

BIG DATA for BUSINESS

2.4 Introducción a Data Science y Machine Learning

Conecta Empleo

Contenido desarrollado por
Synergic Partners



Índice del módulo

2.4 INTRODUCCIÓN DE DATA SCIENCE Y MACHINE LEARNING


- ¿Qué es Data Science?
- Etapas de un proyecto analítico
- ¿Qué es Machine Learning?
- Tipos de aprendizaje
- Técnicas básicas de Machine Learning
- Evaluación y selección de modelos analíticos



¿Qué es Data Science?

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Introducción

A decorative graphic on the left side of the slide, consisting of a network of blue dots connected by thin, light blue lines, forming a complex, interconnected web-like structure.

En los últimos años se ha producido una explosión en el volumen de los **datos** disponibles. La tecnología actual permite expresar las actividades comerciales, industriales, e individuales en datos digitales que son almacenados y procesados con el fin de extraer **valor** para las empresas y sus clientes.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

¿Qué es Data Science?

“

Data Science es un campo multidisciplinar que combina principios, procesos, y técnicas que permiten entender fenómenos mediante el análisis automatizado de grandes volúmenes de datos.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Diferencias con otras disciplinas

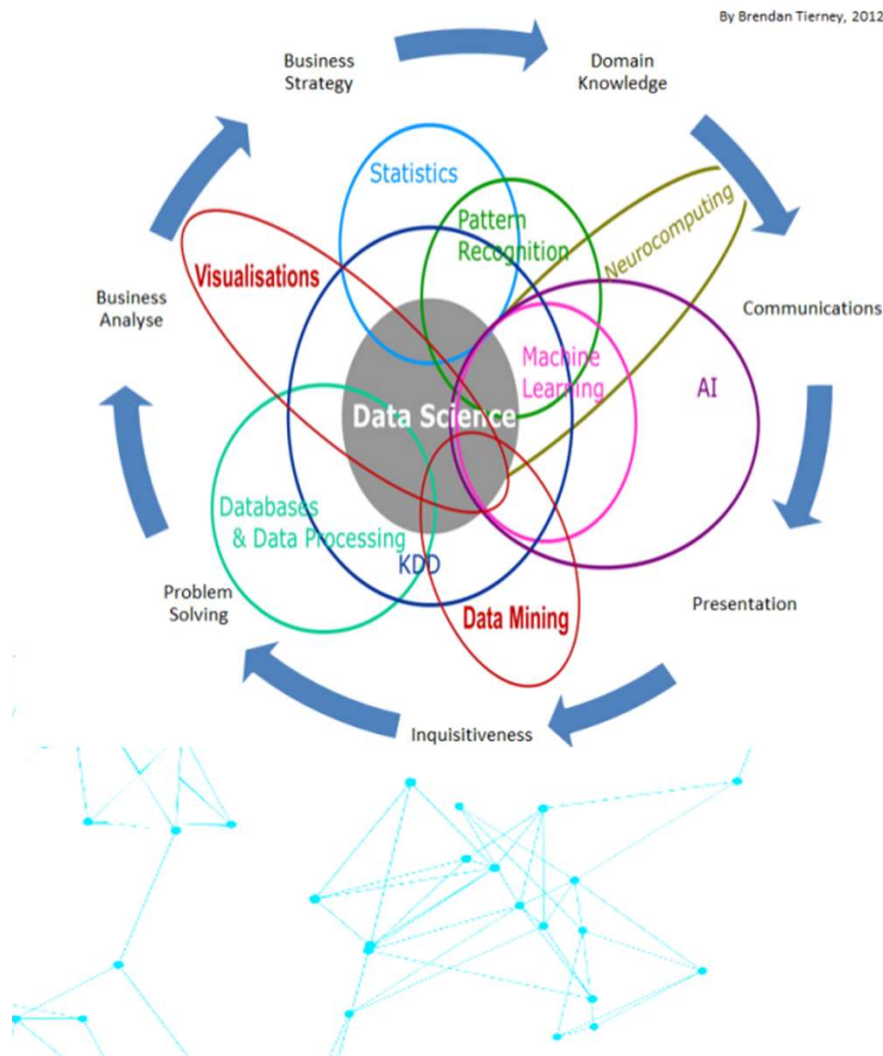
“

Data mining es el proceso de extraer patrones ocultos y desconocidos a partir de un conjunto de datos crudos, con el objetivo de transformar grandes cantidades de datos en información útil.

Ambos son términos que se refieren al proceso de transformar grandes volúmenes de datos en conocimiento; sin embargo, data science se entiende como un conjunto de principios fundamentales que sirven de guía para la extracción de conocimiento, mientras data mining se refiere a las tecnologías que incorporan estos principios.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Disciplinas de Data Science

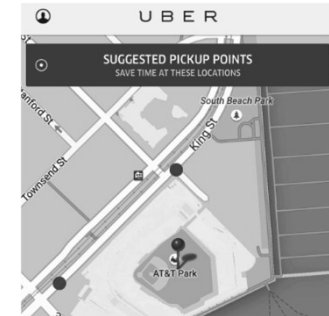
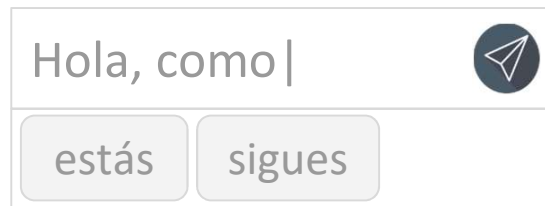


En entornos **Big Data** el análisis de datos debe combinar métodos y técnicas de distintas disciplinas, como:

- Estadística
- Bases de Datos
- Informática
- Machine Learning
- Visualización, entre otras.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Data Science en el *día a día*



Texto predictivo

Anticipar la entrada de usuario presentándole sugerencia de la palabra que escribirá a continuación.

Web search / Internet

Los motores de búsqueda y el contenido personalizado en redes sociales representan sistemas interactivos inteligentes.

Data Analytics

Sugerir el mejor punto de encuentro entre el conductor y el pasajero.

Fuente:

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Data Scientist

CHARACTERISTICS OF A DATA WHISPERER

Data scientists aren't born—they're made. IT pros from all backgrounds are working to gain the types of skills companies need as the demand for data scientists outpaces the supply of qualified candidates. These are some common personality traits and skills of a data scientist.



Personality traits:

Intellectual curiosity combined with skepticism and good intuition. A tireless problem-solver driven to find a needle in a haystack. Creativity to guide further investigation with the goal of uncovering new information.



Interpersonal skills:

A storyteller who knows how to present data insights to drive business value and who can communicate with people at all levels of an organization.



Business skills:

Data scientists need knowledge far beyond data analysis and statistics. They need the business savvy to discover patterns that can be used to identify risks and opportunities and the leadership skills to influence business leaders to make data-driven decisions.



Education:

Bachelor's degree in statistics, data science, computer science or mathematics.



Specialized skills:

Data mining, machine learning and distributed computing. Ability to integrate structured and unstructured data. Experience with statistical research techniques, including modeling, data mining, clustering and segmentation.



Tools of the trade:

Familiarity with Hadoop, Pig, Hive, Spark and MapReduce. Comfortable with SQL, Python, Perl or other scripting languages, as well as statistical computing languages such as R.



