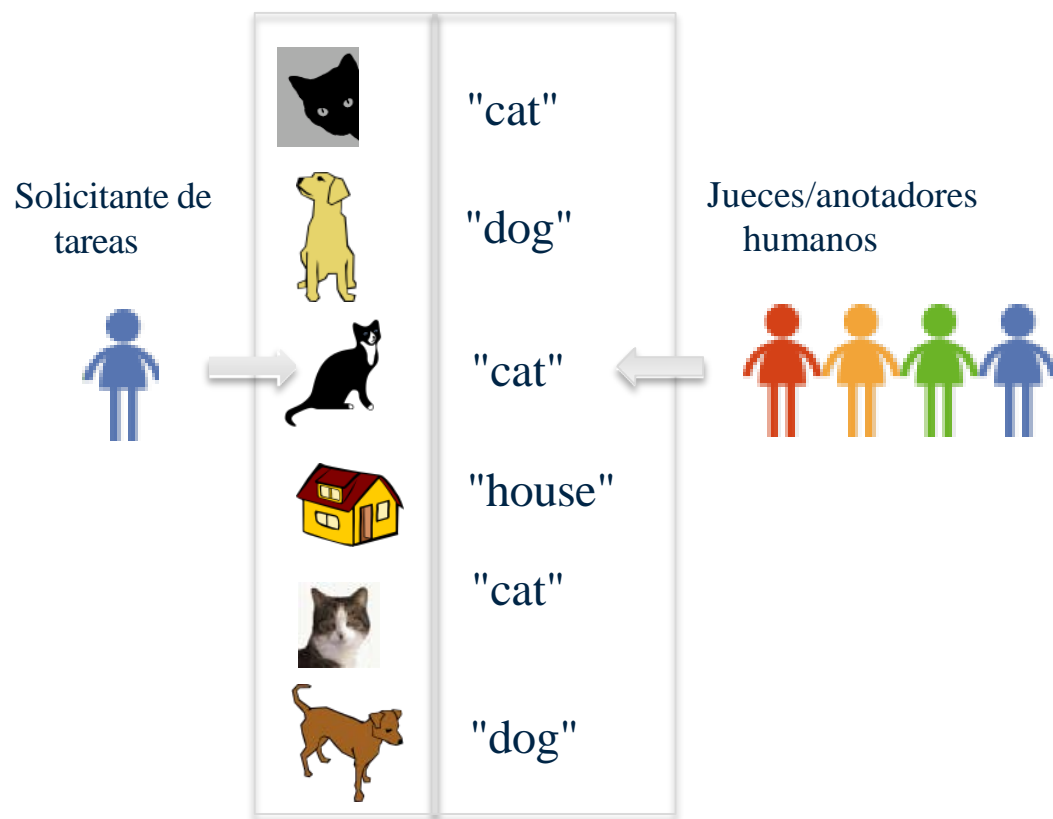
## Aprendizaje Supervisado (Clasificación)

X Sample		Y Target Value (Label)	
	$x_1$	Apple	$y_1$
	$x_2$	Lemon	$y_2$
	$x_3$	Apple	$y_3$
	$x_4$	Orange	$y_4$



