Minería de datos

Flavio E. Spetale

2024

Equipo docente: Vasquez, Facundo
Spetale, Flavio E.

Horarios de la Materia: Lunes 20:00 a 22Hs – Virtual
Miércoles 20:00 a 22Hs - Presencial

Previas: Fundamentos de Ciencia de Datos

Clases Teóricas-Practicas en Python

Promoción de la materia

01

Trabajos Prácticos

Se deben tener los 3 TP de la materia aprobados

02

Examen Globalizador

Se deben tener aprobado, nota **mayor a 6**, el examen

03

Evaluación continua

Se evaluará en forma continua el desarrollo de los estudiantes durante las clases

Bibliografía

01 **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**
Aurélien Géron. O'Reilly Media, Inc. 3er Edition. 2023

02 **Tree-Based Methods for Statistical Learning in R**
Brandon M. Greenwell. CRC Press. 1er Edition. 2022

03 **An Introduction to Statistical. Learning, With Applications in Python**
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor. Springer. 2023

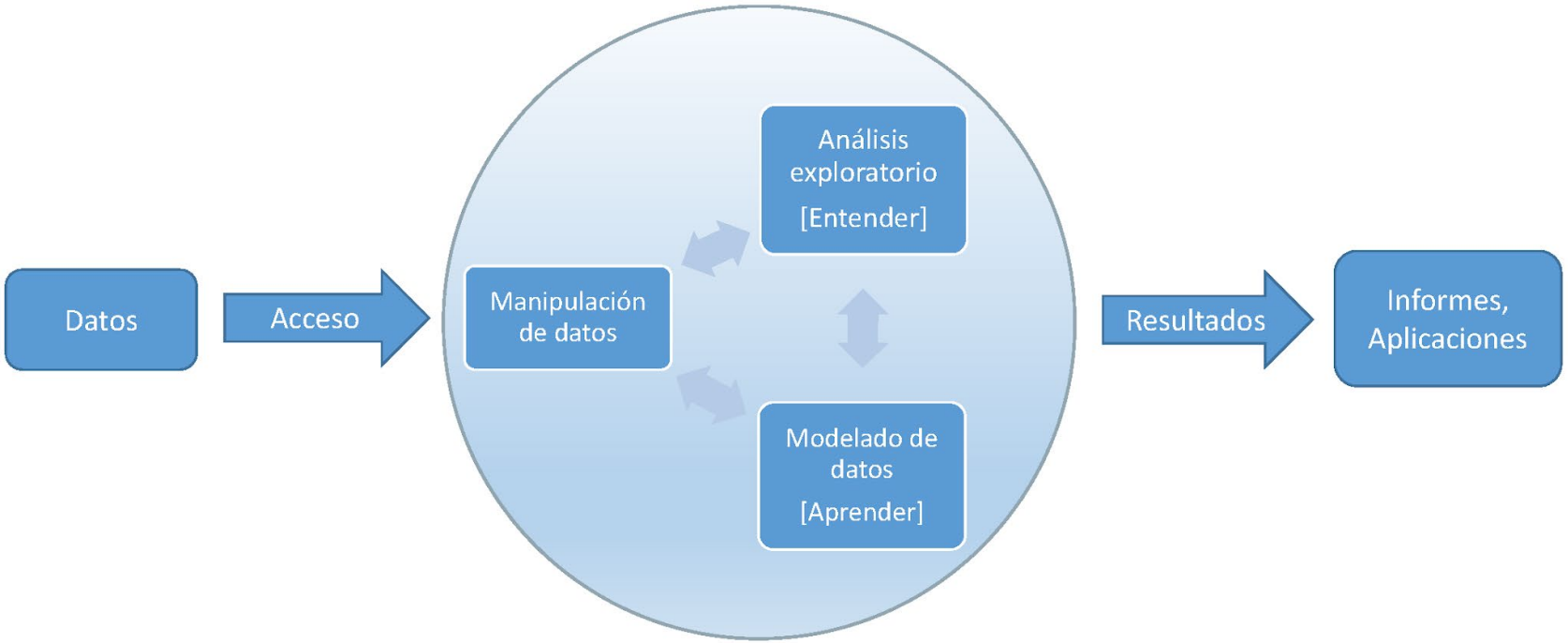
04 **Data Mining and Machine Learning: Fundamental Concepts and Algorithms**
Mohammed J. Zaki, Wagner Meira, Jr. Cambridge University Press. 2do Edition. 2020

