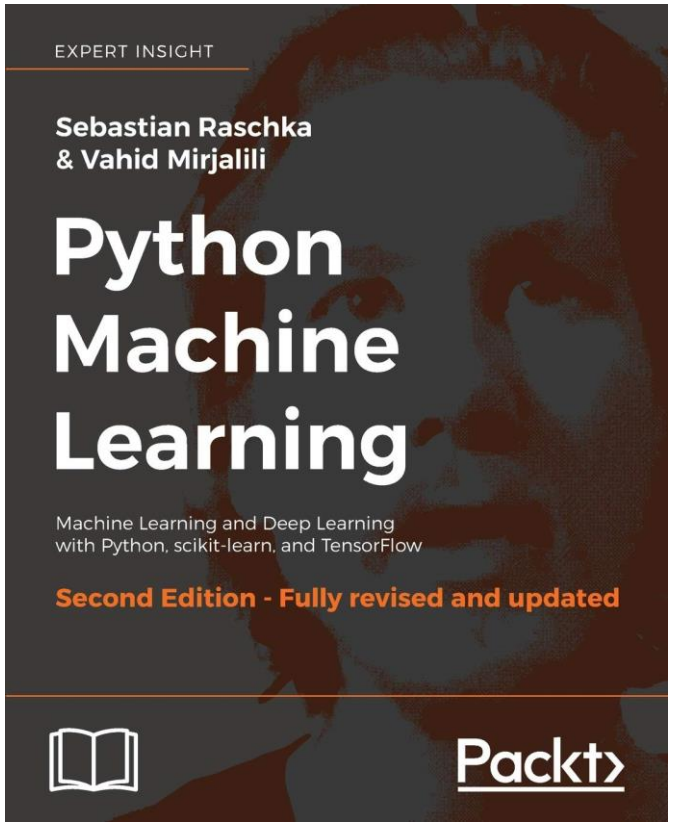
Tipos y métodos de Machine Learning

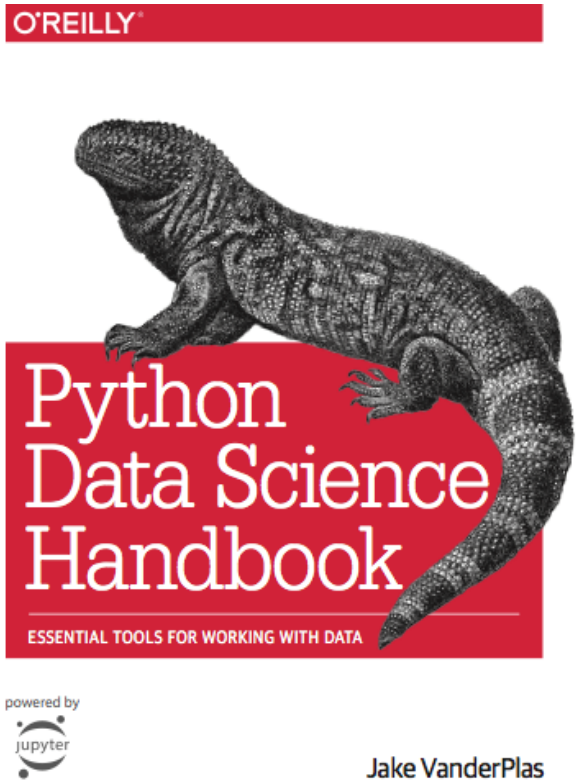
Equipo Innovación Copec

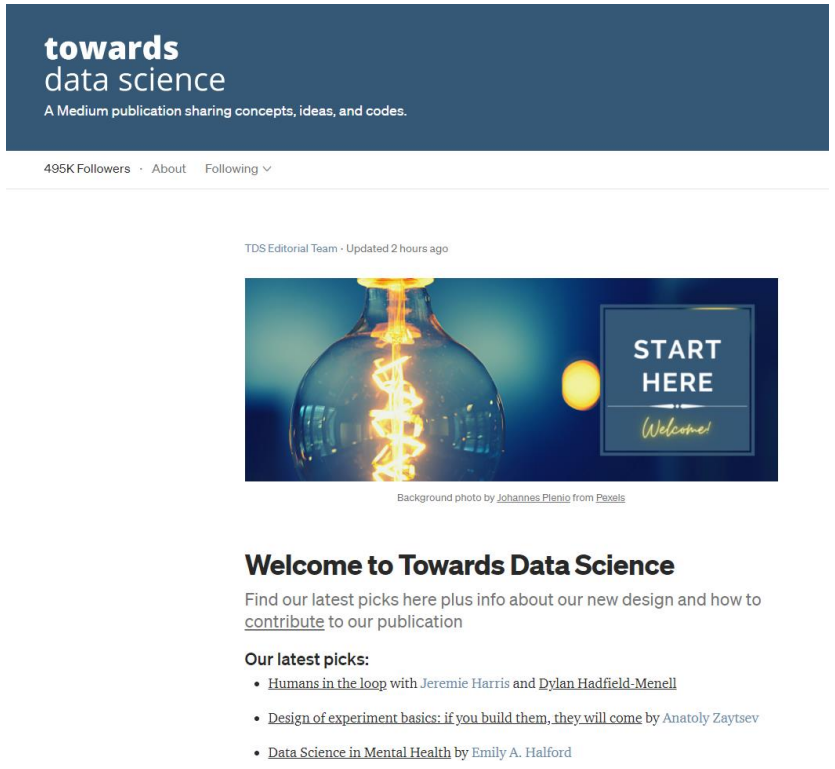
Roberto Muñoz
Senior Data Scientist
Digital Data Analytics
EY