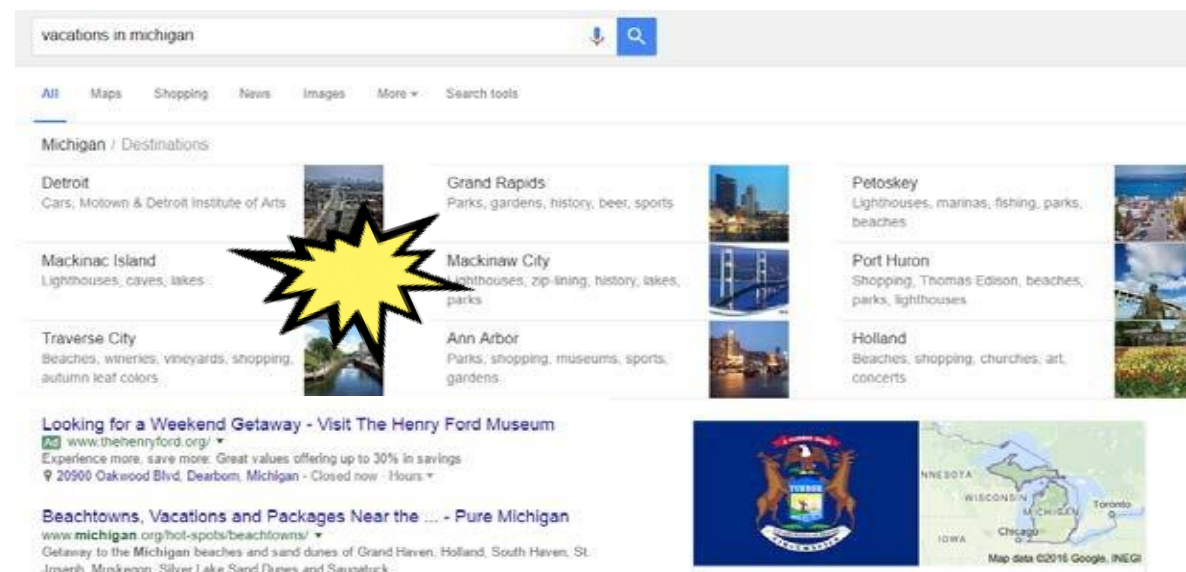
# Ejemplos de tipos de etiquetas

## Etiquetas explícitas



Plataforma de crowdsourcing

## Etiquetas implícitas



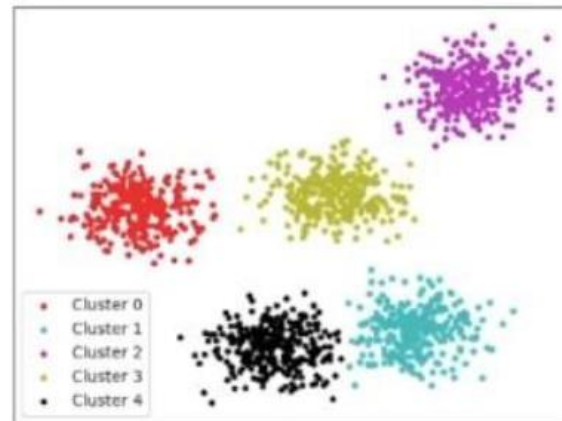
Hacer clic y leer el resultado de "Mackinac Island" puede ser una etiqueta implícita para que el motor de búsqueda aprenda que "Mackinac Island" es especialmente relevante para la consulta [vacaciones en Michigan] para ese usuario específico.

# Aprendizaje NO Supervisado

APRENDIZAJE NO  
SUPERVISADO



No se conoce la variable target





## Aprendizaje no supervisado

- Encontrar una estructura o conocimiento útil en los datos cuando no hay etiquetas disponibles.
- Encontrar la estructura en datos sin etiquetar
  - Busca grupos de instancias similares en los datos (agrupación en clústeres)
  - Búsqueda de patrones inusuales (detección de valores atípicos)