Minería de datos

- El arte y la ciencia del **análisis inteligente de los datos**.
- El conjunto de herramientas para **entender** y **modelizar** conjuntos (complejos) de datos.
- El proceso de **descubrir patrones** y obtener conocimiento a partir de grandes conjuntos de datos.

Minería de datos

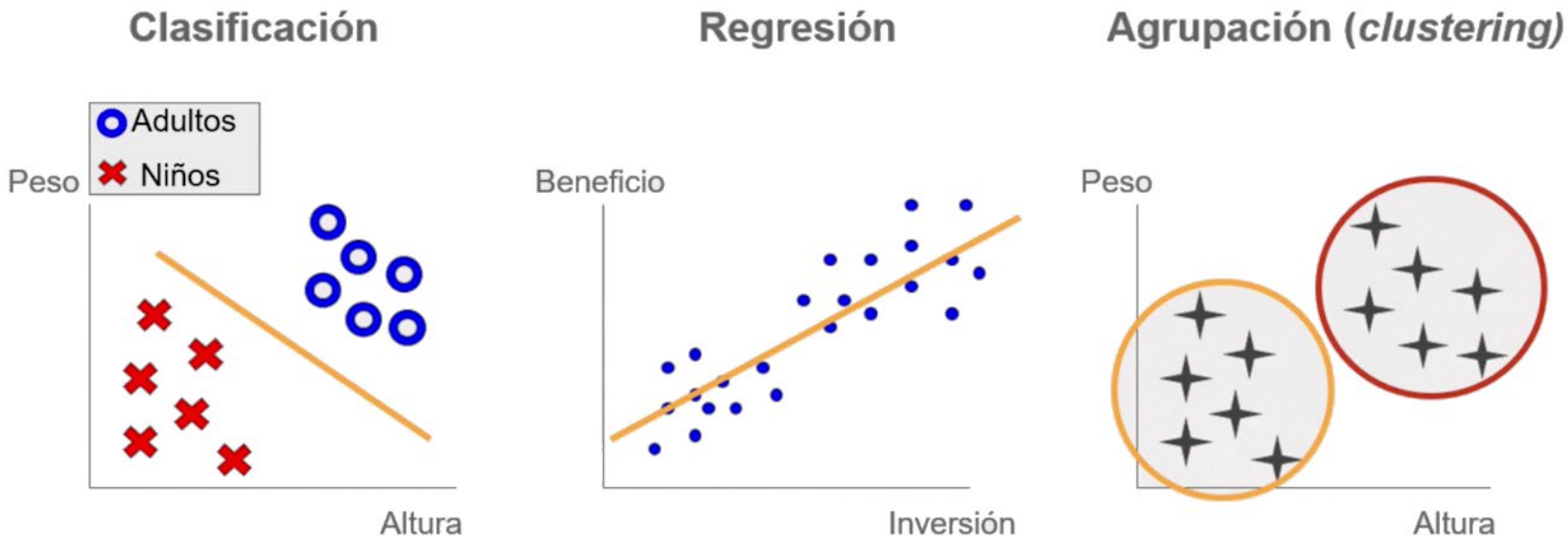
Conjunto de conocimientos y herramientas utilizados en las distintas etapas del análisis de datos



Minería de datos

La clasificación, la regresión y el agrupamiento se basan en el paradigma del aprendizaje inductivo.

La esencia de cada una de ellas es derivar inductivamente a partir de los datos (que representan la información del entrenamiento) un modelo (que representa el conocimiento) que tiene utilidad predictiva, es decir, que puede aplicarse a nuevos datos.



Clasificación

Es uno de los procesos cognitivos importantes, tanto en la vida cotidiana como en los negocios.

