Minería de Datos

Paradigmas de Aprendizaje, Problemas Canónicos y Datasets

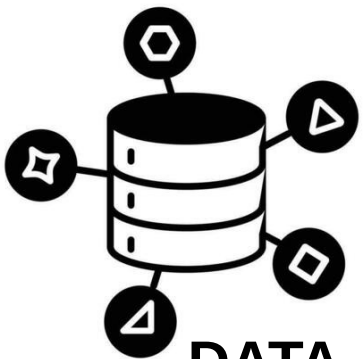
Máster en Ciencia de Datos

Con la colaboración de:

Rodrigo García Manzanas (rodrigo.manzanas@unican.es)
Departamento de Matemática Aplicada y Ciencias de la Computación
Universidad de Cantabria

Contenidos

- Introducción
- Paradigmas de Aprendizaje
- Problemas Canónicos
- Datasets

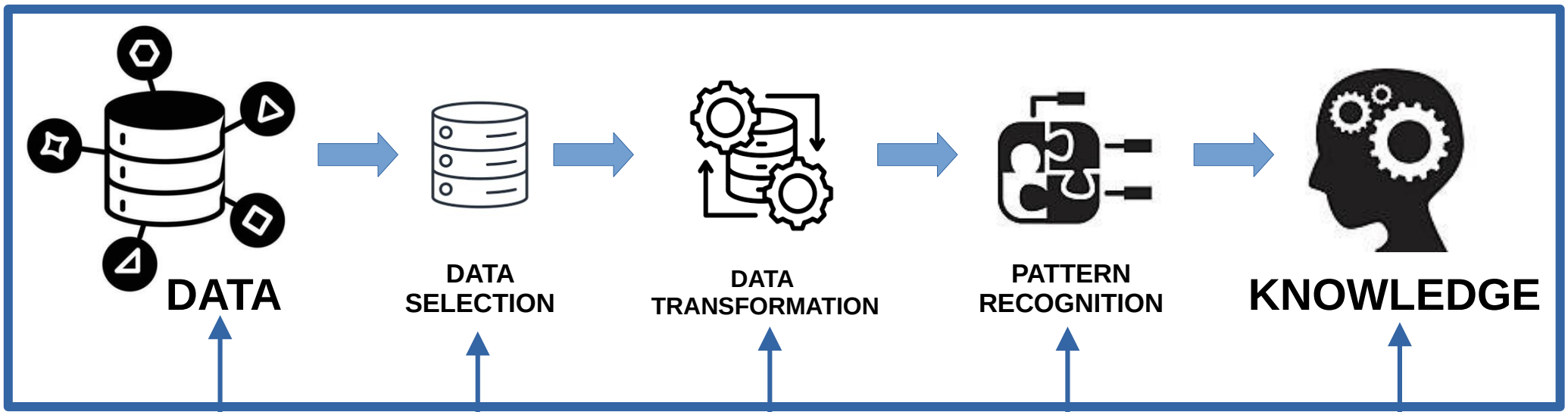


DATA

?

KNOWLEDGE

Knowledge Discovery in Databases (KDD) ~ DATA MINING



- Heterogeneous data types and sources
- Data retrieval
- Data collection
- Data storage
- Data curation