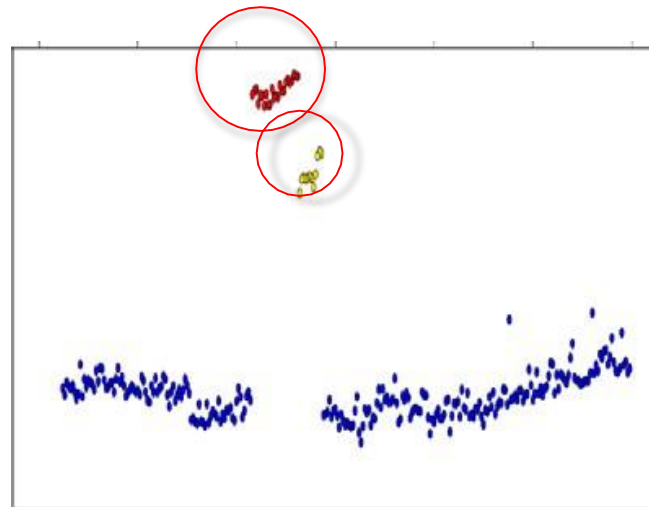


## Aprendizaje no supervisado

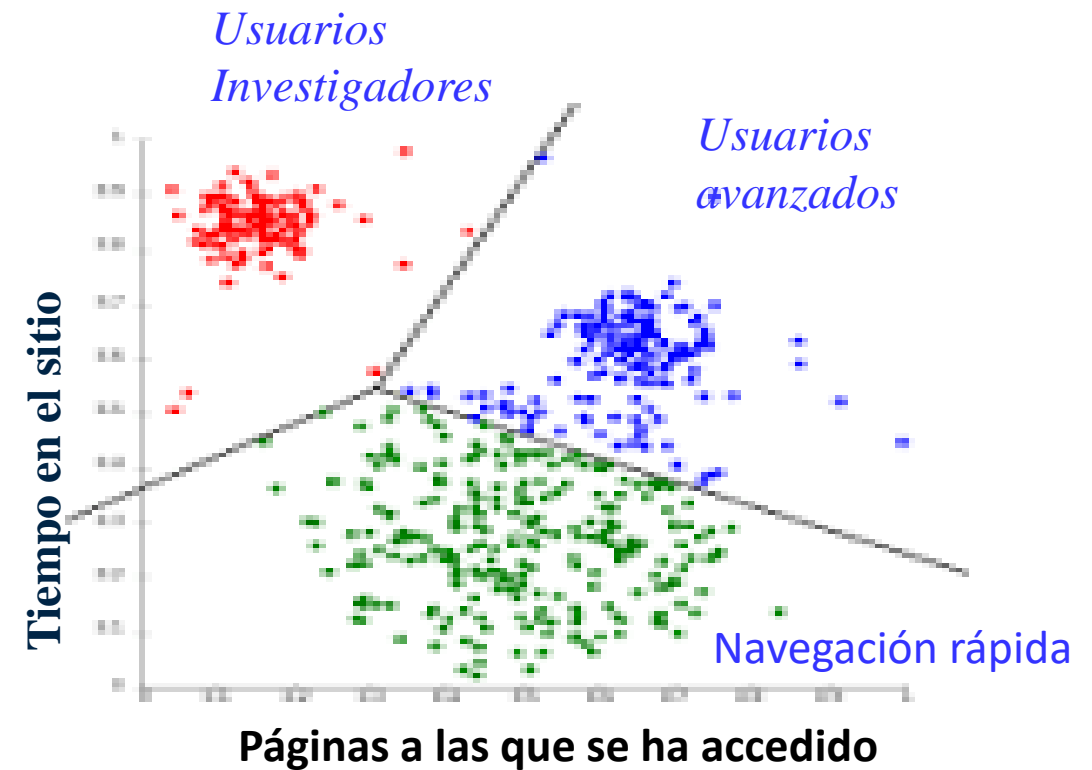
- Búsqueda de clústeres de usuarios similares (agrupación en clústeres)
- Detección de patrones de acceso anormales al servidor (detección de valores atípicos no supervisados)