Bibliografía

Libros y medios digitales



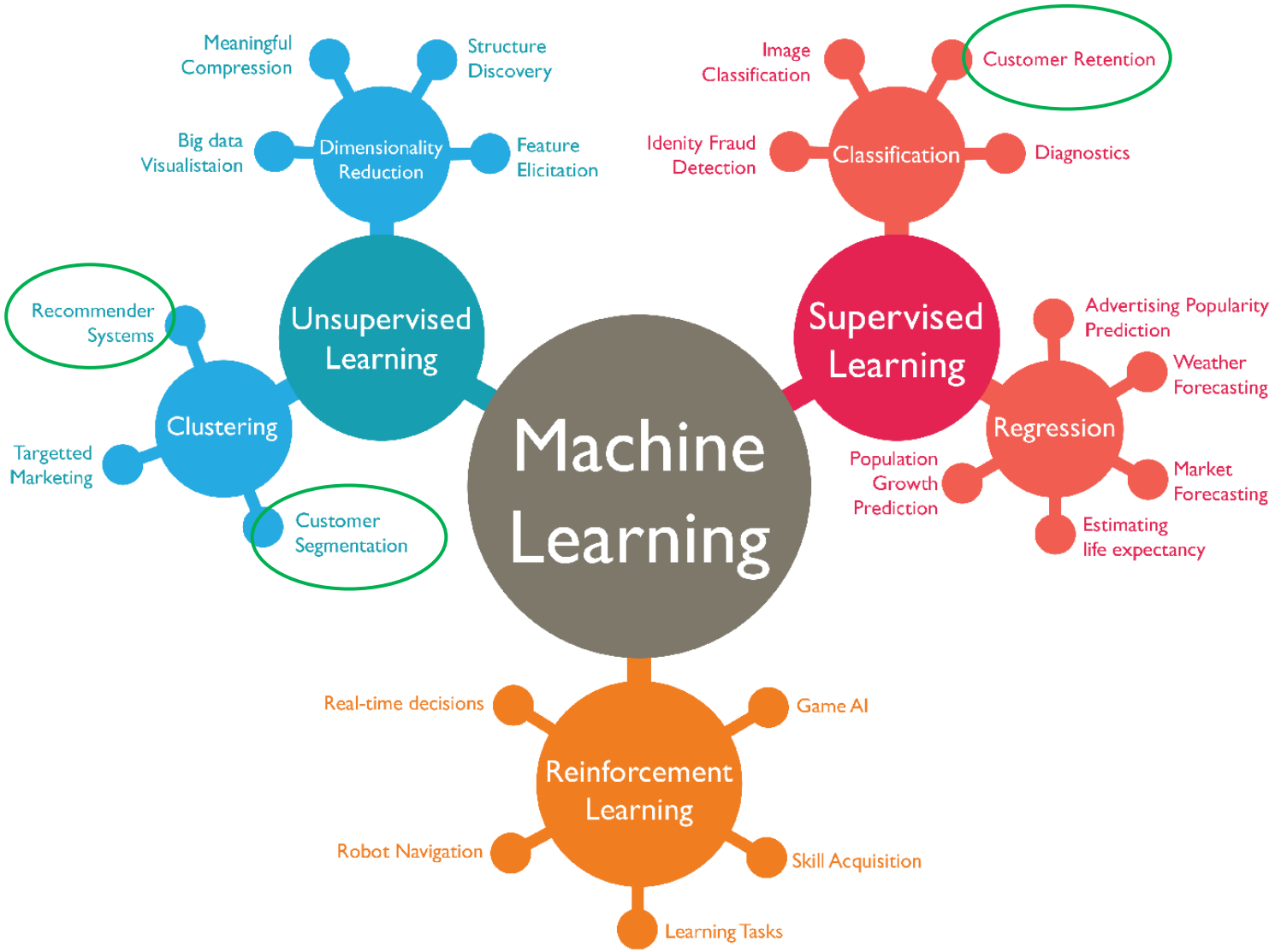




https://jakevdp.github.io/PythonDataScienceHandbook/index.html

https://towardsdatascience.com/

Tipos de aprendizaje



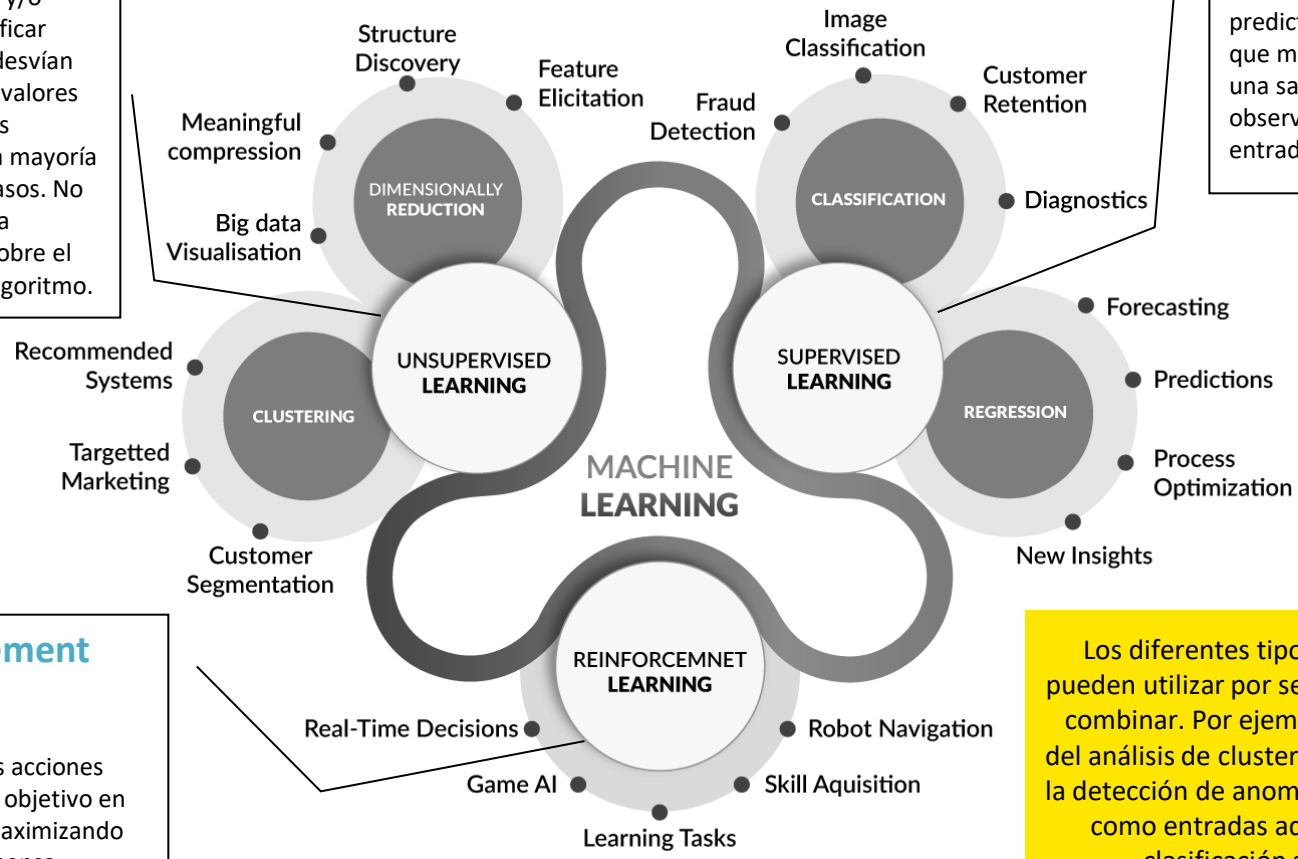
Tipos de aprendizaje

Unsupervised learning

Los modelos pueden agrupar datos no estructurados y/o pueden identificar casos que se desvían fuertemente (valores atípicos) de los patrones