La tarea de clasificación consiste en asignar instancias de un dominio dado, descritas por un conjunto de atributos discretos o de valor continuo, a un conjunto de clases, que pueden ser consideradas valores de un atributo discreto seleccionado, generalmente denominado clase.

La función de clasificación puede verse como:

$$c : X \rightarrow C$$

donde ***c*** representa la función de clasificación, ***X*** el conjunto de atributos que forman una instancia y ***C*** la etiqueta de clase de dicha instancia.

Un tipo de clasificación particularmente simple, hace referencia a los problemas de clasificación binarios, es decir, problemas con un conjunto de datos pertenecientes a dos clases, por ejemplo, C = {0, 1}.

Regresión

Es una tarea de aprendizaje inductivo para predecir valores numéricos en lugar de etiquetas de clase discretas.

La tarea de regresión consiste en asignar valores numéricos a instancias de un dominio dado, descritos por un conjunto de atributos discretos o de valor continuo.

La función de regresión se puede definir como:

$$f : X \rightarrow R$$

donde ***f*** representa la función de regresión, ***X*** el conjunto de atributos que forman una instancia y ***R*** un valor en el dominio de los números reales.

Agrupamiento

El agrupamiento (clustering) es una tarea de aprendizaje inductiva que, a diferencia de las tareas de clasificación y regresión, no dispone de una etiqueta de clase a predecir.

Puede considerarse como un problema de clasificación, pero donde no existen un conjunto de clases predefinidas, y estas se «descubren» de forma autónoma por el método o algoritmo de agrupamiento, basándose en patrones de similitud identificados en los datos.

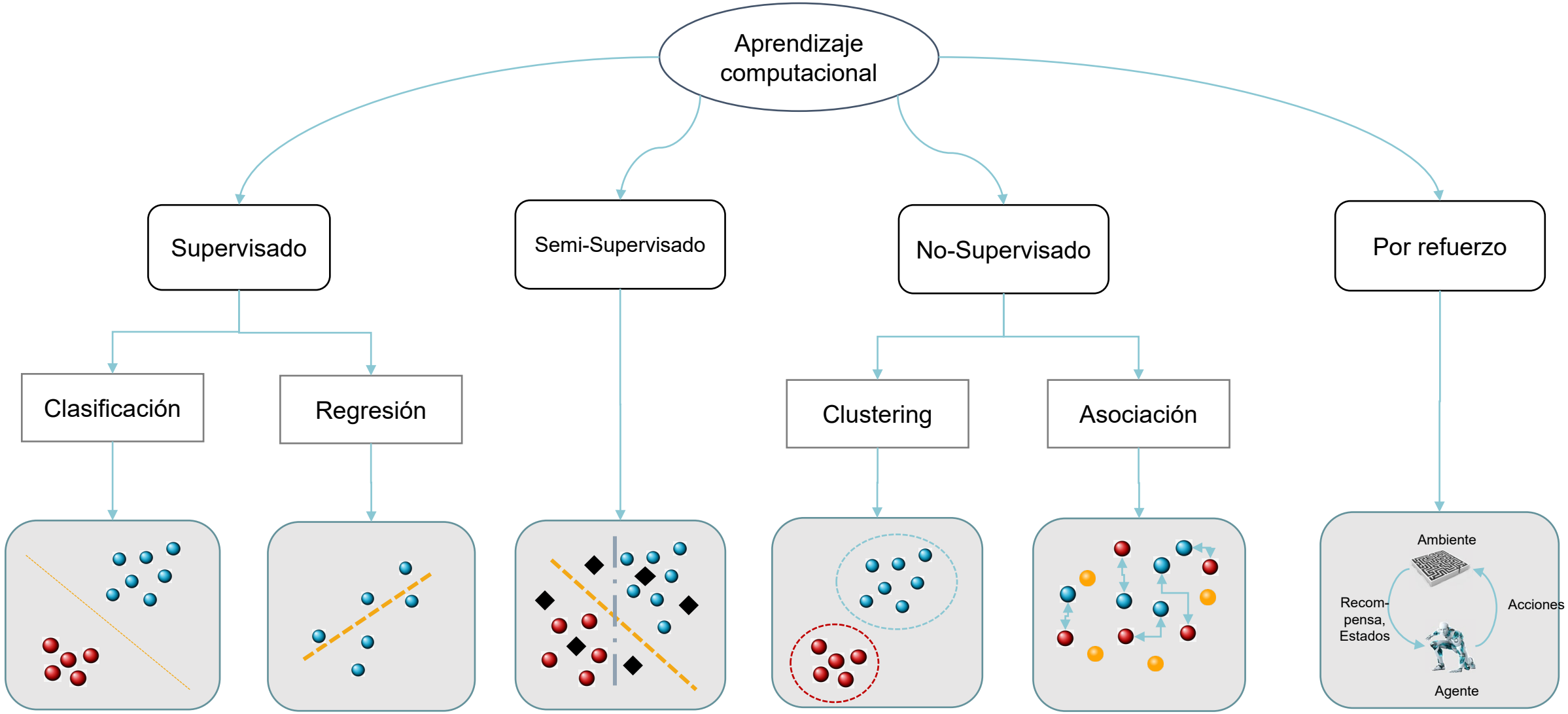
La tarea de agrupamiento consiste en dividir un conjunto de instancias de un dominio dado, descrito por un número de atributos discretos o de valor continuo, en un conjunto de grupos (clústeres) basándose en la similitud entre las instancias, y crear un modelo que puede asignar nuevas instancias a uno de estos grupos.