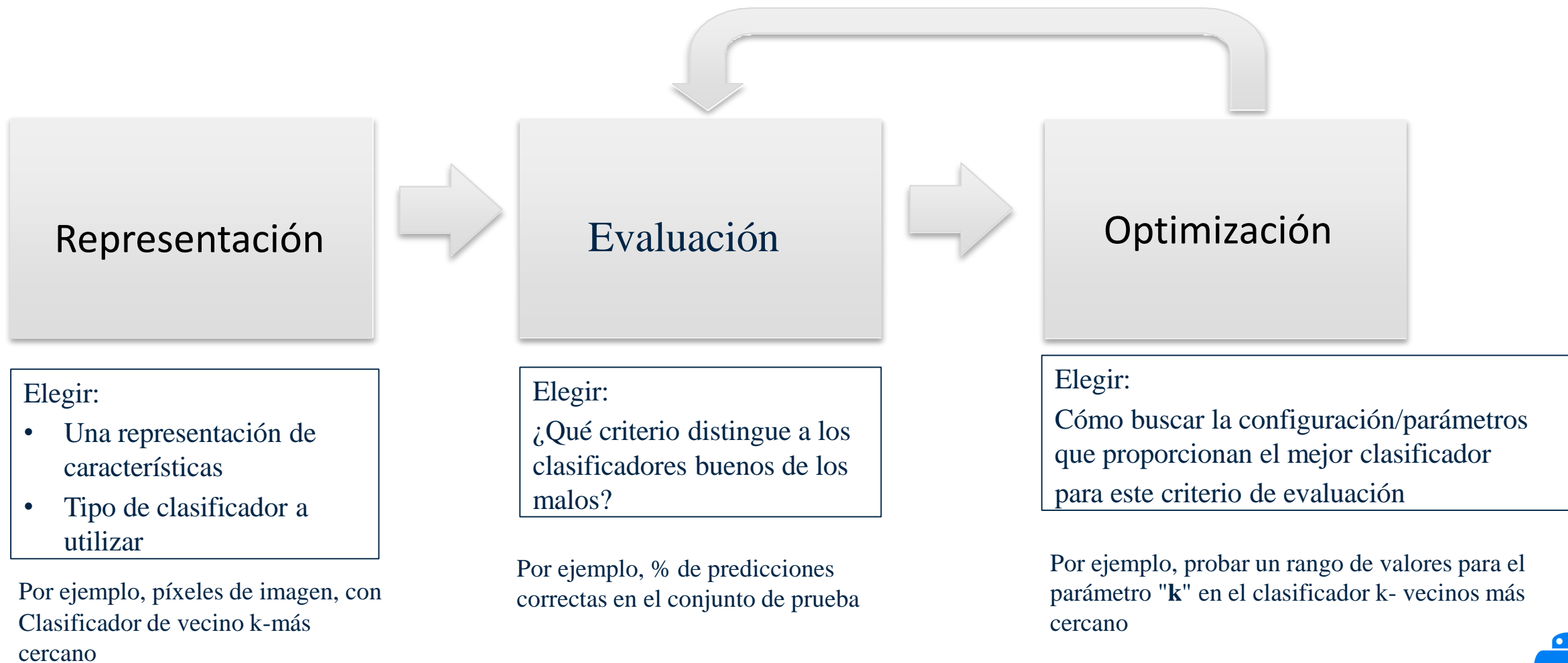
Accesos al servidor



Tiempo

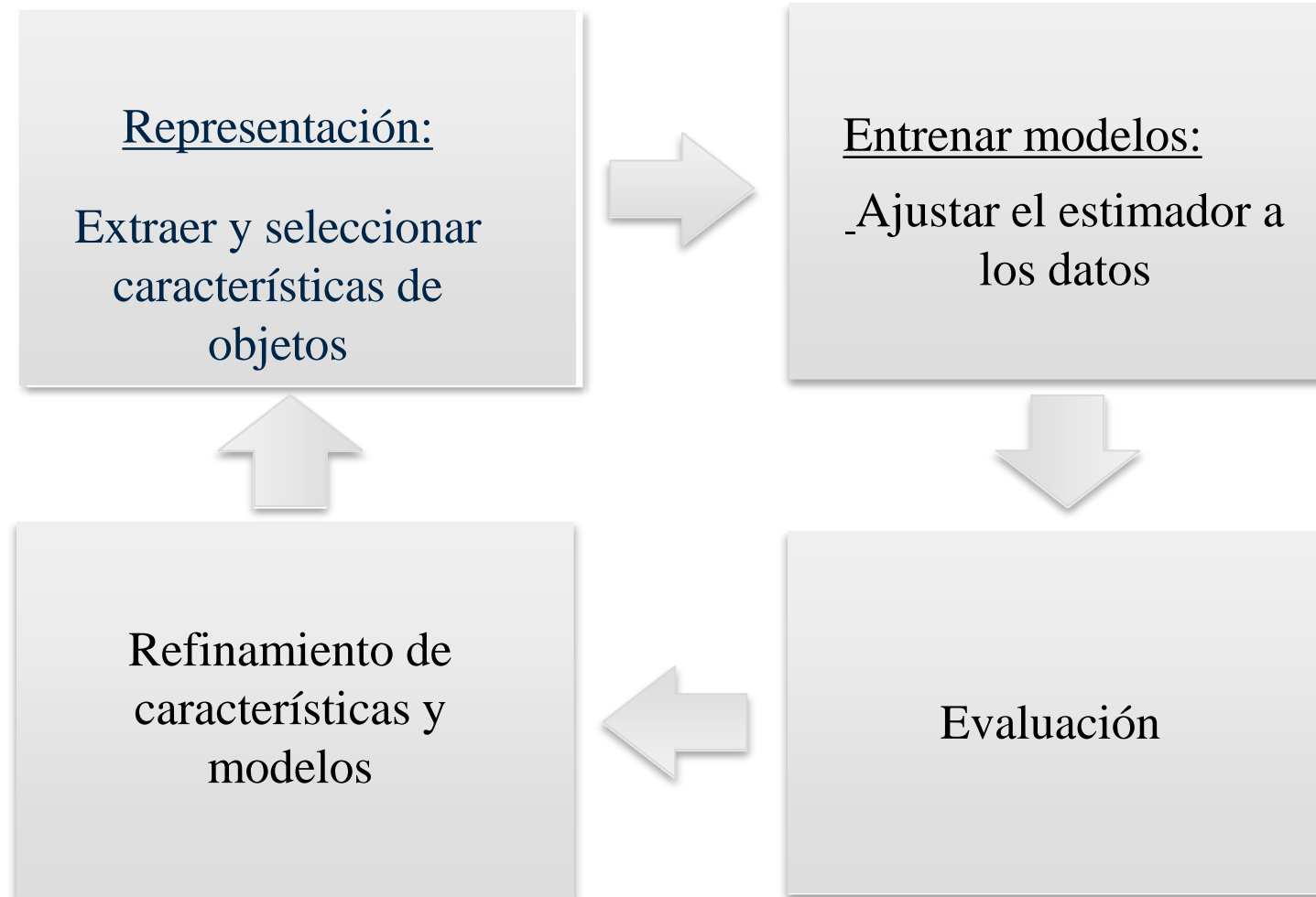


# Un flujo de trabajo básico de aprendizaje automático





## Representar / Entrenar / Evaluar / Refinar el ciclo



# Representaciones de características

## Representación de características

### Correo electrónico

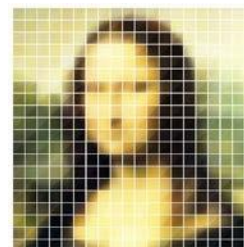
To: Chris  
Brooks From:  
Daniel Romero  
Subject: Next course  
offering Hi Daniel,  
Could you please send the  
outline for the next course  
offering? Thanks! -- Chris



<u>Feature</u>	<u>Count</u>