CONTENT: BRIGIT BOHLEN, EDITORIAL DIRECTOR, BUSINESS APPS & INFO MANAGEMENT/SALES, JESSICA LUNA ADAMS, PHOTO AND GRAPHIC DESIGN

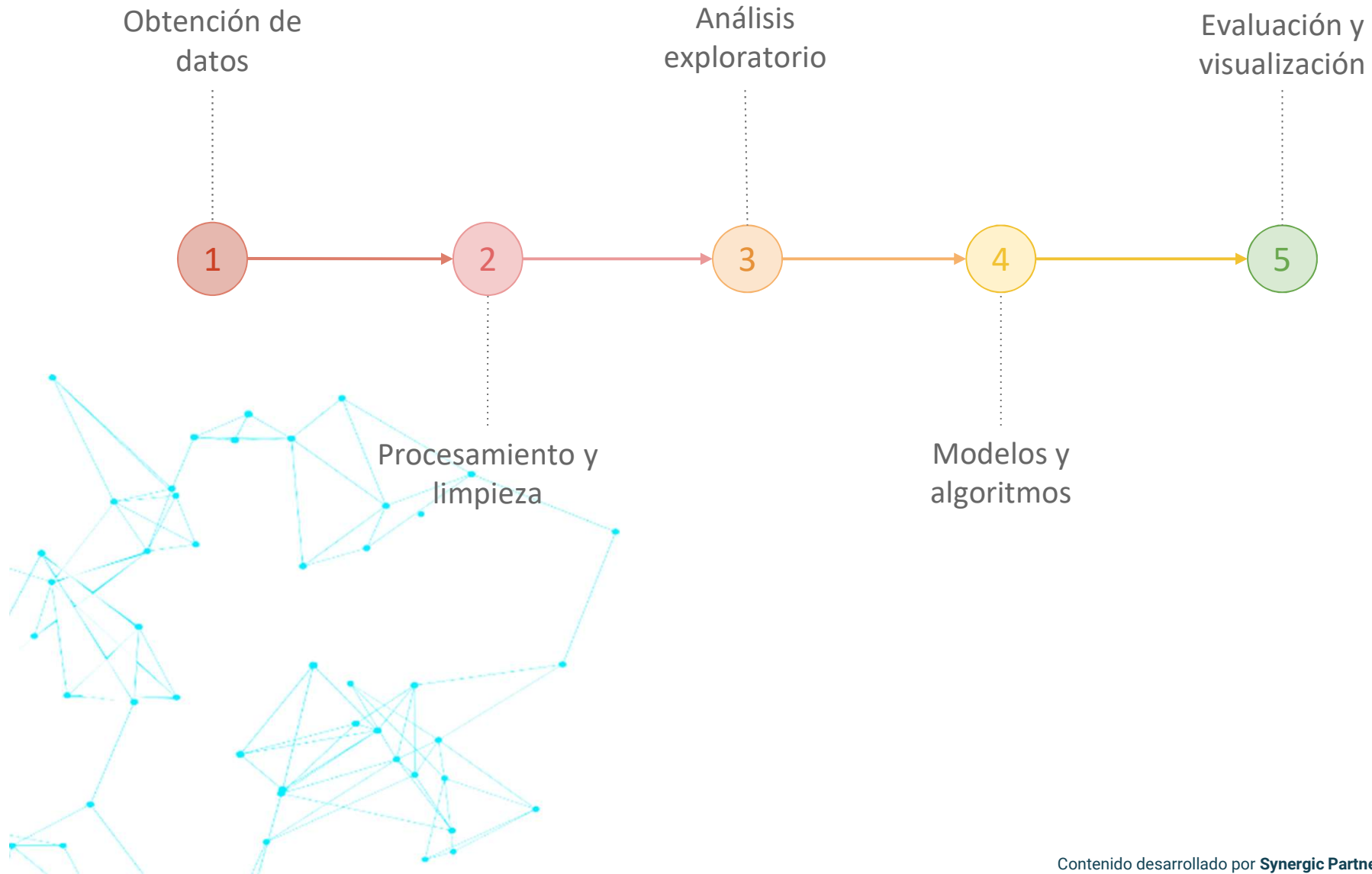
PHOTOGRAPH BY: BOHLEN/ADAMS

An abstract network diagram composed of teal-colored dots (nodes) and thin teal lines (edges). The nodes are scattered across the left side of the slide, with some forming small, dense clusters and others being isolated. The lines connect various nodes, creating a web-like structure that suggests a complex system or process.

Etapas de un proyecto analítico

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

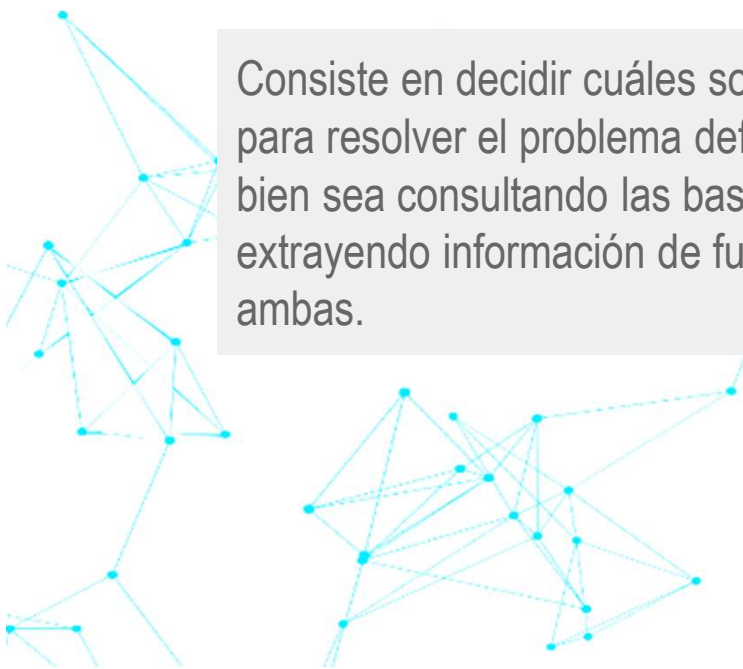
Etapas del proceso analítico



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Etapas del proceso analítico

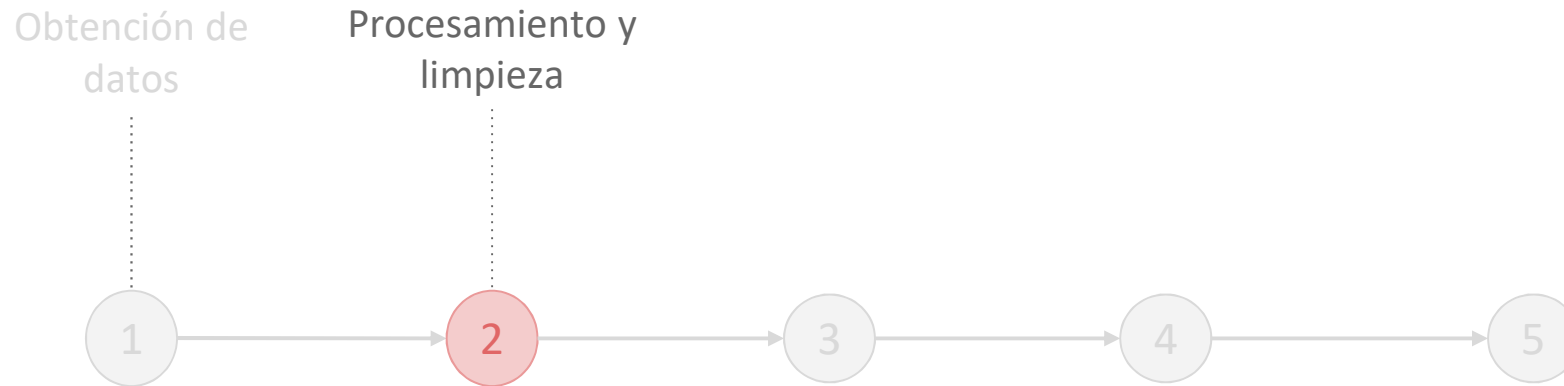
Obtención de
datos



Consiste en decidir cuáles son los datos necesarios para resolver el problema definido y recopilarlos, bien sea consultando las bases de datos internas, extrayendo información de fuentes externas, o ambas.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