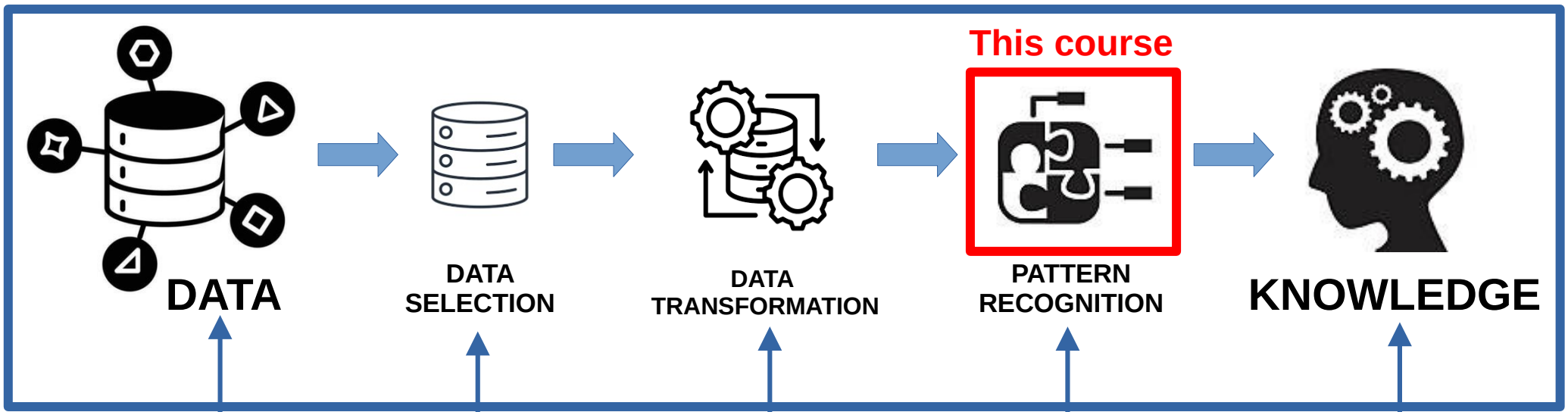
- Sector- and/or application-tailored
- Accurate definition of the problem and objective

- Missing data
- Outliers
- Noise
- Imbalanced data
- Standardization
- Dimensionality reduction

- Models' construction and optimization
- Models' evaluation and intercomparison

- Deployment
- Generation of results (reports, visuals, etc.)
- Decision making

Knowledge Discovery in Databases (KDD) ~ DATA MINING



- Heterogeneous data types and sources
- Data retrieval
- Data collection
- Data storage
- Data curation

- Sector- and/or application-tailored
- Accurate definition of the problem and objective

- Missing data
- Outliers
- Noise
- Imbalanced data
- Standardization
- Dimensionality reduction