de la mayoría de los otros casos. No se proporciona información sobre el resultado al algoritmo.

Supervised learning

Construyendo modelos predictivos o de clasificación que mapean una entrada a una salida en base a observaciones de pares de entrada-salida anteriores.

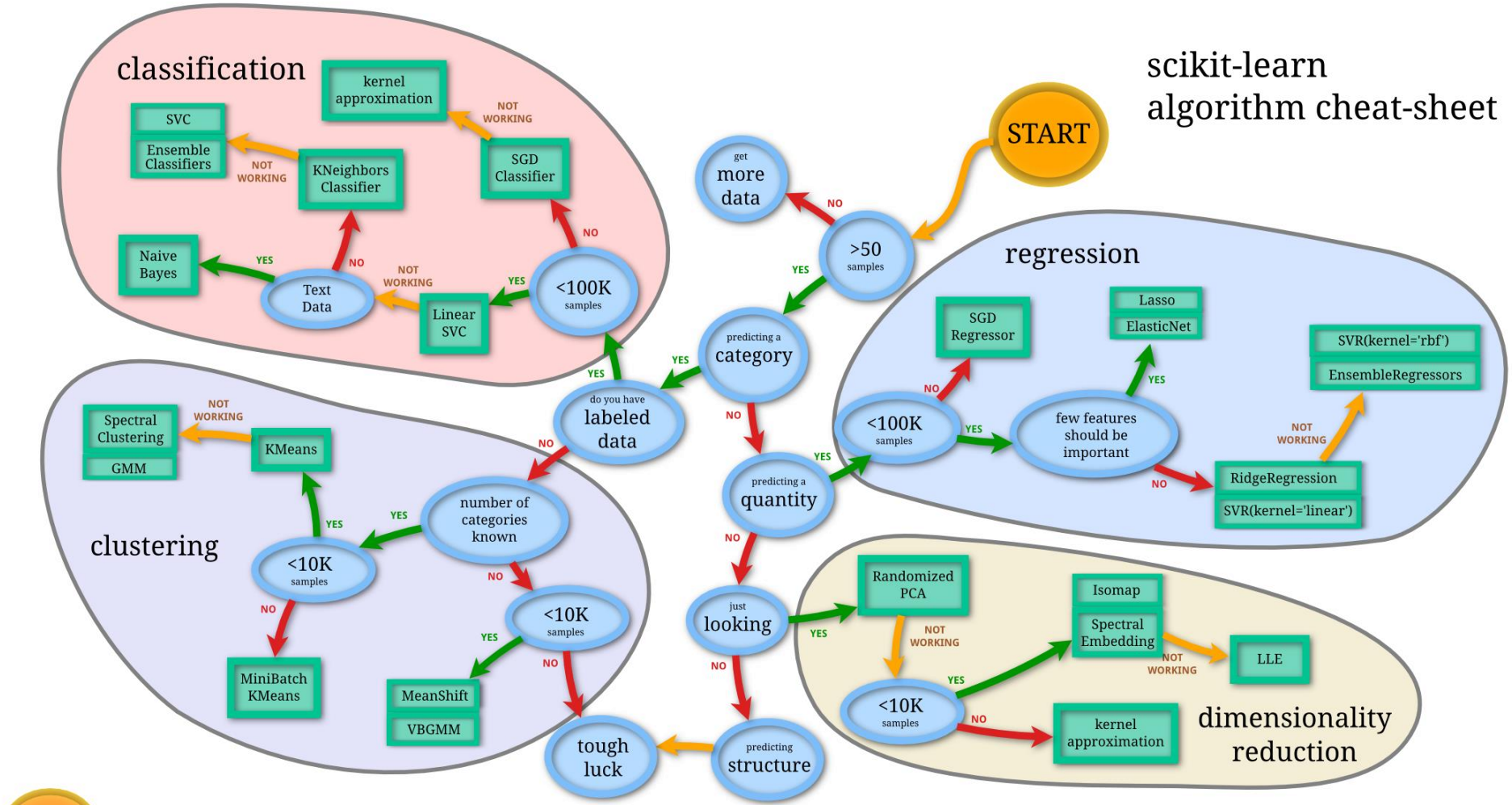


Reinforcement learning

Determinar las acciones para lograr un objetivo en un entorno, maximizando alguna recompensa.