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2

Una lista de palabras con  
sus recuentos de frecuencia

### Imagen



Una matriz de valores de  
color (píxeles)

### Criaturas marinas

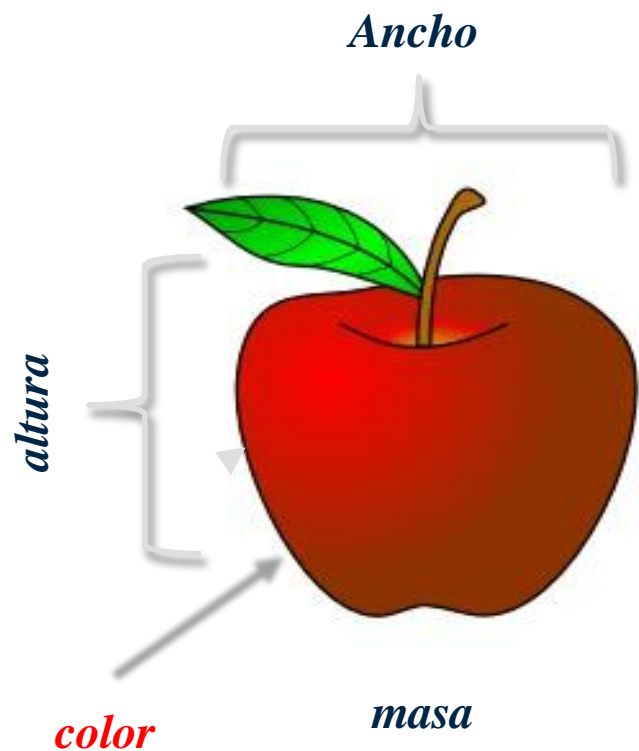


<u>Feature</u>	<u>Value</u>
DorsalFin	Yes
MainColor	Orange
Stripes	Yes
StripeColor1	White
StripeColor2	Black
Length	4.3 cm

Un conjunto de valores de  
atributo



# Representar una pieza de fruta como una matriz de características (más la información de la etiqueta)



## 1. Representación de características

Información de etiqueta  
(disponible solo en datos de  
entrenamiento)