La función de agrupamiento se puede modelar mediante:

$$h : X \rightarrow Ch$$

donde ***h*** representa la función de agrupamiento, ***X*** el conjunto de atributos que forman una instancia y ***Ch*** un conjunto de grupos.

Tipología de algoritmos



Tipología de algoritmos

Métodos	Supervisado		No supervisado
	Clasificación	Regresión	Agrupamiento
Agrupamiento jerárquico			**X**
k-means y derivados			**X**
k-NN	**X**		
SVM	**X**	**X**	
Redes neuronales	**X**	**X**	
Árboles de decisión	**X**	**X**	
Métodos probabilísticos	**X**	**X**	

Aprendizaje computacional - Aplicaciones

- Reconocimiento y procesamiento de lenguaje natural, oral y escrito.
- Bots de soporte.
- Visión por computadora.
- Reconocimiento de imágenes.
- Seguridad de datos y personal.
- Comercio.
- Detección de fraudes.
- Fidelización de clientes.
- Marketing y recomendaciones personalizadas.
- Búsqueda en línea.
- Control automático.
- Conducción automática de vehículos.
- Antivirus y detección de intrusos.
- Ayuda a disciplinas científicas (genómica, biología, ingeniería, física, geografía, etc).

Preparación de los datos

- Tareas de limpieza de datos, que permiten corregir o eliminar ruido o datos no válidos.
- Tareas de normalización de datos, que facilita la presentación de los datos en el mismo rango.
- Tareas de discretización, entendidas como procesos de conversión de variables continuas a categóricas.
- Tareas de reducción de la dimensionalidad, que nos ayudará a desarrollar modelos con juegos de datos reducidos.

Preparación de los datos

Limpieza de datos

En el proceso de limpieza de datos se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos.

A nivel de valores de atributos se gestionan los valores ausentes, los erróneos y los inconsistentes.

Un ejemplo podrían ser los valores fuera de rango (outliers).

El proceso de integración de datos puede ser una de las principales fuentes de incoherencias en los datos.

Preparación de los datos

Normalización de datos

La normalización de datos consiste en modificar los datos para lograr que estén en una escala de valores equivalentes que simplifique la comparación entre ellos.