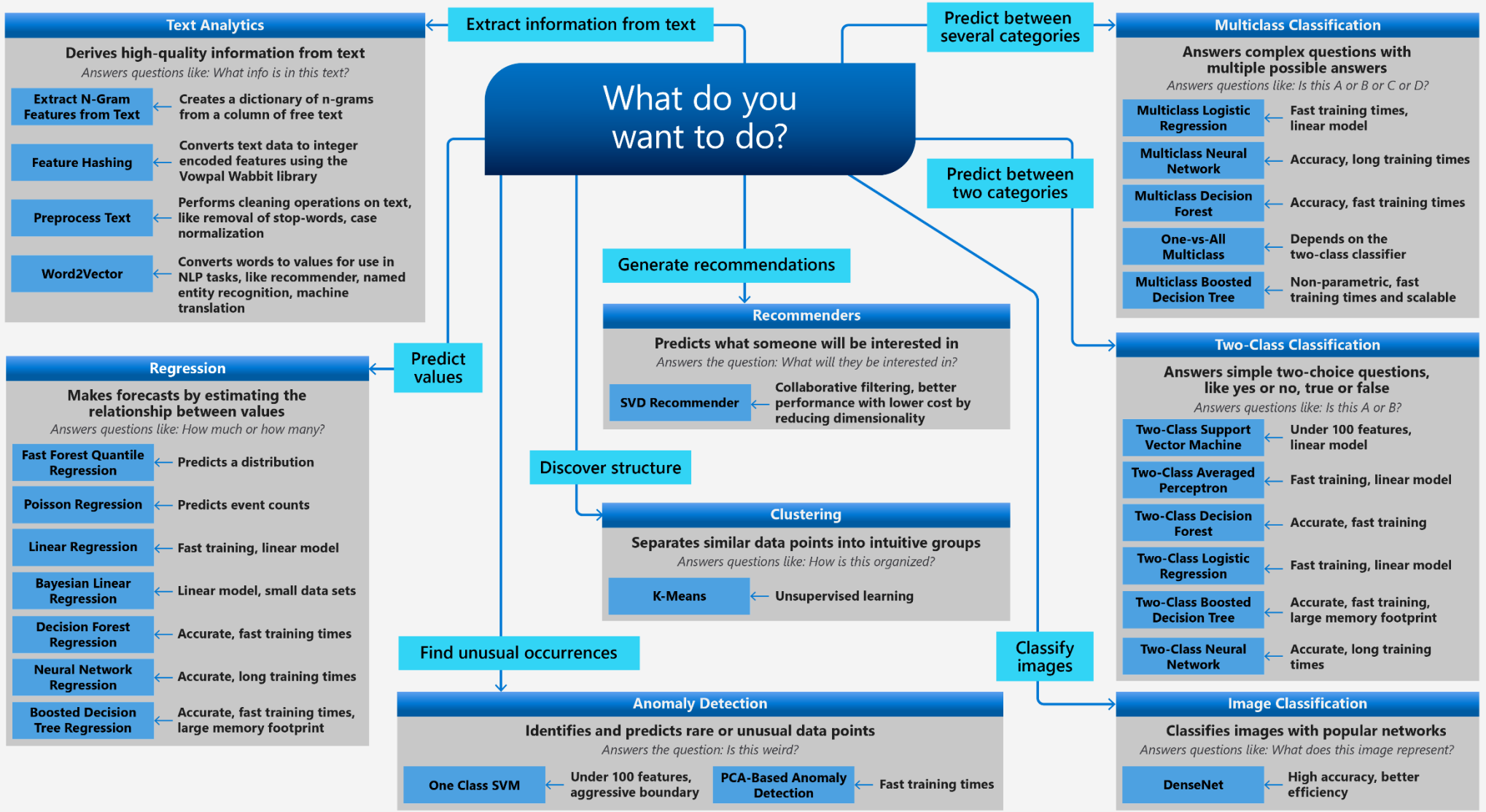
Los diferentes tipos de enfoques se pueden utilizar por separado o se pueden combinar. Por ejemplo: Los resultados del análisis de clustering no supervisado y la detección de anomalías se pueden usar como entradas adicionales en una clasificación supervisada.

Scikit-learn methods



Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.



 Share this poster: aka.ms/mlcheatsheet

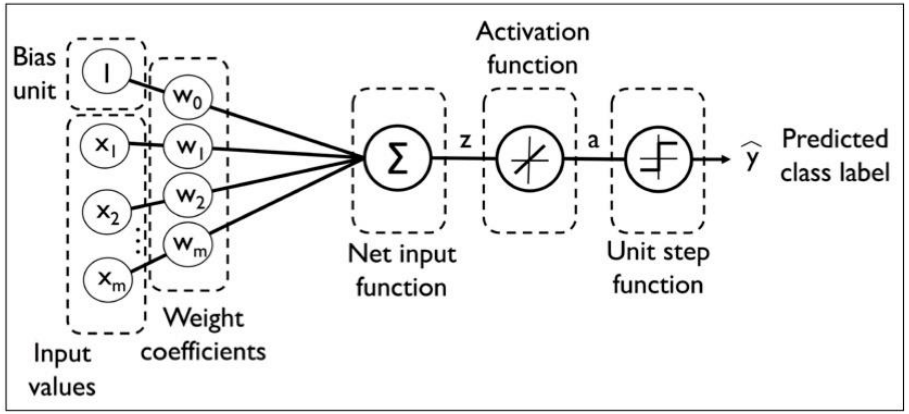
Microsoft

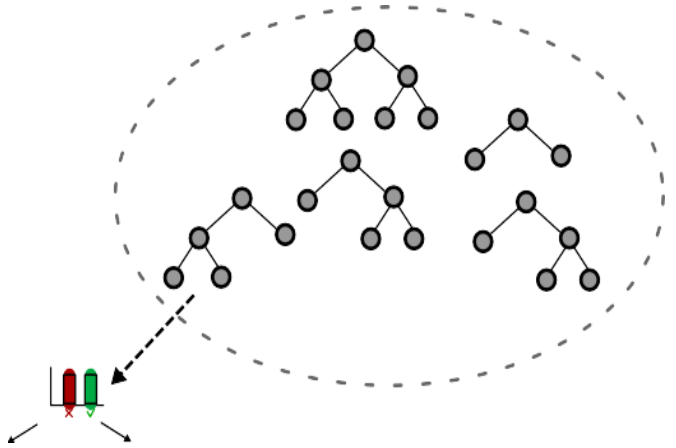
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet

Supervised learning

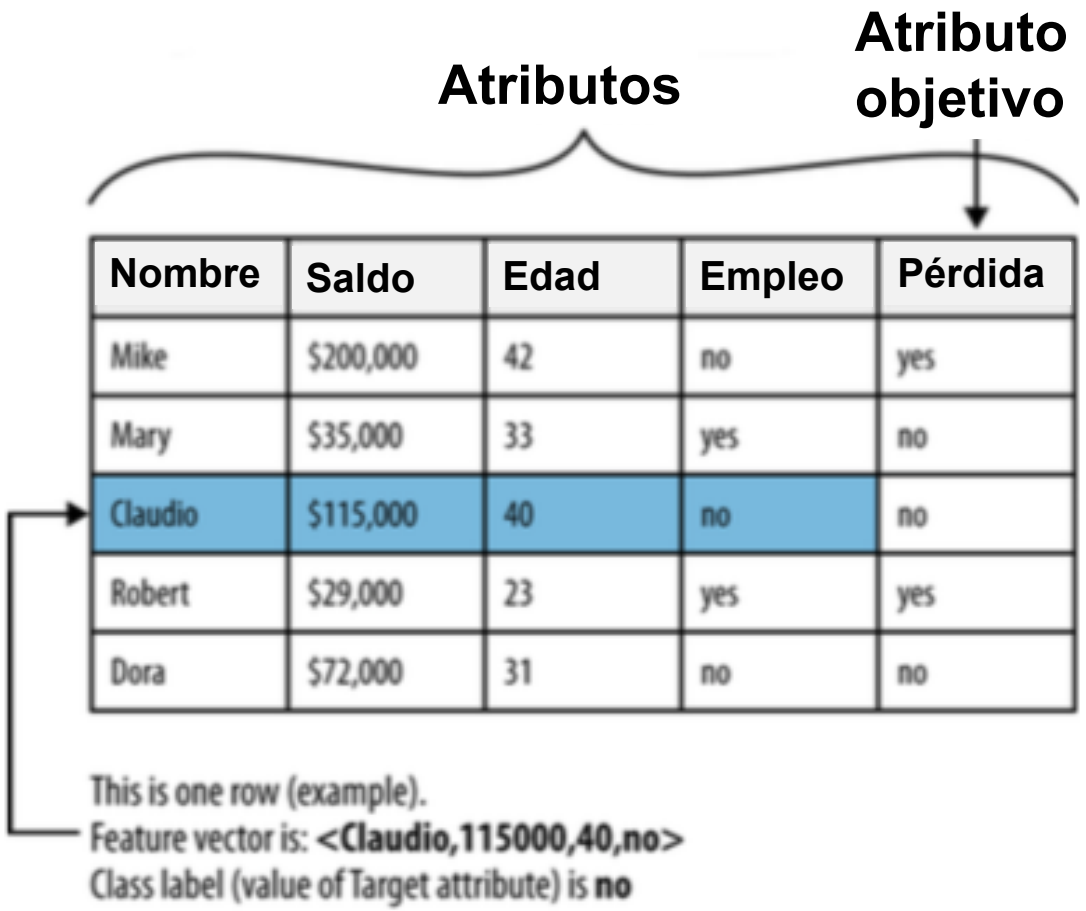
Clasificación

- Los métodos más usados para resolver los problemas de clasificación en ML son
  - Support Vector Machine (SVM)
  - Arbol de decisión
  - Random forest
  - Deep Neural Networks





Evaluación de créditos bancarios



Clasificación

1)  → perro

2)  → perro

3)  → gato

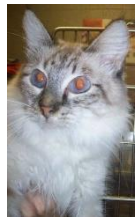
4)  → perro

5)  → perro

6)  → gato

7)  → perro

8)  → gato

9)  → perro

10)  → gato

Matriz de confusión

id	observado	prediccion
1	perro	perro
2	perro	perro
3	perro	gato
4	perro	perro
5	perro	perro
6	perro	gato
7	perro	perro
8	gato	gato
9	gato	perro
10	gato	gato

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

Matriz de confusión

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Inglés

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Español

Recall, Precision, Accuracy

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Recall**: Sensibilidad
- **Precision**: Precisión
- **Accuracy**: Exactitud

- Entrenamiento: Ajustar los parámetros del algoritmo de forma tal de que se minimicen la cantidad de predicciones que no correspondan a la etiqueta original.
- Recall: Porcentaje de clasificados correctamente como positivos sobre todos los que realmente eran positivos.
- Precision: Porcentaje de clasificados correctamente como positivos sobre todos los clasificados como positivos.
- Accuracy: Porcentaje de clasificados correctamente.