- Models' construction and optimization
- Models' evaluation and intercomparison

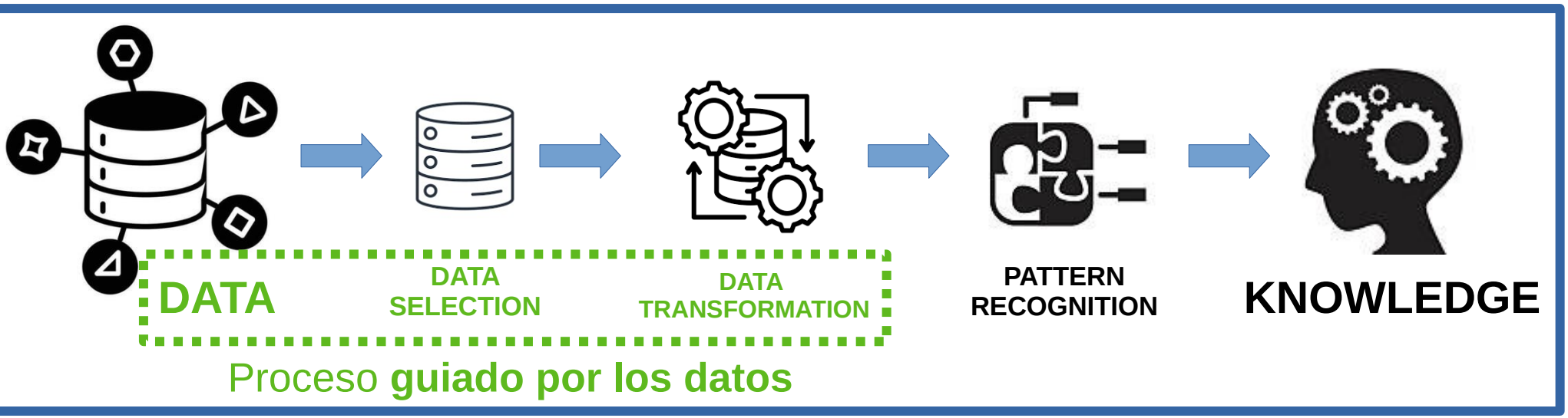
MACHINE LEARNING

- Deployment
- Generation of results (reports, visuals, etc.)
- Decision making

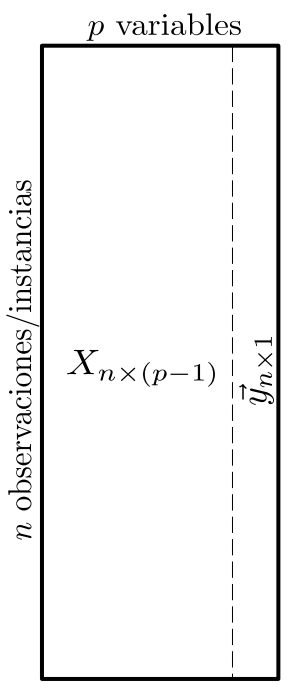
~80% of work

~20% of work

KDD ~ DATA MINING



Típicamente, para poder modelizar los problemas de minería de datos en un ordenador es necesario disponer de una matriz ***ready-to-use*** en la cual cada fila corresponde a una **observación** (o **instancia**) y cada columna a una **variable.**



De acuerdo a su tipología:

- **Cuantitativos**: Continuos (numéricos).
- **Cualitativos/categóricos**: Discretos (binarios o multiclase).

De acuerdo a su función

- **Predictores** (o covariables), **X**: Variables independientes, necesarias para el desarrollo de **modelos**, tanto **predictivos** como **explicativos**.
- **Predictando**, **y**: Variable dependiente, necesaria únicamente para el desarrollo de **modelos predictivos**.