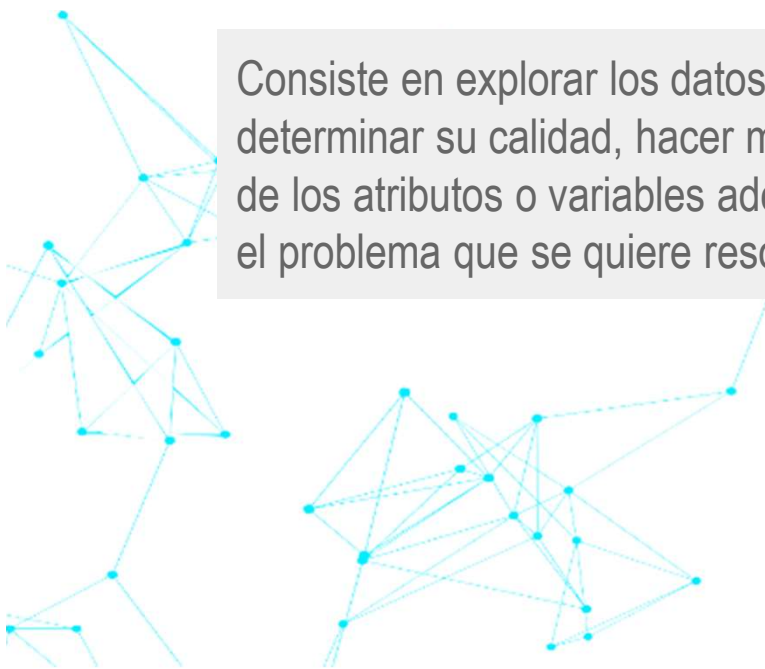
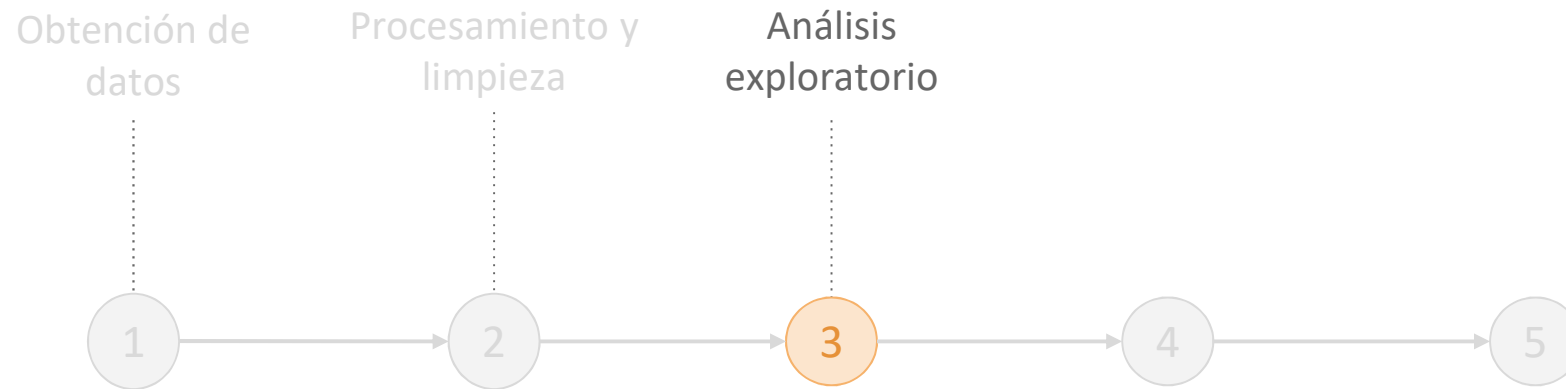
Etapas del proceso analítico



Una vez obtenidos los datos se deben utilizar procesos de ETL para organizar, limpiar, y dar formato, para finalmente hacer la ingesta de los datos.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

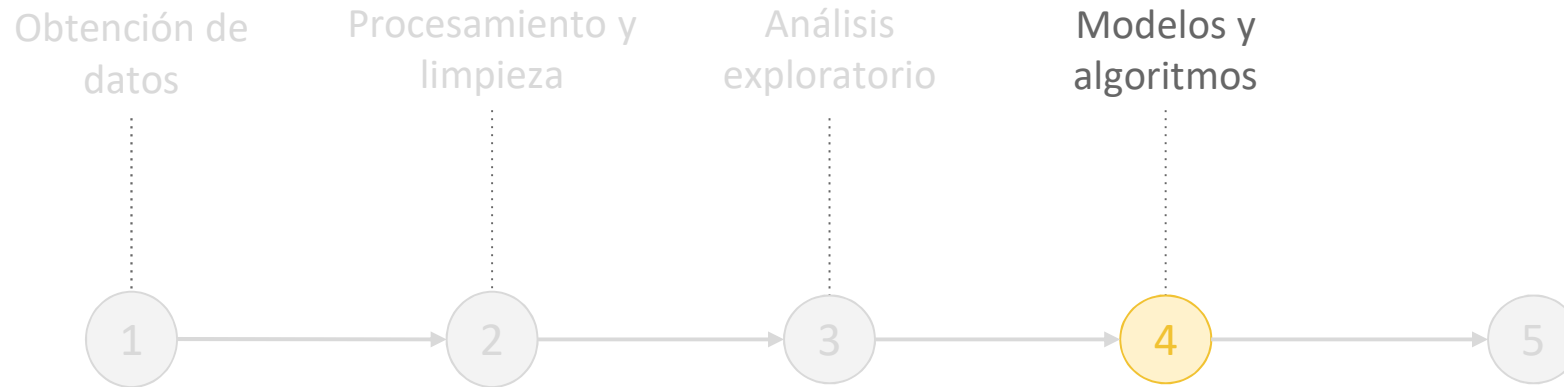
Etapas del proceso analítico



Consiste en explorar los datos recolectados, determinar su calidad, hacer muestreo, y selección de los atributos o variables adecuadas para resolver el problema que se quiere resolver.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

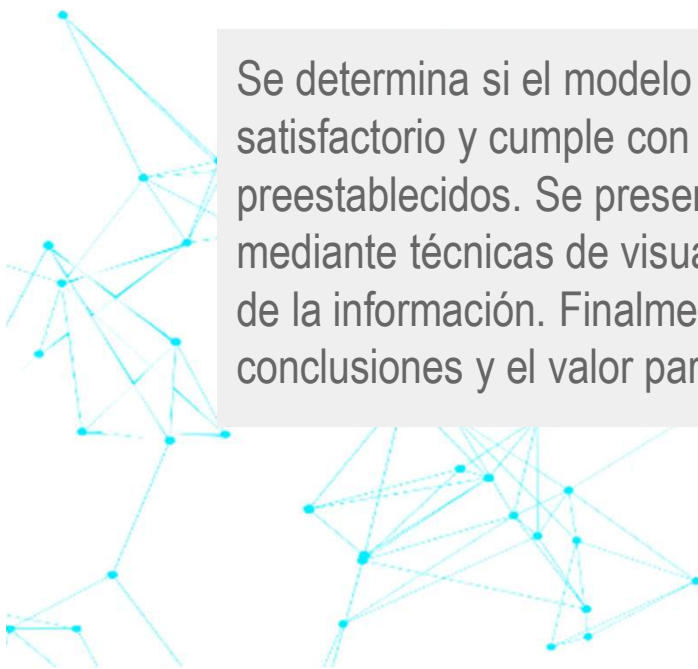
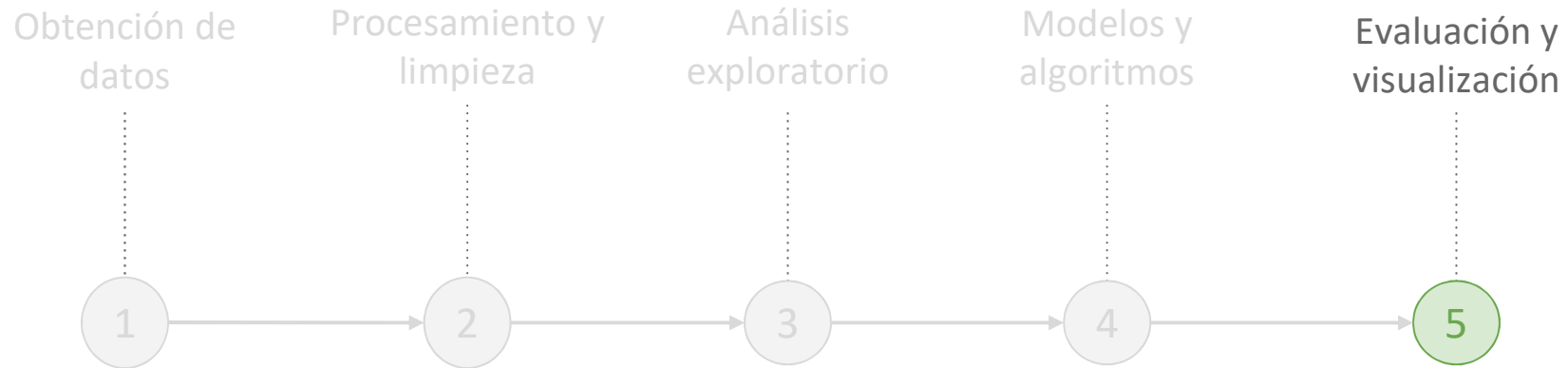
Etapas del proceso analítico