Unsupervised learning

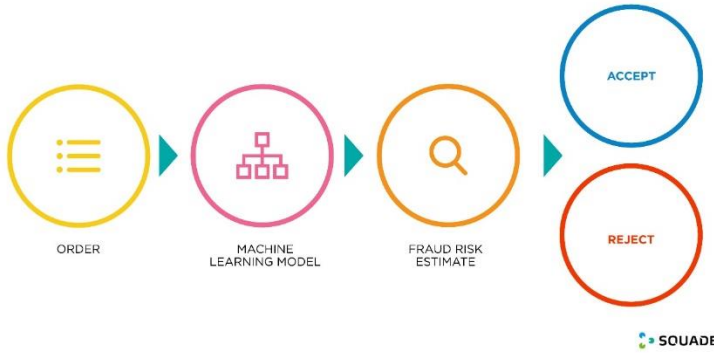
Créditos: Interstellar

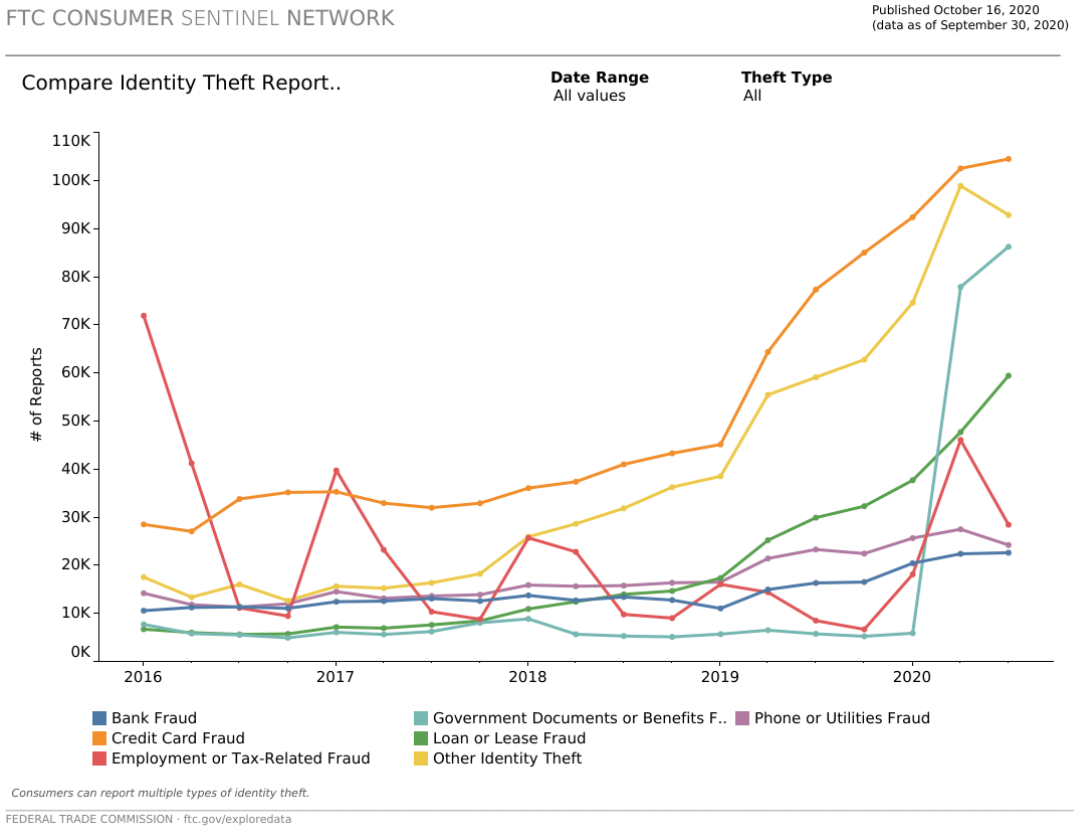
Casos de uso

- **Detección de fraudes**

Los fraudes con tarjeta de crédito aumentaron un 104% entre Q1 2019 y Q2 2020.

Usar modelo para identificar operaciones fraudulentas.





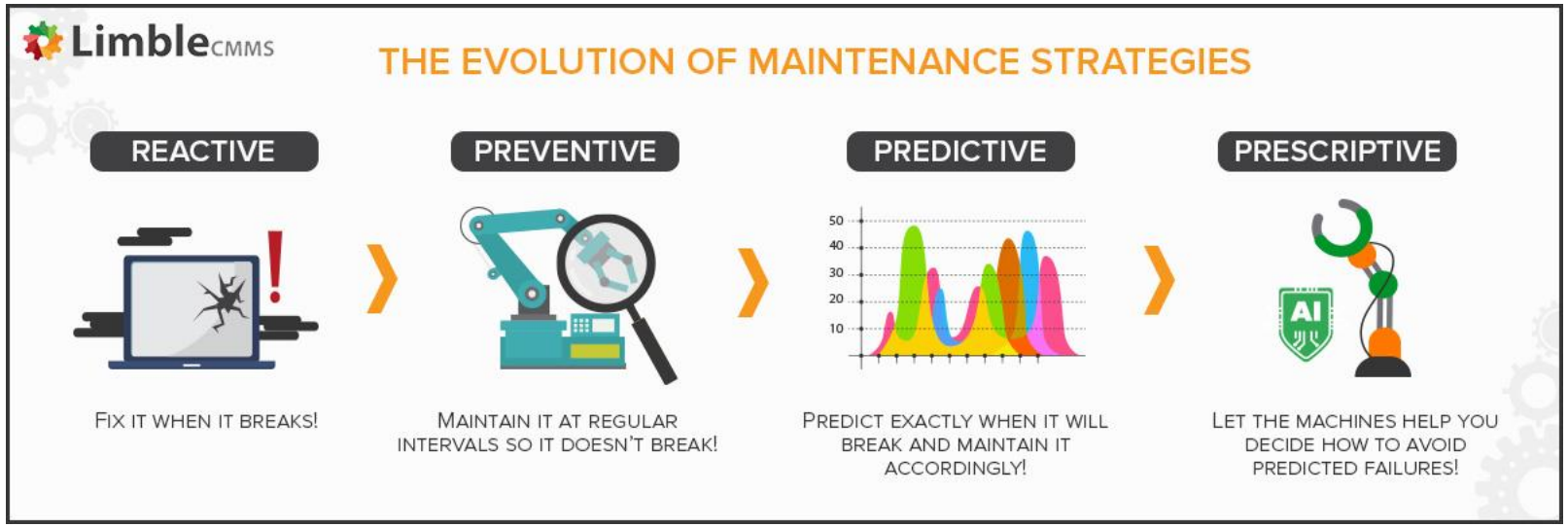
https://public.tableau.com/profile/federal.trade.commission#!/vizhome/IdentityTheftReports/TheftTypesOverTime