Modelos **predictivos**

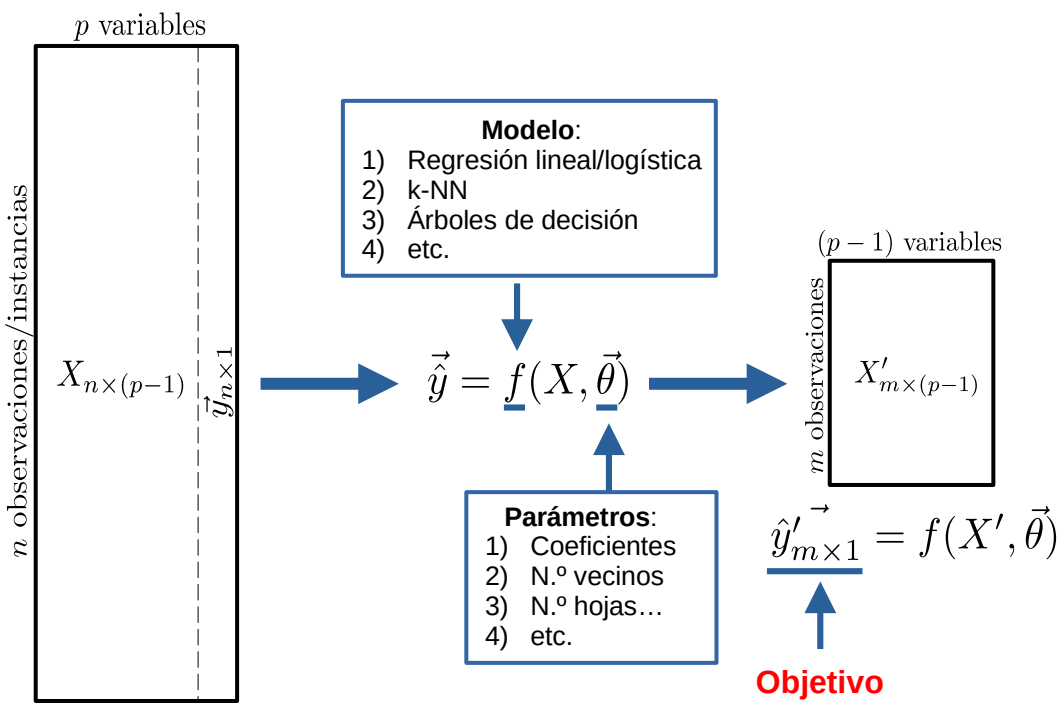
Aprendizaje supervisado

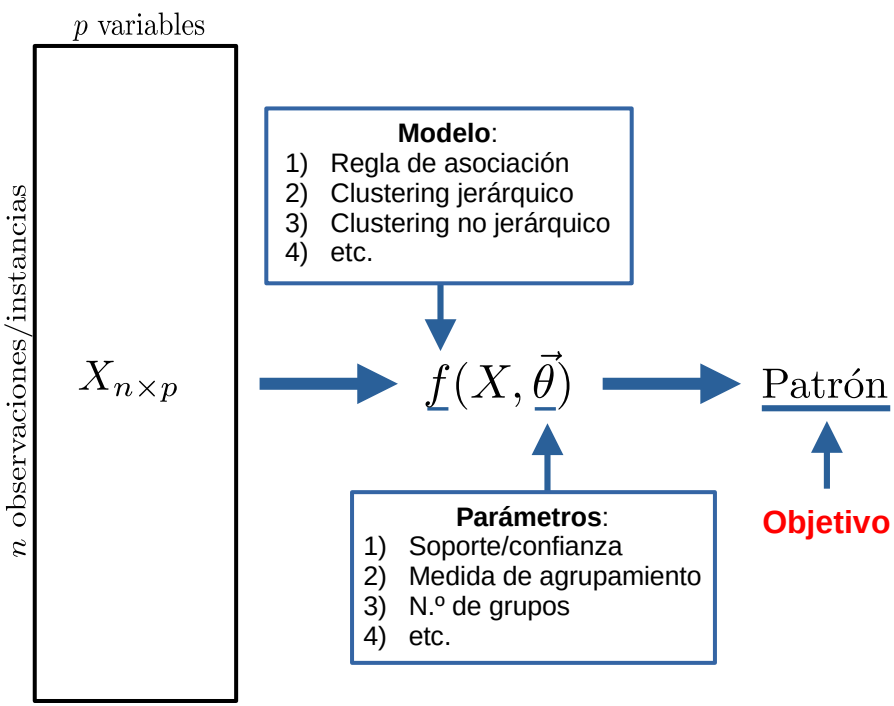
Hay una variable objetivo, y, que está **etiquetada** (bien a través de un valor numérico o una categoría) y depende de una serie de variables explicativas, X. Se trata de encontrar la forma en la que y depende de X (**modelo**), para posteriormente poder predecir la etiqueta y' que le correspondería a nuevos datos explicativos X'.

Modelos **explicativos**

Aprendizaje no supervisado

No hay una variable objetivo, sólo una serie de variables explicativas X. El objetivo es descubrir, a partir de X, algún patrón/estructura presente en los datos que pueda resultar de interés.





Modelos **predictivos**

Aprendizaje supervisado

Modelos **explicativos**

Aprendizaje no supervisado

Offline (o batch)

El modelo se entrena una única vez, con todos los datos que estén disponibles en ese momento, y se congela (modelo **estático**)

- No es útil para aplicaciones en tiempo real.
- No captura tendencias recientes: El modelo no tiene capacidad de especialización.
- Se puede hacer una estimación de los recursos computaciones que se van a necesitar. Suelen ser razonables.

Online

El modelo se va reentrenando continuamente a medida que aparecen nuevos datos (modelo **dinámico**)

- Ideal para aplicaciones en tiempo real.
- Captura tendencias recientes: El modelo se puede ir especializando en la realización de tareas concretas.
- Es difícil hacer una estimación de los recursos computacionales que se van a necesitar. En general, son muy altos.