En esta etapa se utilizan distintas técnicas de análisis y algoritmos de machine learning para ajustar modelos a partir de un conjunto de datos, mediante un proceso de aprendizaje que nos permite descubrir patrones, hacer predicciones, o describir los datos.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

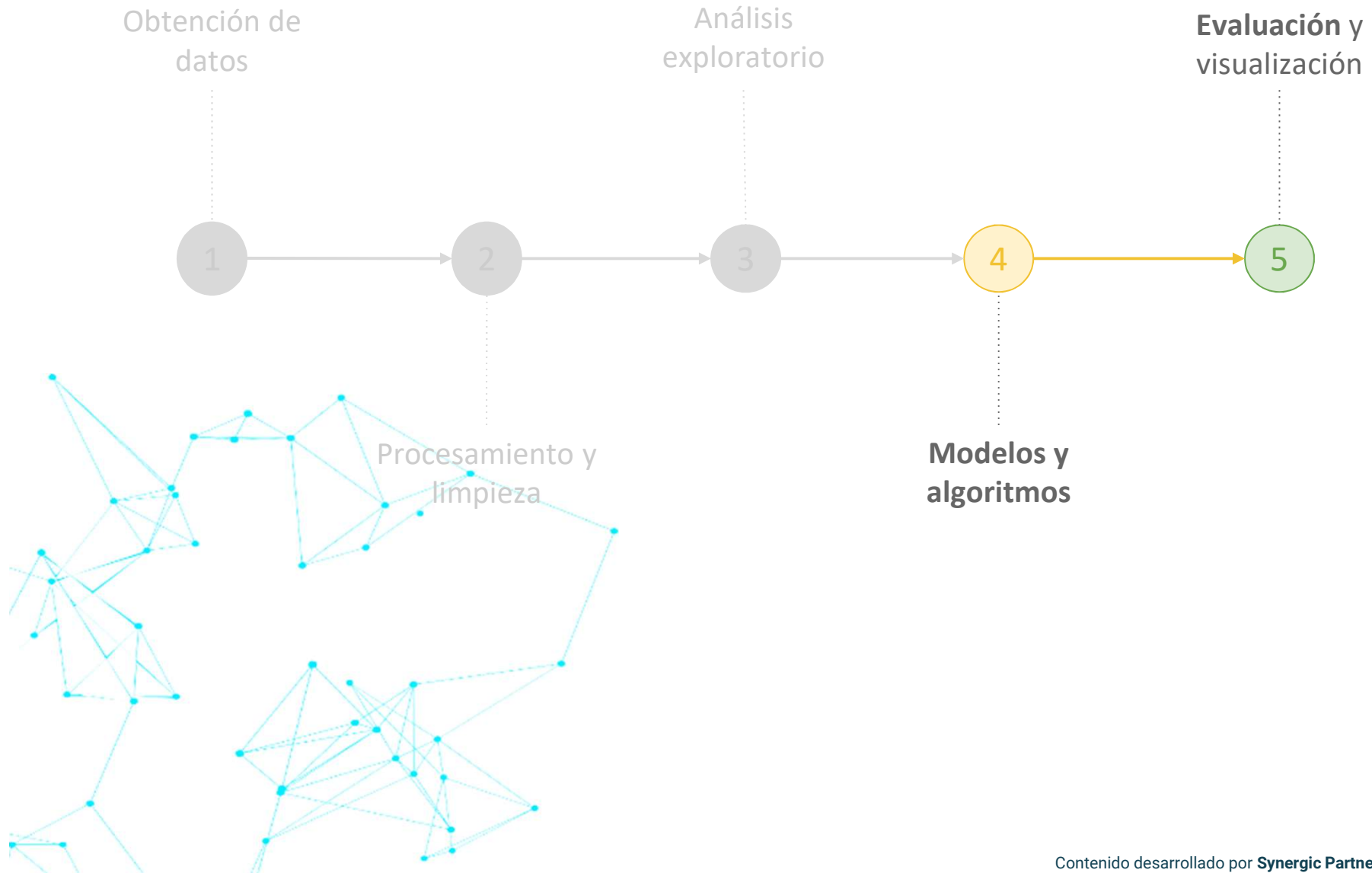
Etapas del proceso analítico



Se determina si el modelo obtenido ha sido satisfactorio y cumple con los objetivos preestablecidos. Se presentan los resultados mediante técnicas de visualización y representación de la información. Finalmente se extraen conclusiones y el valor para el negocio.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Etapas del proceso analítico



An abstract network diagram consisting of numerous teal-colored nodes connected by thin teal lines. The nodes are scattered across the left side of the slide, with some forming small clusters and others standing alone. The lines represent connections between these nodes, creating a complex web-like structure.

¿Qué es Machine Learning?

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

¿Qué es machine learning?

“

Machine Learning es una disciplina científica cuyas técnicas permiten a los ordenadores **aprender de forma automática** a partir de un conjunto de datos, de tal forma que seamos capaces de hacer predicciones sobre un proceso o describirlo de forma compacta.

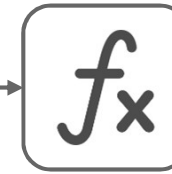
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

¿En qué consiste el aprendizaje?

Dataset de
entrenamiento



Algoritmo de
aprendizaje



Hipótesis
o Modelo

En los sistemas de *machine learning* se dispone un conjunto de datos de entrenamiento o *training dataset*, que son utilizados para obtener una *hipótesis* (modelo o función) mediante un proceso de aprendizaje.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

¿En qué consiste el aprendizaje?

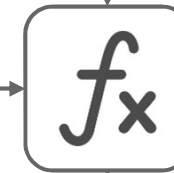
Dataset de
entrenamiento



Algoritmo de
aprendizaje



Nuevos datos



Hipótesis
o Modelo



Salida/
Predicción

Esta hipótesis es a priori la que mejor se ajusta a los datos de entrenamiento y que permite **generalizar la salida** ante **nuevos datos** de entrada.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

¿En qué consiste el aprendizaje?

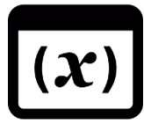


Datos

En general, los datos utilizados para entrenar un modelo de *machine learning* se organizan de forma tabular.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Terminología



Atributo

Es cualquier aspecto distintivo o característica medible de un sistema o entidad.



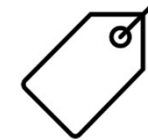
Instancia

Entidad sobre lo que se quiere hacer la predicción. Consiste en una colección de valores que representa a una observación, generalmente mediante un vector de atributos.



Respuesta

Es la variable de salida u objetivo a predecir.



Etiqueta

Es el valor real de salida en una tarea de predicción.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Terminología



Ejemplo

Es una instancia (conjunto de atributos) y la etiqueta correspondiente.



Modelo ó hipótesis

Es una estructura que resume un conjunto de datos para predicción o descripción.



Métrica

Una medida de interés que cuantifica el desempeño del modelo.



Objetivo

Es una métrica que optimiza el algoritmo de aprendizaje.

Terminología

(x) Atributo	(x) Atributo	$[Y]$ Respuesta
Minutos llamadas	Monto factura	Baja
135.0	18.8	true
26	57.79	true
545	16.73	false
...
...