Casos de uso

- Mantención preventiva de fallas

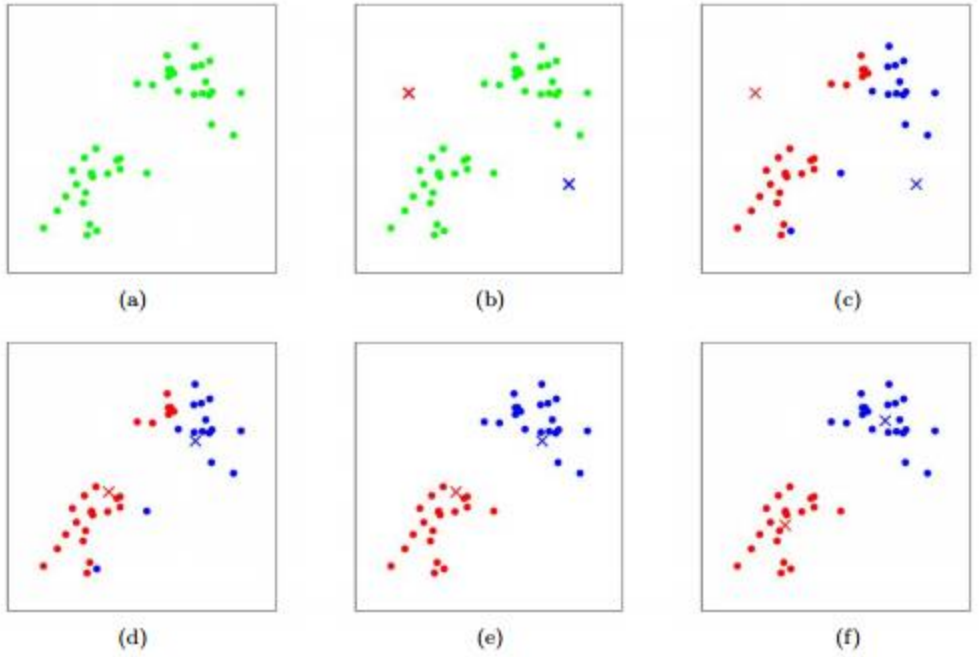
Las industrias de las telecomunicaciones y manufactura están constantemente recolectando datos de sus operaciones. Cuentan con máquinas equipadas con múltiples sensores.

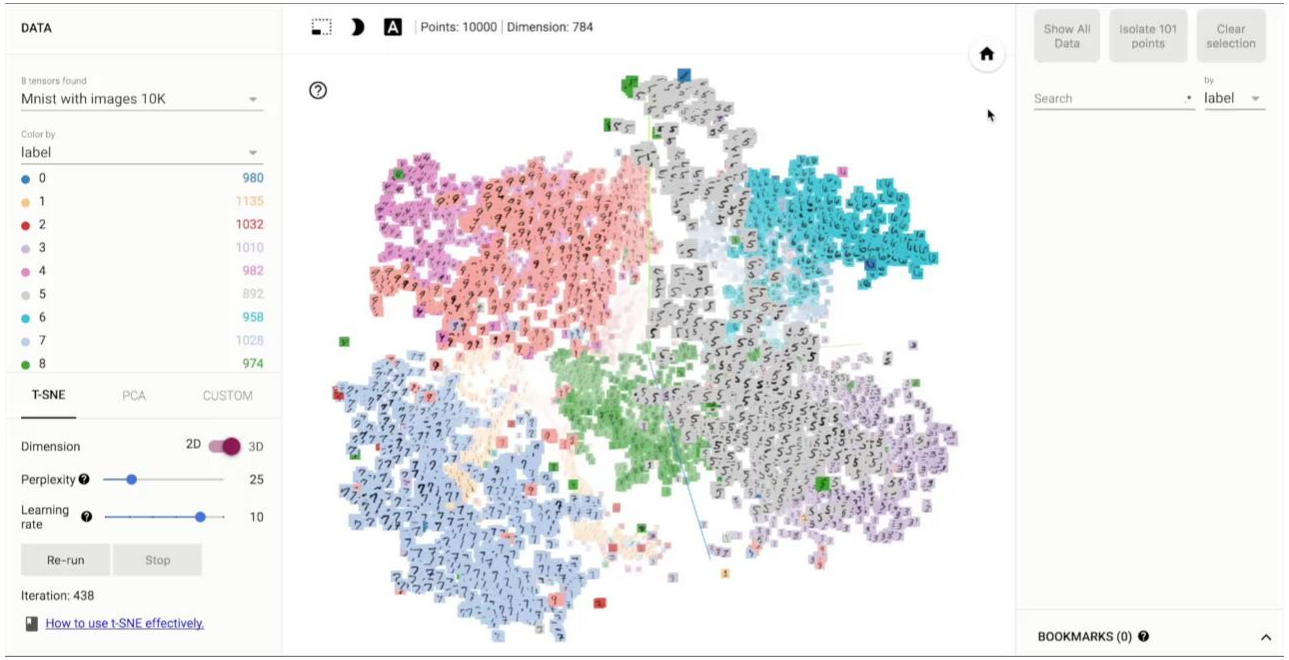
Usar modelos para identificar fallas de manera temprana y hacer mantenciones a la maquinaria.



Tipos de ML

- Aprendizaje no supervisado
  - Entrenar un modelo usando datos que no han sido clasificados previamente.
  - El sistema debe poder reconocer patrones y generar sus propias etiquetas. El modelo debe clasificar los nuevos datos de entrada.