Aprendizaje por refuerzo

A mitad de camino entre el aprendizaje supervisado y el no supervisado. El objetivo es **aprender en base a la experiencia**, en un proceso de prueba y error, en el que intervienen un *agente* y un *ambiente*.

1) El agente realiza una acción (de entre un número de opciones posibles).
2) Como consecuencia de esa acción, el ambiente cambia de estado.
3) Si el nuevo estado es mejor que el anterior, el agente recibe una recompensa (junto con el nuevo estado del ambiente). En caso contrario, recibe una penalización.
4) El objetivo es **maximizar las recompensas**.

- Brazos mecánicos
- Brokers digitales
- Coche autónomo
- etc.

(en entornos simulados)

Q-learning: Para evitar que el proceso de aprendizaje se detenga hay que:

- Evitar la aplicación repetitiva de acciones conservativas (dilema exploración-explotación)
- Realizar un gran número de simulaciones

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Aprendizaje supervisado

Clasificación (Cla)

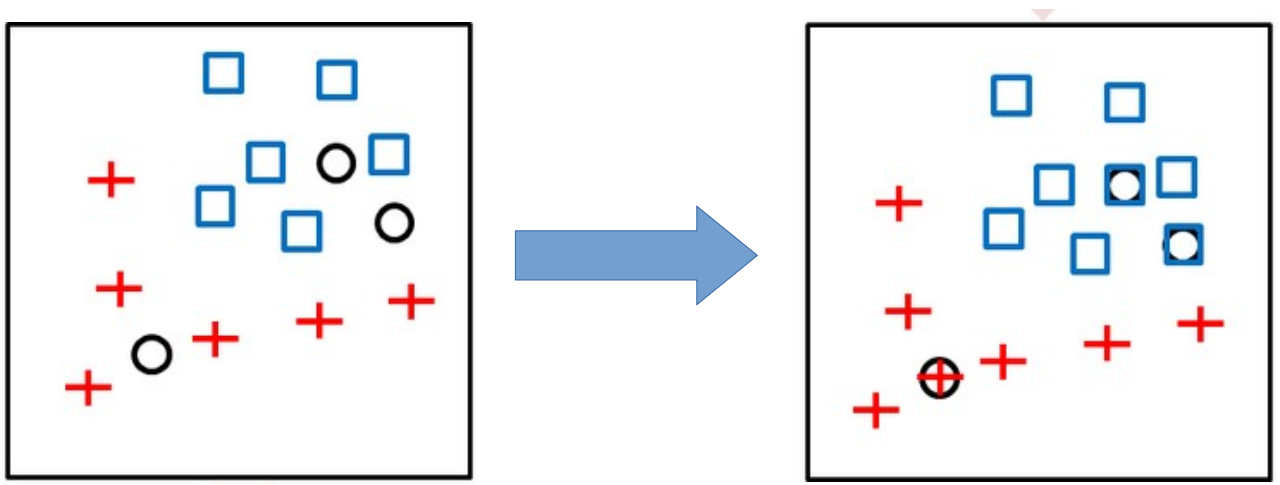
Regresión/ predicción (Reg)

Aprendizaje no supervisado

Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)



X: Categóricas/continuas
y: **Categórica** (binaria o multiclase)

Aprendizaje supervisado

Clasificación (Cla)