Etiqueta

Ejemplo

Instancia

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Tenemos acceso a datos de ventas del mercado inmobiliario de la ciudad de *Nueva York*, con las siguientes variables:

- Nombre del barrio
- Categoría catastral del edificio
- Categoría fiscal
- Dirección
- Código Postal
- Superficie
- Año de construcción
- Precio de venta
- Fecha de venta

Con estos datos queremos construir un **modelo** predictivo que nos permita **estimar** el **precio de venta** de la propiedad.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Atributos

- Nombre del barrio
- Categoría catastral del edificio
- Categoría fiscal
- Dirección
- Código Postal
- Superficie
- Año de construcción
- Precio de venta
- Fecha de venta

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Instancia

- Nombre del barrio → *Chinatown*
- Categoría catastral del edificio → *Unifamiliar*
- Categoría fiscal → *Clase 1 (residencial)*
- Dirección → *53-55 Division Street*
- Código Postal → *10002*
- Superficie → *1708 ft²*
- Año de construcción → *1920*
- Precio de venta
- Fecha de venta → *27/06/2013*

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Salida

- Nombre del barrio
- Categoría catastral del edificio
- Categoría fiscal
- Dirección
- Código Postal
- Superficie
- Año de construcción
- Precio de venta
- Fecha de venta

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Etiqueta

- Nombre del barrio
- Categoría catastral del edificio
- Categoría fiscal
- Dirección
- Código Postal
- Metros cuadrados
- Año de construcción
- Precio de venta → \$2.800.000
- Fecha de venta

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Ejemplo

- Nombre del barrio → *Chinatown*
- Categoría catastral del edificio → *Unifamiliar*
- Categoría fiscal → *Clase 1 (residencial)*
- Dirección → *53-55 Division Street*
- Código Postal → *10002*
- Superficie → *1708*
- Año de construcción → *1920*
- Precio de venta → *\$2.800.000*
- Fecha de venta → *27/06/2013*

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ejemplo. Mercado inmobiliario



Dataset

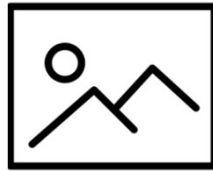
Nombre del barrio	Categoría catastral	...	Fecha de venta	Precio de venta
Chinatown	Unifamiliar	...	27/06/2013	2800000
.
.
.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

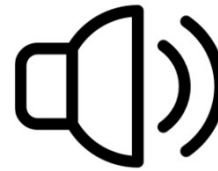
Tipos de datos

AaI

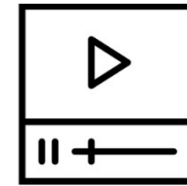
Texto. Tipo de datos más simple, se compone de cadenas de caracteres (números y letras).



Imágen. Contiene información bidimensional, además puede contener información del color (RGB, HSV). Suele requerir pre-procesamiento para poder ser usado en sistemas de aprendizaje automático.



Sonido. Una señal de sonido puede contener varios canales por lo que su complejidad a la hora de ser usada en machine learning también es considerable, requiere preprocesamiento.



Video. Consiste en un conjunto de imágenes (frames) consecutivas a una determinada tasa de imágenes por segundo, introduce una nueva dimensión, el tiempo.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de variables

Variables Cuantitativas

Son variables numéricas que se pueden ordenar y medir. Las variables cuantitativas pueden ser:

- Continuas (en intervalos)
- Discretas (valores concretos)

Superficie (m ²)	232.26, 993.78, 469.62
Precio de venta (\$)	5380000, 53183682, 41250000
Año de construcción	1905, 1920, 1991, ...

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de variables

Variables Categóricas

También se denominan variables cualitativas o variables de atributos. Los valores de una variable categórica se pueden colocar en un número contable de categorías o grupos diferentes. Los datos categóricos pueden tener o no algún orden lógico.

Categoría Fiscal	1: Residencial 2: Residencial Coop 3:
Código Postal	10038, 10013, 10002, ...
Barrio	Alphabet City, Chinatown, Chelsea, ...

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de variables

Variables Indicadoras (*Dummies*)

Las variables dummy son variables cualitativas. Sólo pueden asumir los valores 0 y 1, indicando respectivamente ausencia o presencia de una cualidad o atributo.

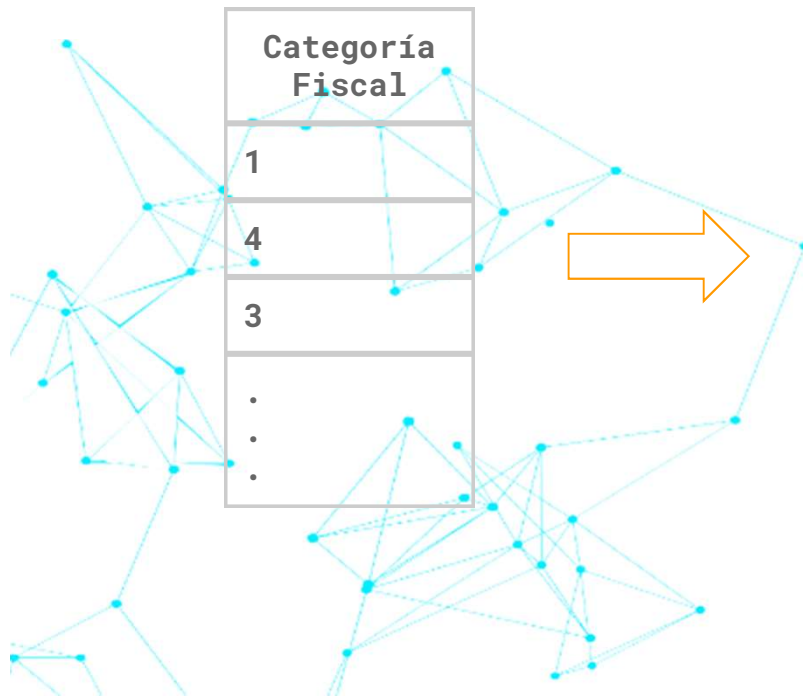
Derecho a usufructo	0: No 1: Si
----------------------------	----------------

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de variables

Variables Indicadoras (*Dummies*)

Es usual generar variables *dummy* a partir de una variable categórica.



Categoría Fiscal 1	Categoría Fiscal 2	Categoría Fiscal 3	Categoría Fiscal 4
1	0	0	0
0	0	0	1
0	0	1	0
⋮	⋮	⋮	⋮

An abstract network diagram composed of numerous teal-colored dots (nodes) connected by thin teal lines (edges). The connections form various geometric shapes, including triangles and polygons, scattered across the left side of the slide. The overall effect is a complex, interconnected web of points and lines.

Tipos de Aprendizaje

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de aprendizaje

Existen diferentes **estrategias de aprendizaje automático**. En general el proceso de aprendizaje consiste en modificar los parámetros del modelo en función de los datos. Se suelen agrupar en:

Aprendizaje
supervisado

Aprendizaje
no
supervisado

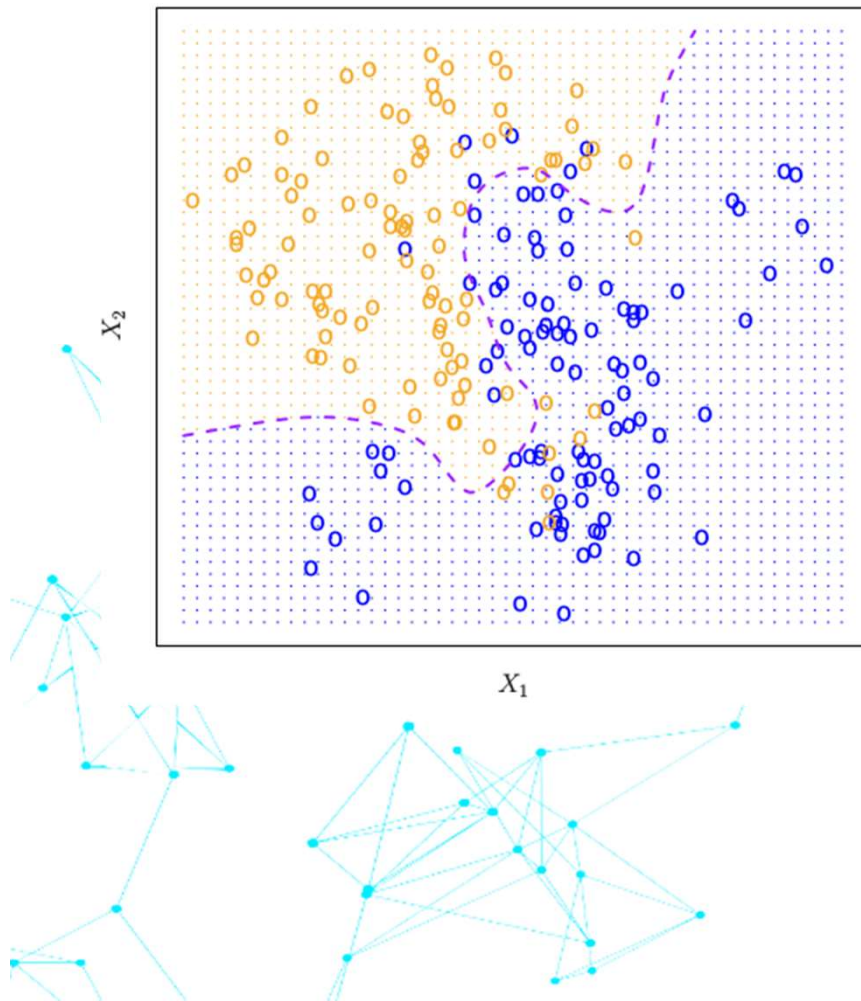
Aprendizaje
por refuerzo

Aprendizaje
semi-
supervisado

Otros...

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de aprendizaje. Aprendizaje supervisado



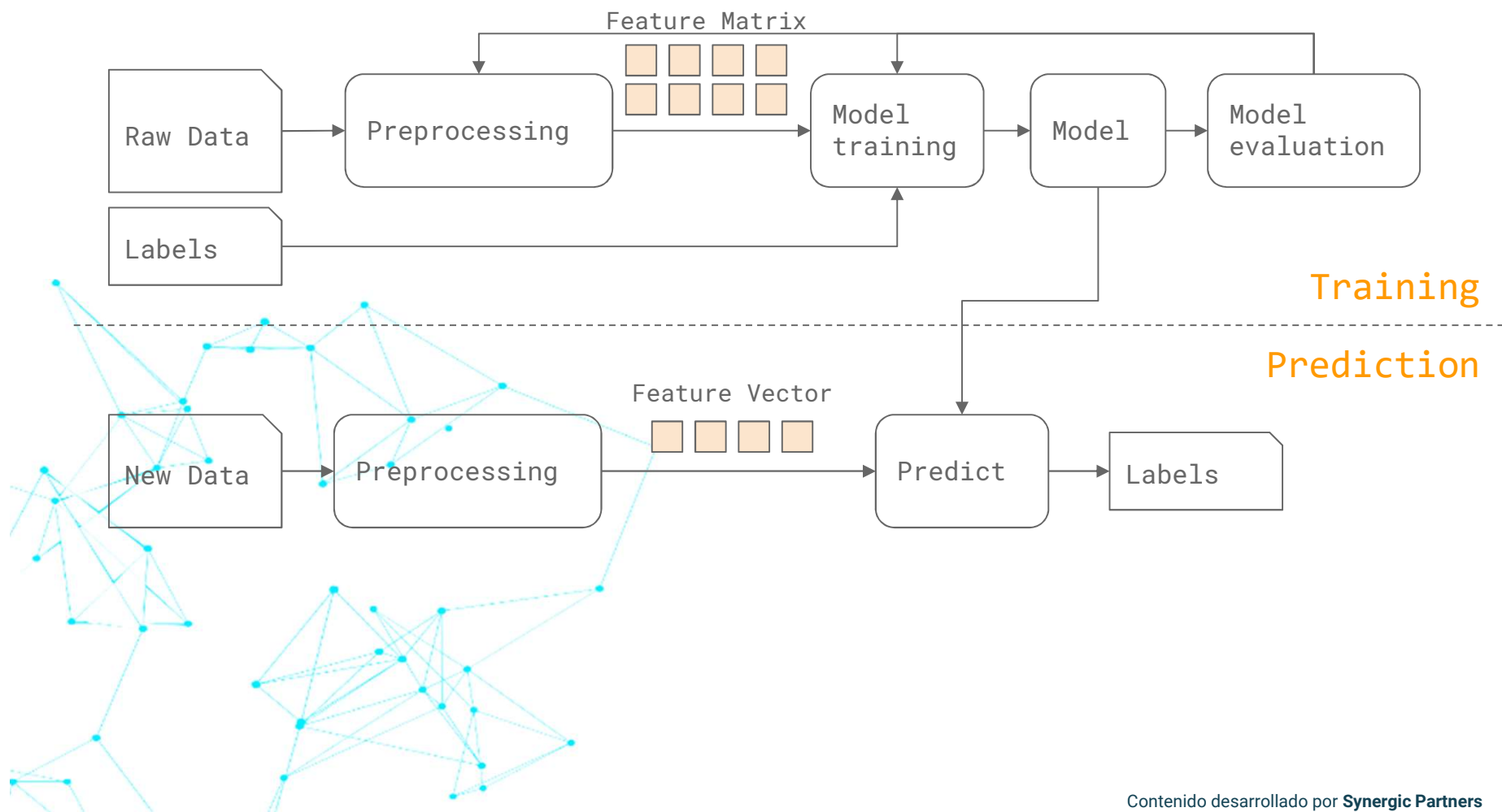
Aprendizaje Supervisado

Consiste en inferir a partir de ejemplos una función que relacione un conjunto de atributos con una variable de respuesta. El objetivo es predecir (generalizar) la respuesta ante futuras observaciones de los atributos.

Ejemplo. Clasificación binaria. El mapa de colores muestra la frontera de decisión aprendida por el algoritmo.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de aprendizaje. Aprendizaje supervisado



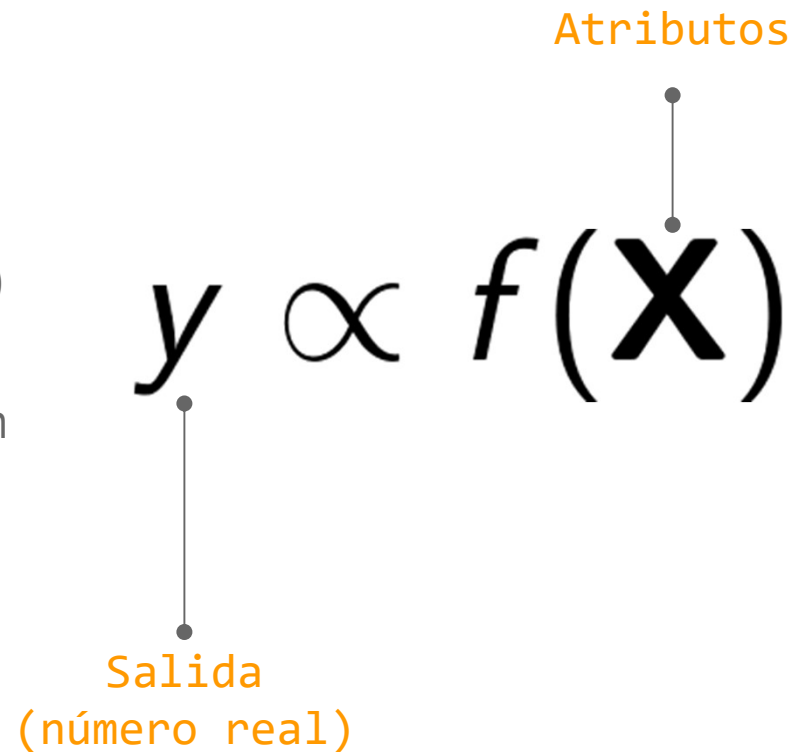
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Aprendizaje supervisado. Regresión

Regresión

Los métodos de regresión estudian la construcción de modelos para explicar la dependencia entre una variable respuesta dependiente (Y) y la(s) variable(s) explicativa(s) o dependiente(s), X .

Es un modelo óptimo para patrones de demanda con tendencia (creciente o decreciente), es decir, patrones que presenten una relación de linealidad entre la demanda y el tiempo.