Tipos de tareas en Aprendizaje no supervisado

- **Clustering o Agrupación**

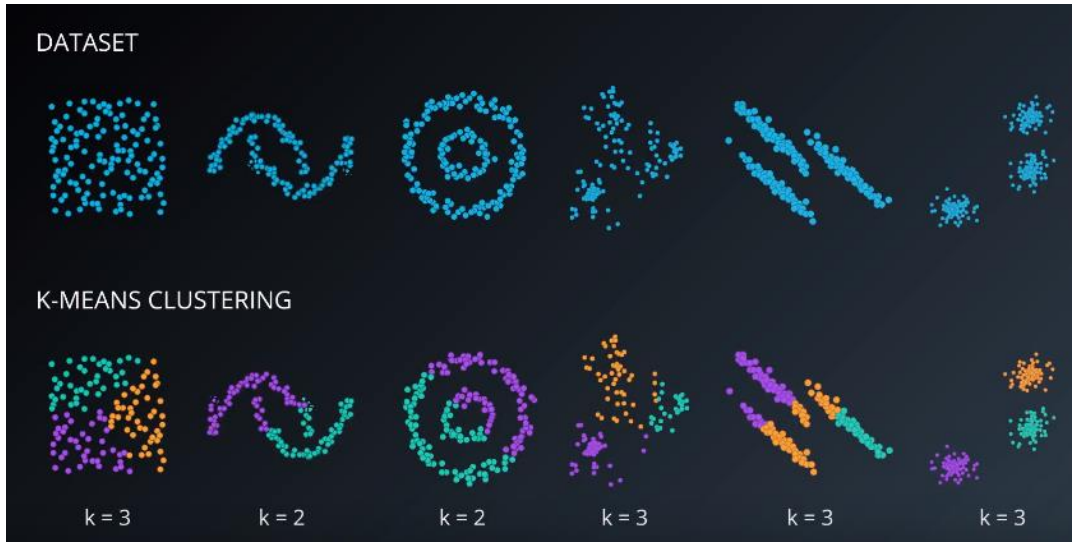
  Encontrar diferentes grupos o segmentos presentes en los datos. Usado principalmente en segmentación de clientes y sistemas de recomendación.

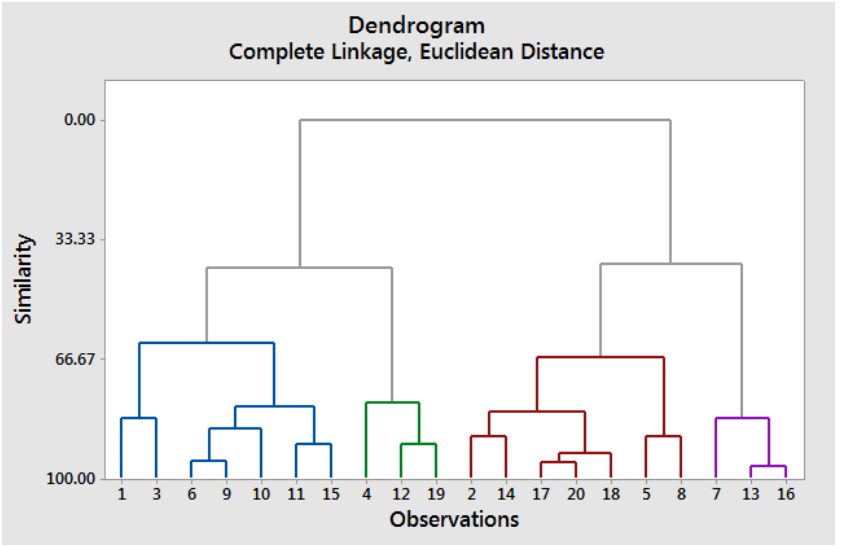
- **Detección de anomalías**

  Analizar el comportamiento y los patrones regulares en los datos y detectar desviaciones respecto al comportamiento normal. Usado principalmente en análisis de series de tiempo y análisis de imágenes.

Métodos de clustering

- K-means
- Mean-shift
- Density-Based Spatial Clustering of Applications with Noise
- Gaussian mixture models
- Clustering jerquizado

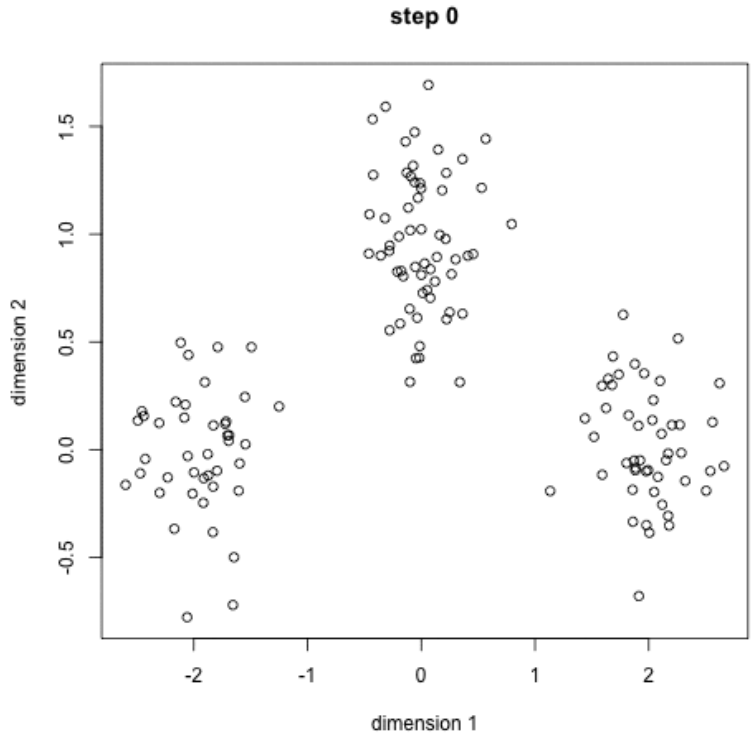




Método de K-means

- K-Means tiene como objetivo encontrar y agrupar en clases los puntos de datos que tienen una alta similitud entre ellos.

1. Primero, necesitamos elegir k, el número de clusters que queremos que nos encuentren.
2. Luego, el algoritmo seleccionará aleatoriamente los centroides de cada grupo.
3. Se asignará cada punto de datos al centroide más cercano (utilizando la distancia euclídea).
4. Se calculará la inercia del conglomerado.
5. Los nuevos centroides se calcularán como la media de los puntos que pertenecen al centroide del paso anterior. En otras palabras, calculando el error cuadrático mínimo de los puntos de datos al centro de cada cluster, moviendo el centro hacia ese punto.
6. Volver al paso 3.



Taller