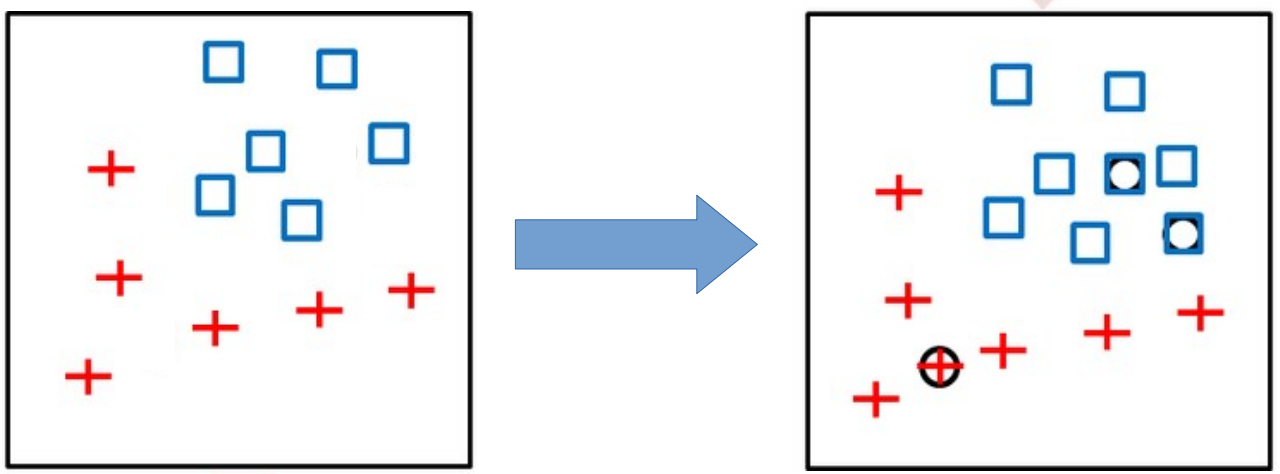
Regresión/ predicción (Reg)

Aprendizaje no supervisado

Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)



X: Categóricas/continuas
y: **Categórica** (binaria o multiclase)

- Detección de spam
- Reconocimiento de objetos
- Diagnóstico médico
- Análisis de sentimientos (en textos)
- etc.

Aprendizaje supervisado

Clasificación (Cla)

Regresión/ predicción (Reg)

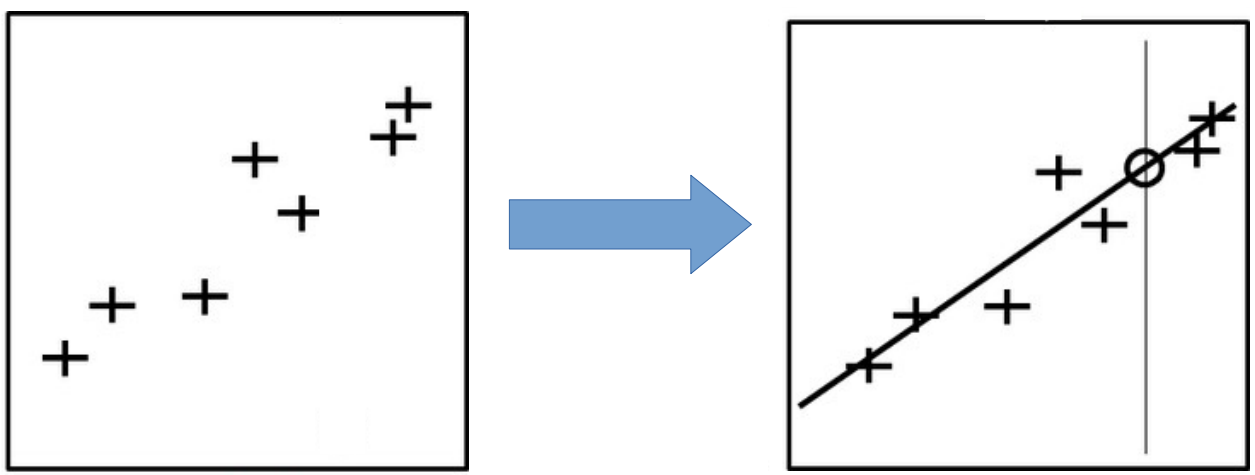
Aprendizaje no supervisado

Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)

Nota: Aunque es frecuente referirse a este problema canónico con el término *regresión*, realmente sería más correcto usar el de *predicción*, puesto que hay un gran abanico de técnicas, más allá de la propia regresión, que se usan en este contexto.



X: Categóricas/continuas
y: **Continua**

- Evolución de las ventas de un comercio
- Evolución del precio de las acciones
- Duración de los ingresos hospitalarios
- etc.