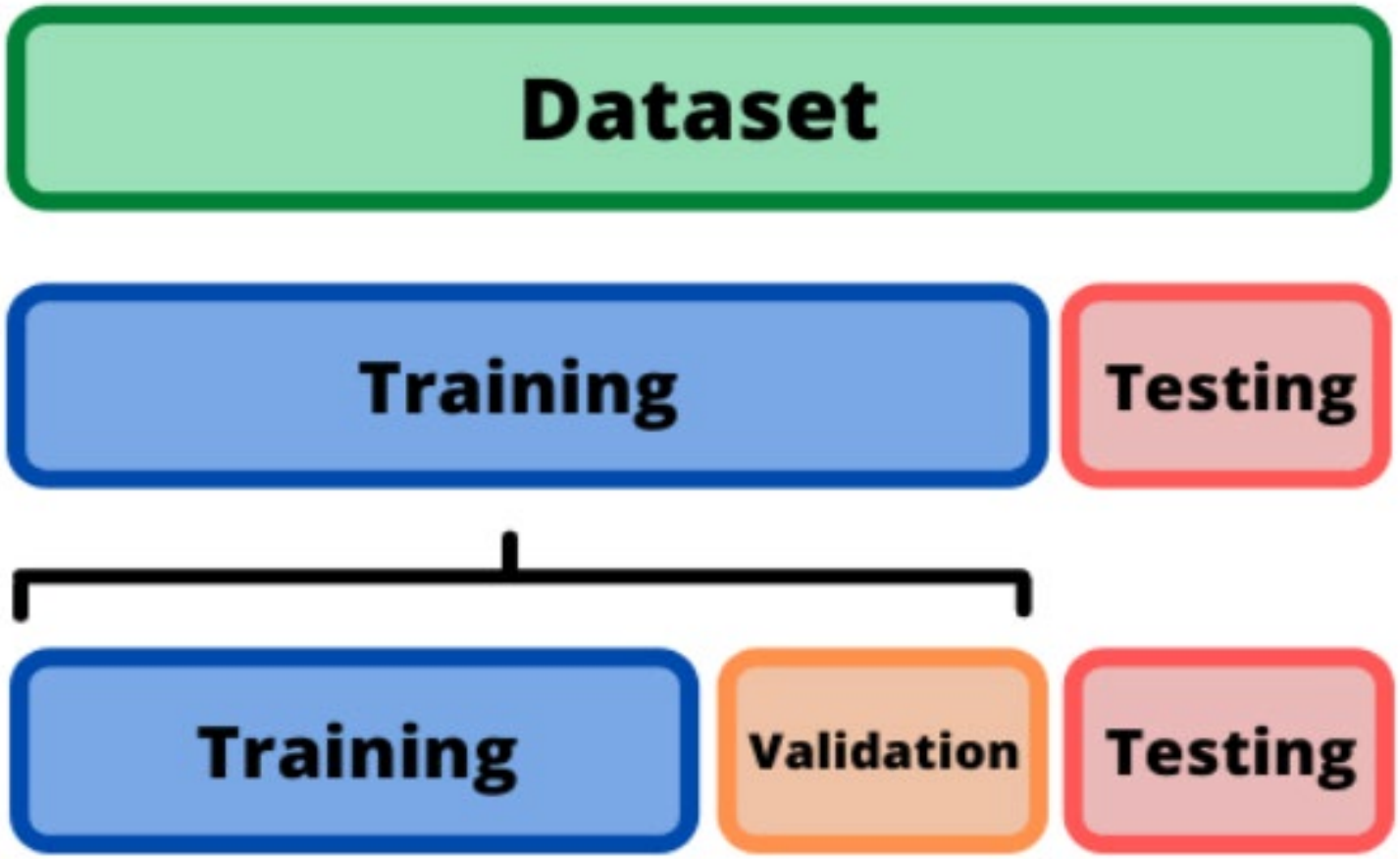
La normalización es útil para varios métodos de minería de datos, que tienden a quedar sesgados por la influencia de los atributos con valores más altos, distorsionando de esta forma el resultado del modelo