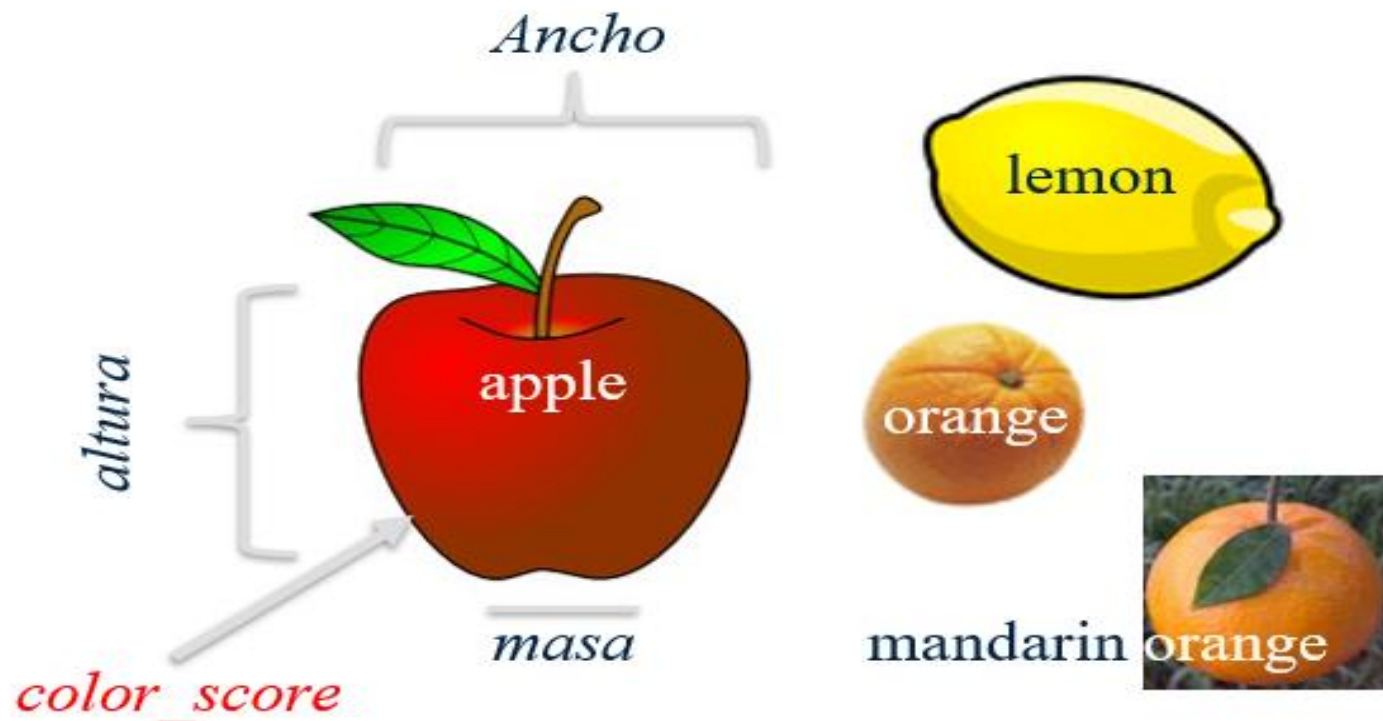
Información de etiqueta (disponible solo en datos de entrenamiento)				Características			
	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
18	1	apple	cripps_pink	162	7.5	7.1	0.83

## 2. Modelo de aprendizaje

Clasificador

Clase predicha  
(manzana)

# The Fruit Dataset



	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	golden_delicious	156	7.6	7.5	0.67

`fruit_data_with_colors.txt`

Créditos: Versión original del conjunto de datos de frutas creado por el Dr. Iain Murray, Univ. de Edimburgo

## Los datos de entrada en forma de tabla

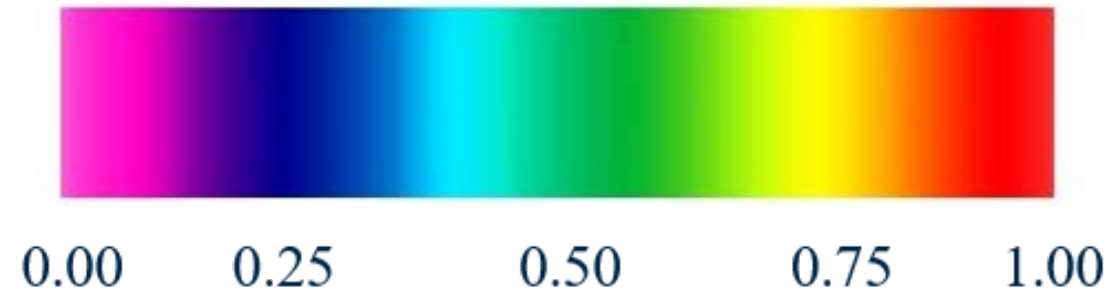
Cada fila corresponde a una  
única instancia de datos  
(ejemplo)