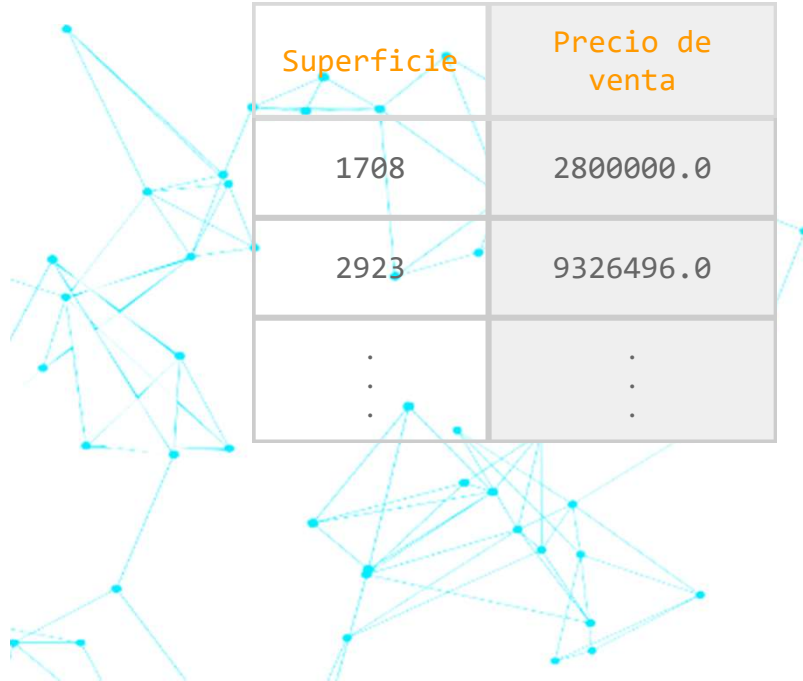
A diagram illustrating the regression equation $y \propto f(X)$. The variable y is labeled "Salida (número real)" (Output (real number)). The variable X is labeled "Atributos" (Attributes). The function f is represented by the symbol \propto .

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

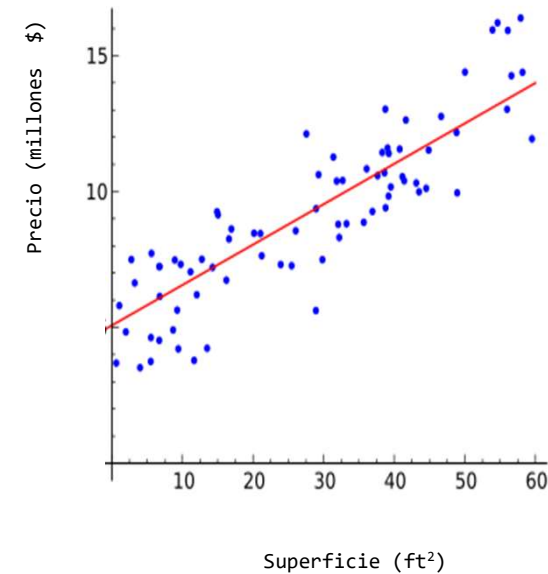
Aprendizaje supervisado. Regresión



Ejemplo. Predecir precio de una vivienda en función de los ft^2 .

A decorative network graph with blue nodes and lines, partially overlapping the table.

Superficie	Precio de venta
1708	2800000.0
2923	9326496.0
⋮	⋮



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Aprendizaje supervisado. Clasificación

Clasificación

Un sistema de clasificación predice una categoría o etiqueta; es decir, asigna una clase a un objeto.

La categoría puede ser binaria (dos valores, por ejemplo: churn/no churn, residencial/no residencial) o un rango de valores (A, B, C, D, ..).



$$y \propto f(\mathbf{X})$$

Atributos

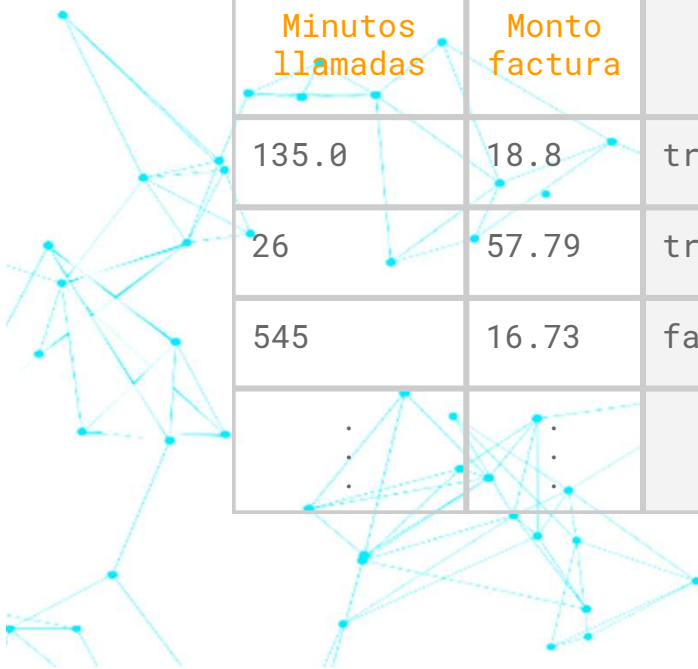
Salida
(categoría o clase)

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

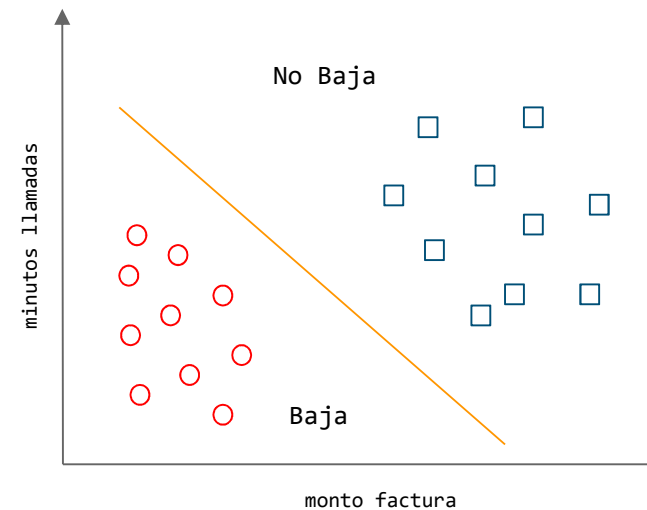
Aprendizaje supervisado. Clasificación



Ejemplo. Predicción de bajas en un servicio de telefonía. El objetivo del modelo es clasificar a los clientes actuales como “baja” o “no baja” del servicio en función de una probabilidad (scoring) calculada a partir de un modelo desarrollado con el histórico de bajas y no bajas de clientes.