Normalización por el máximo: $z_i = \frac{x_i}{x_{\max}}$

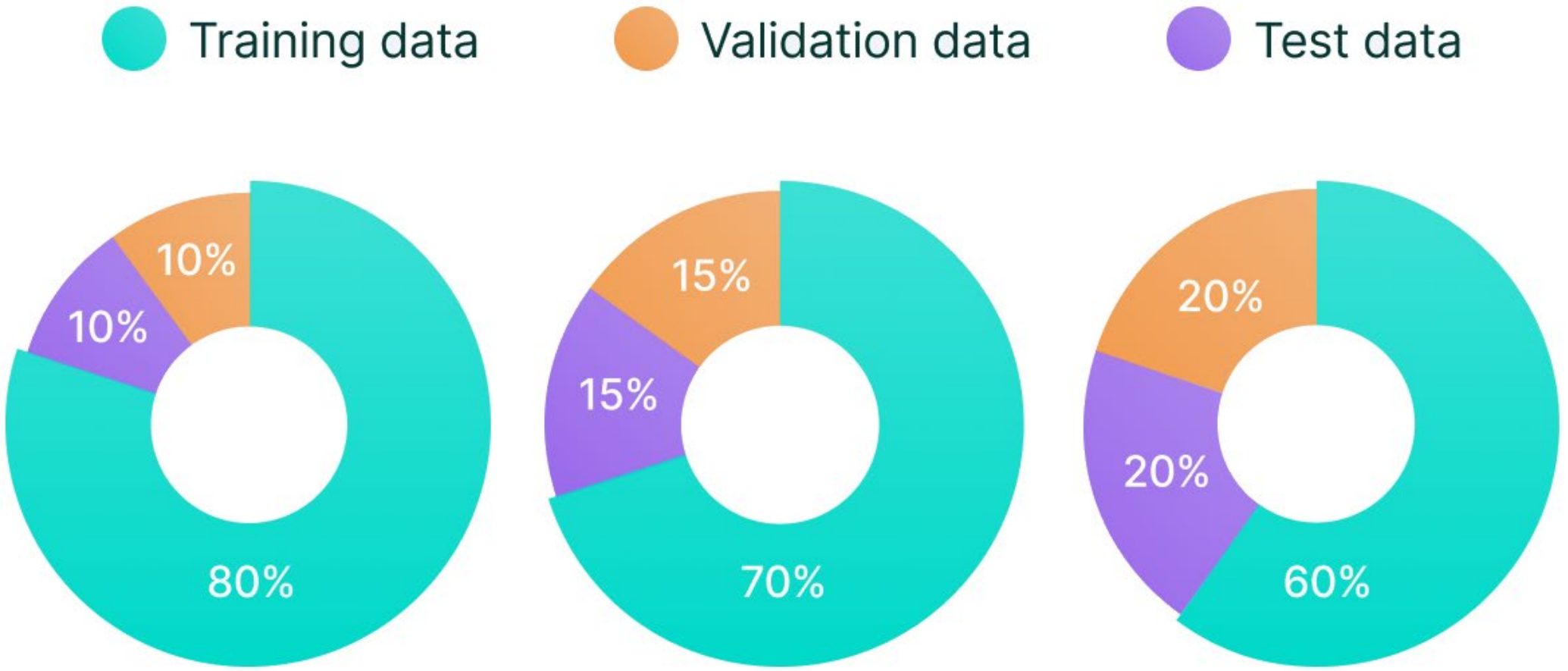
Normalización por la diferencia: $z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$