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.50
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	golden_delicious	156	7.6	7.5	0.67
17	1	apple	golden_delicious	168	7.5	7.6	0.73
18	1	apple	cripps_pink	162	7.5	7.1	0.83
19	1	apple	cripps_pink	162	7.4	7.2	0.85
20	1	apple	cripps_pink	160	7.5	7.5	0.88

La columna fruit\_label contiene la  
etiqueta para cada instancia de datos  
(ejemplo)

Las cuatro columnas contienen las  
características de cada instancia  
(muestra)

## La escala de la característica color\_score (simplista) utilizada en el conjunto de datos de frutas



Color category	color_score
Red	0.85 - 1.00
Orange	0.75 - 0.85
Yellow	0.65 - 0.75
Green	0.45 - 0.65



# Creación de conjuntos de entrenamiento y pruebas

X					Y		X_train					y_train		X_test					y_test	
	height	width	mass	color_score				height	width	mass	color_score				height	width	mass	color_score		
0	7.3	8.4	192	0.55	0	1	42	7.2	7.2	154	0.82	42	3	26	9.2	9.6	362	0.74	26	3
1	6.8	8.0	180	0.59	1	1	48	10.1	7.3	174	0.72	48	4	35	7.9	7.1	150	0.75	35	3
2	7.2	7.4	176	0.60	2	1	7	4.0	5.8	76	0.81	7	2	43	10.3	7.2	194	0.70	43	4
3	4.7	6.2	86	0.80	3	2	14	7.3	7.6	152	0.69	14	1	28	7.1	6.7	140	0.72	28	3
4	4.6	6.0	84	0.79	4	2	32	7.0	7.2	164	0.80	32	3	11	7.6	7.1	172	0.92	11	1
5	4.3	5.8	80	0.77	5	2	49	8.7	5.8	132	0.73	49	4	2	7.2	7.4	176	0.60	2	1
6	4.3	5.9	80	0.81	6	2	29	7.4	7.0	160	0.81	29	3	34	7.8	7.6	142	0.75	34	3
7	4.0	5.8	76	0.81	7	2	37	7.3	7.3	151	0.79	37	3	46	10.2	7.3	216	0.71	46	4
8	7.8	7.1	178	0.92	8	1	56	7.3	7.3	151	0.73	56	4	40	7.5	7.1	154	0.78	40	3
9	7.0													22	7.1	7.3	140	0.87	22	1
10	7.3													4	4.6	6.0	84	0.79	4	2
11	7.6																	0.93	10	1
12	7.1																		30	3
13	7.7																			
14	7.3	7.6	152	0.69	14	1	3	4.7	6.2	86	0.80	3	2	30	7.5	7.1	150	0.79	30	3
15	7.1	7.7	156	0.69	15	1	0	7.3	8.4	192	0.55	0	1	41	8.2	7.6	180	0.79	41	3
16	7.5	7.6	156	0.67	16	1	53	8.4	6.0	120	0.74	53	4	33	8.1	7.5	190	0.74	33	3
17	7.6	7.5	168	0.73	17	1	47	9.7	7.3	196	0.72	47	4							
18	7.1	7.5	162	0.83	18	1	44	10.5	7.3	200	0.72	44	4							
19	7.2	7.4	162	0.85	19	1														

```
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

`X_train, X_test, y_train, y_test  
= train_test_split(X, y)`

Conjunto de datos original

Conjunto de entrenamiento

Conjunto de prueba

# Algunas razones por las que es importante revisar los datos inicialmente

- La inspección de los valores de las características puede ayudar a **identificar qué limpieza o preprocesamiento aún debe realizarse** una vez que pueda ver el rango o la distribución de valores que es típica para cada atributo.
- Es posible que observe **datos faltantes o ruidosos**, o incoherencias, como el uso de un **tipo de datos incorrecto** para una columna, **unidades de medida incorrectas** para una columna determinada o que **no hay suficientes ejemplos de una clase determinada**.
- Es posible que se dé cuenta de que su problema se puede resolver sin aprendizaje automático.

## Ejemplos de valores de entidad incorrectos o faltantes

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	192
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	apple	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn		7.4	7.0	0.89
10	1	apple	braeburn		6.9	7.3	0.93
11	1	apple	braeburn		7.1	7.6	0.92
12	1	apple	braeburn		7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69

# Referencias

## Python Intermedio

<https://python-intermedio.readthedocs.io/es/latest/>

## Pandas\_Cheat\_Sheet.

[https://pandas.pydata.org/Pandas\\_Cheat\\_Sheet.pdf](https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)