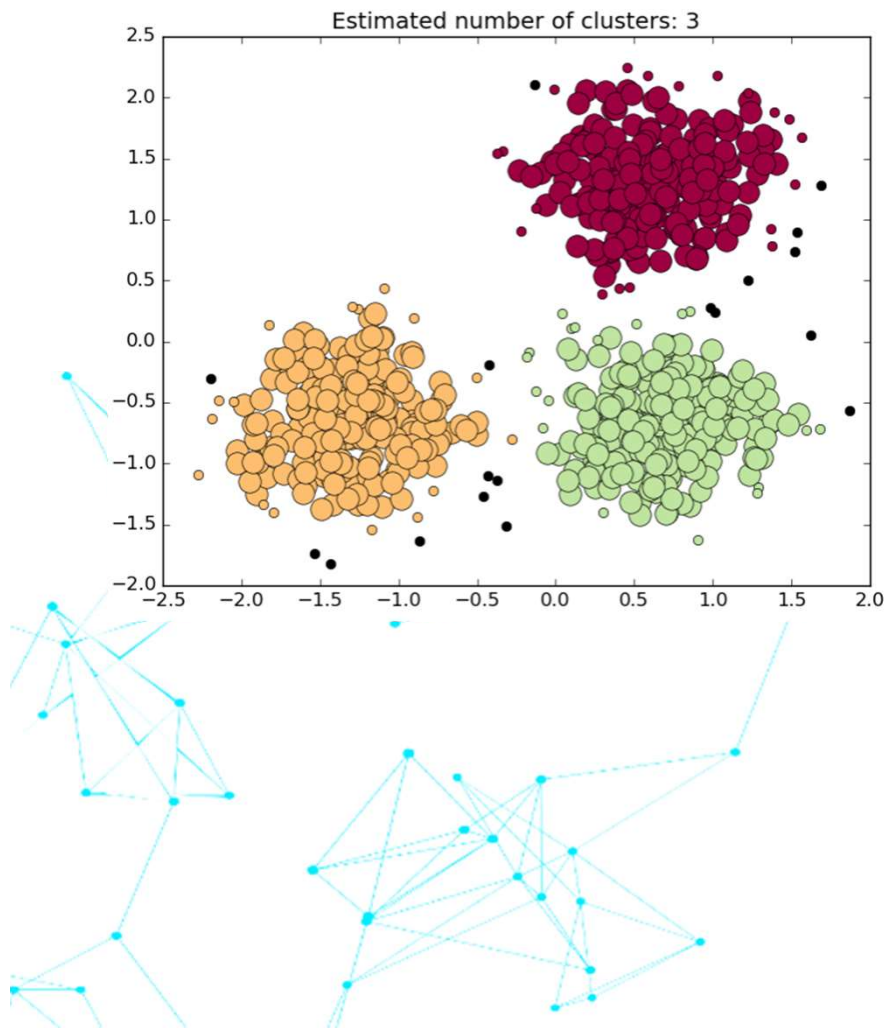


Minutos llamadas	Monto factura	Baja
135.0	18.8	true
26	57.79	true
545	16.73	false
⋮	⋮	⋮



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de aprendizaje. Aprendizaje no supervisado



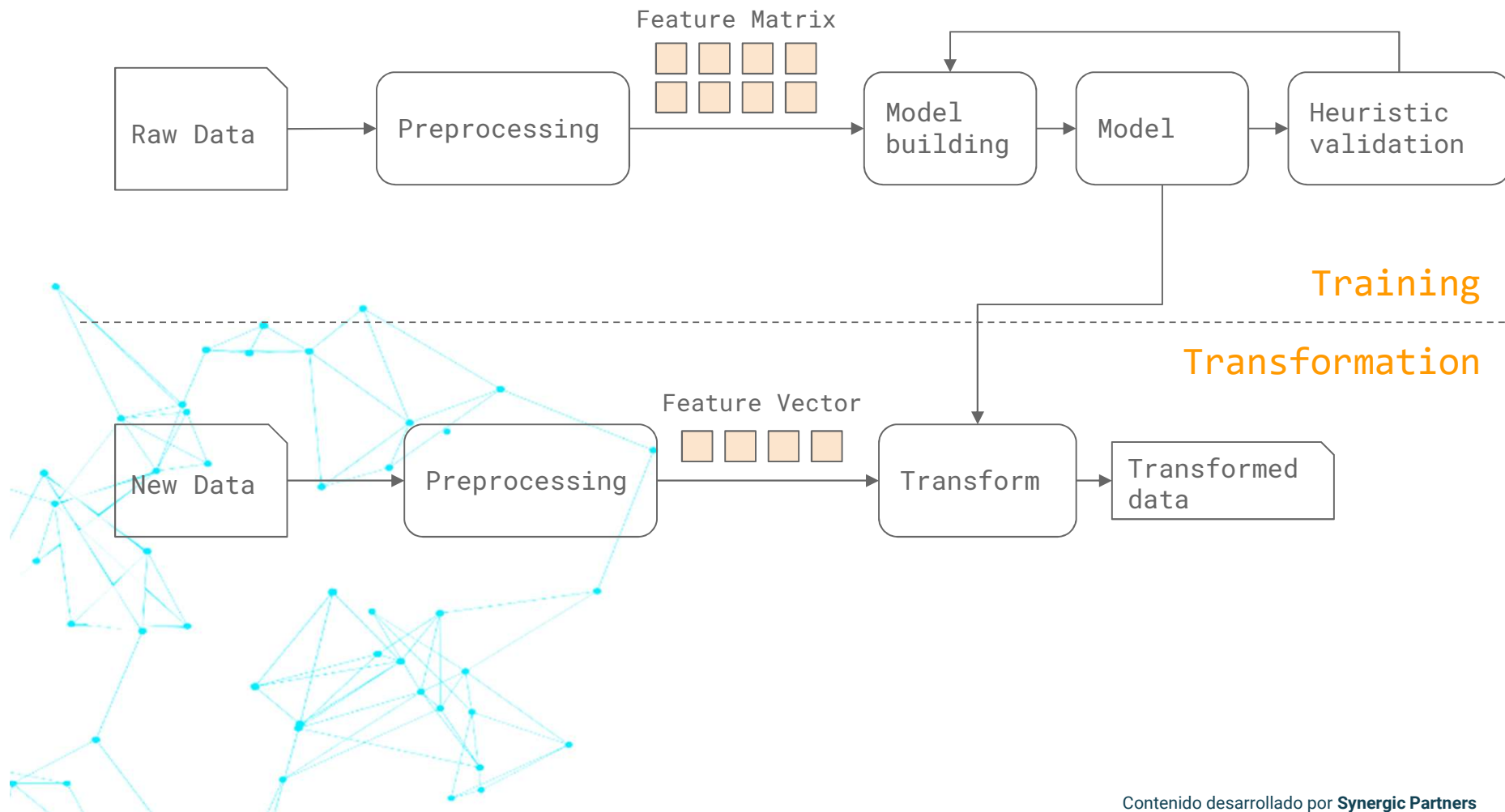
Aprendizaje no Supervisado

Son algoritmos que permiten describir cómo están organizados o agrupados un conjunto de datos sin etiquetas. El método de aprendizaje no supervisado más común es el clustering, que permite agrupar o segmentar datos similares en sus características.

Ejemplo. Agrupamiento basado en densidad (DBSCAN). *Fuente: documentación scikit-learn.*

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de aprendizaje. Aprendizaje no supervisado



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

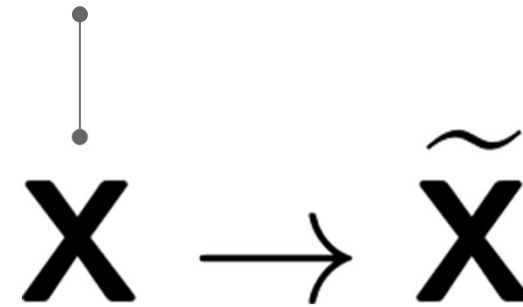
Aprendizaje no supervisado. Clustering

Clustering

Consiste en organizar objetos en grupos según su distribución o similitud.

Es el proceso de agrupar los datos en grupos o en clústeres, de tal forma que, el conjunto de datos o instancias de un mismo clúster presentan una alta similitud entre ellos y a su vez, son muy diferentes de los de otro clúster.

Atributos



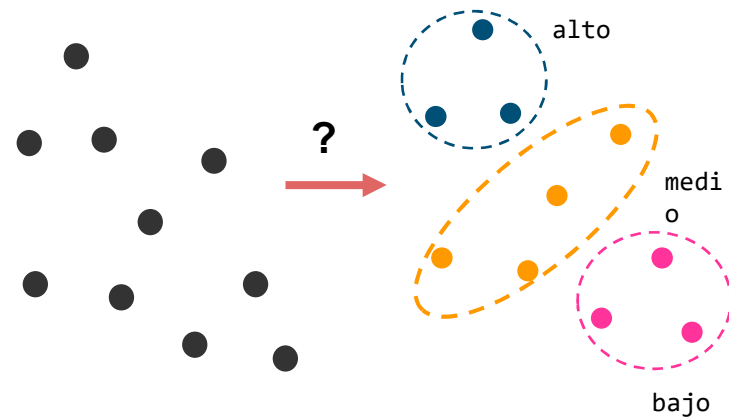
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Aprendizaje no supervisado. Clustering



Ejemplo. Organizar clientes por consumo de datos

Llamadas nacionales (min.)	Llamadas internacionales (min.)	Monto factura
135.0	0	18.8
26	379	57.79
545	10	16.73
⋮	⋮	⋮



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Tipos de problemas a abordar con Machine Learning