Normalización basada en la desviación estándar: $z_i = \frac{x_i - \mu}{\sigma}$

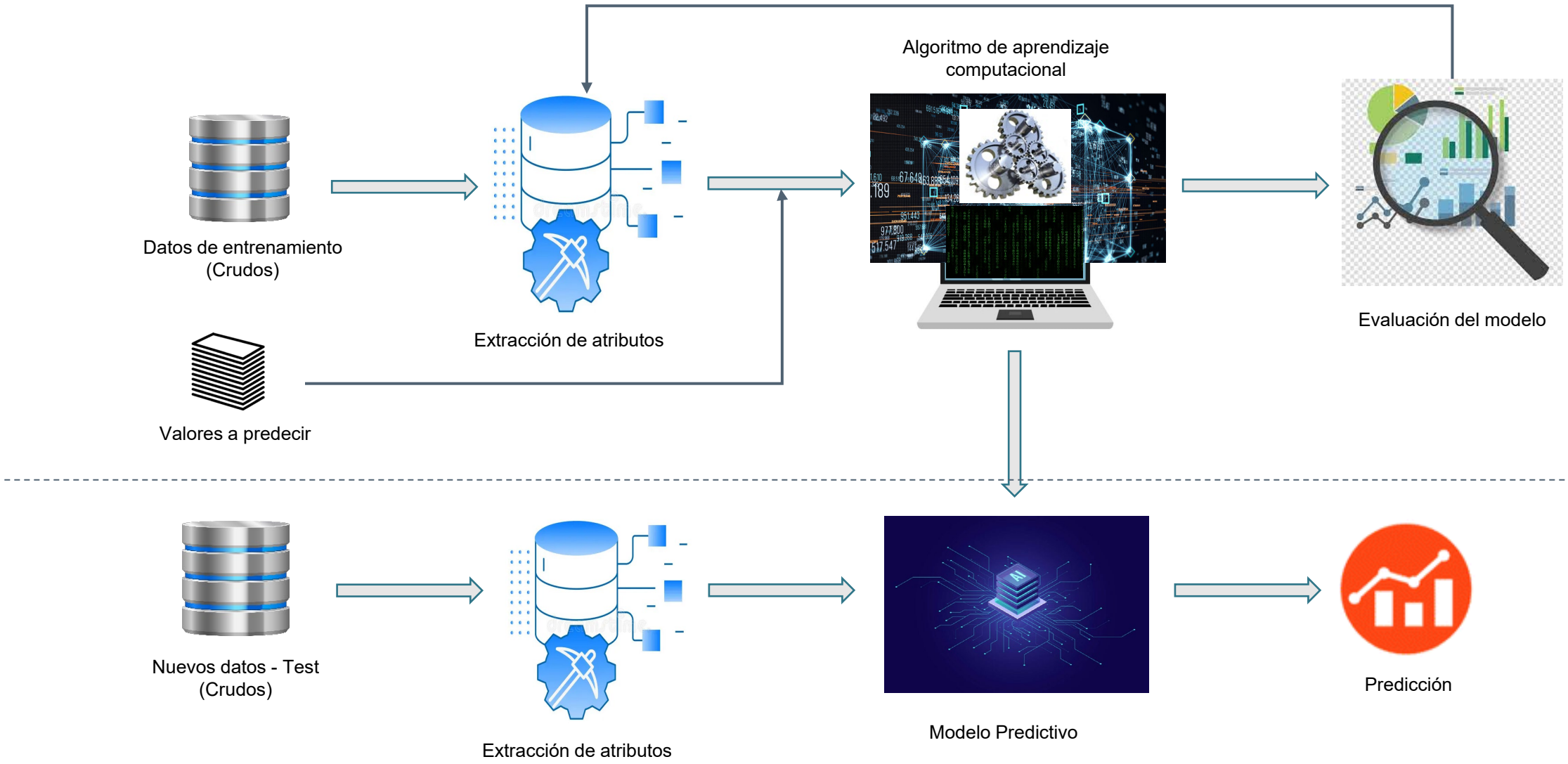
Conjuntos de entrenamiento y test



Conjuntos de entrenamiento y test



Conjuntos de entrenamiento y test



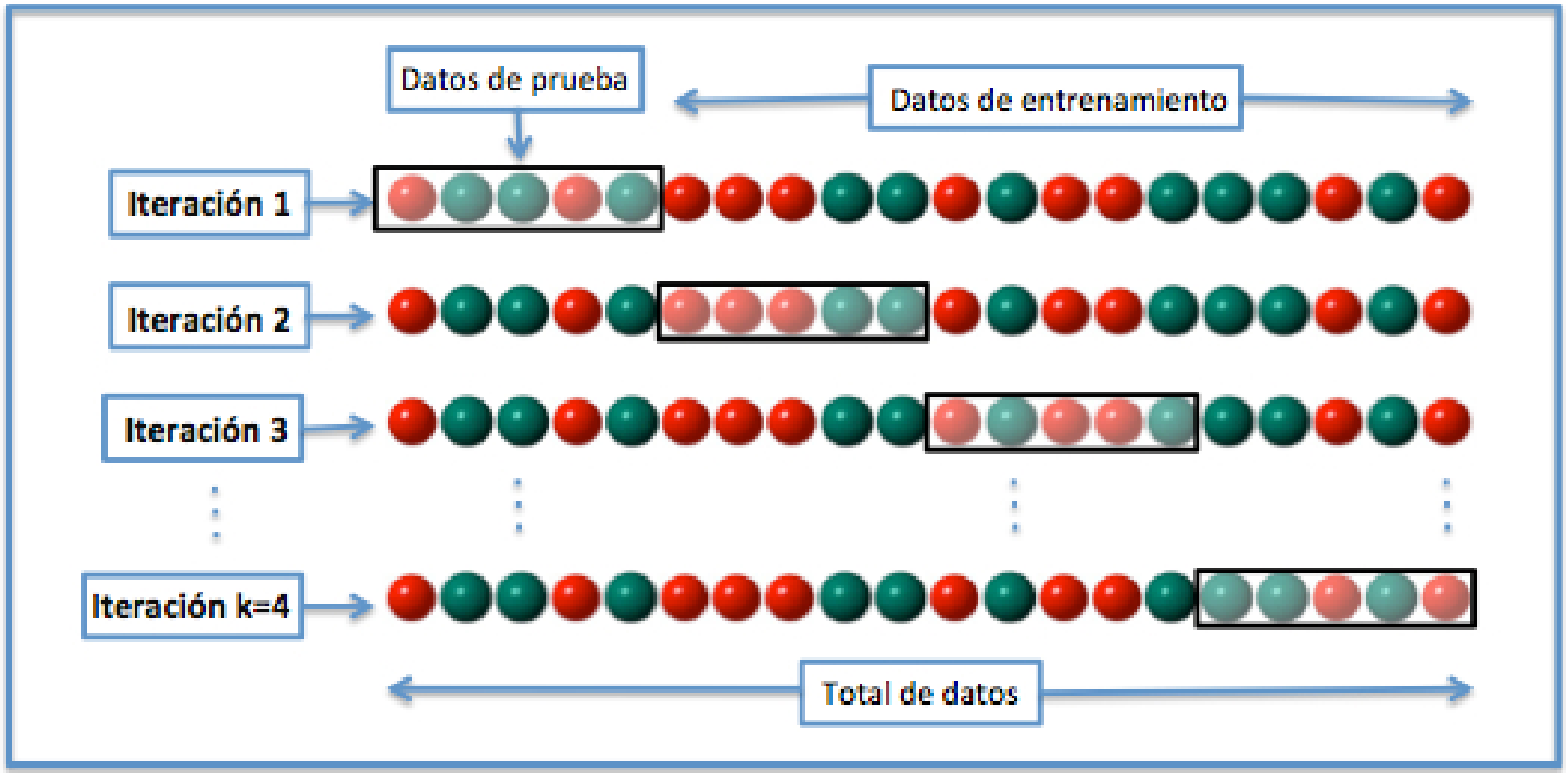
Conjuntos de entrenamiento y test

Habitualmente los conjuntos de datos de entrenamiento y de test suelen ser extracciones aleatorias del juego de datos inicial.

En función del número de datos disponibles, existen diferentes técnicas para la construcción de los conjuntos de entrenamiento y de prueba.

Se trata de un compromiso entre la robustez del modelo construido (a mayor número de datos usados para el entrenamiento, más robusto será el modelo) y su capacidad de generalización (a mayor número de datos usados para la validación, más fiable será la estimación del error cometido).

Validación cruzada k-fold



Conjunto de datos reales

Repositorios libres:

- OpenML.org (https://openml.org/)
- Kaggle.com (https://www.kaggle.com/datasets)
- PapersWithCode.com (https://paperswithcode.com/datasets)
- UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/)
- Amazon’s AWS datasets (https://registry.opendata.aws/)