Aprendizaje supervisado

Clasificación (Cla)

Regresión/ predicción (Reg)

Aprendizaje no supervisado

Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)

54% of times

X: Categóricas (binarias)
y: **No hay**

- Colocación de productos (tienda física)
- Recomendación de productos (tienda virtual)
- Diseño de promociones
- etc.

Aprendizaje supervisado

Clasificación (Cla)

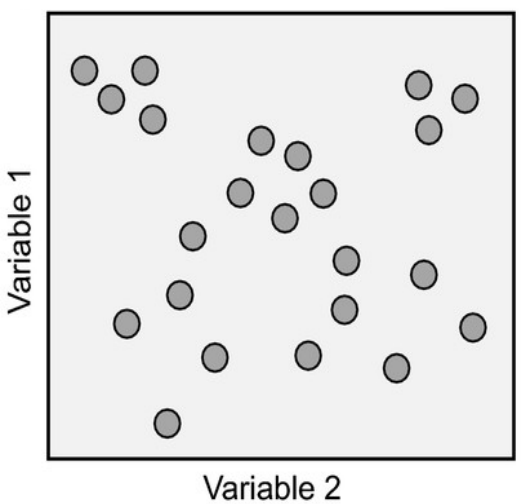
Regresión/ predicción (Reg)

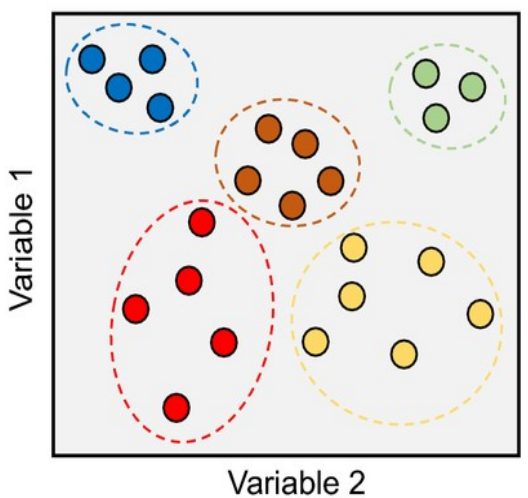
Aprendizaje no supervisado

Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)





X: Continuas/categóricas
y: **No hay**

Segmentación de:
- Clientes (comercio)
- Pacientes (medicina)
- Especies (biología)
- etc.

Aprendizaje supervisado

Clasificación (Cla)

Regresión/ predicción (Reg)

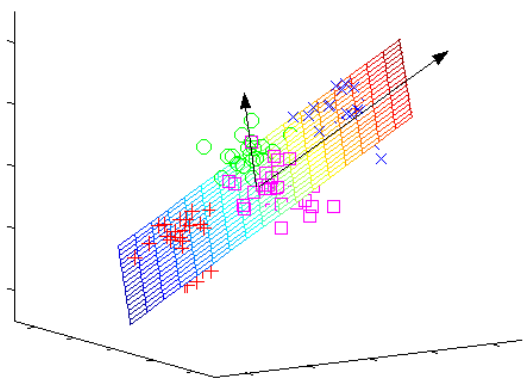
Aprendizaje no supervisado

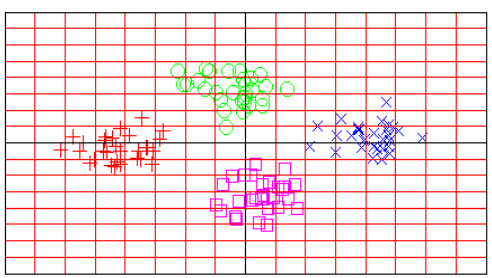
Asociación (Aso)

Clustering (Clu)

Reducción de la dimensionalidad (RDim)

	Samples	Genes	Classes
Leukemia	72	7129	2
Prostate cancer	102	12600	2





X: Continuas/categóricas
y: **No hay**

Filtrado de información irrelevante/redundante en:

- Análisis genómicos
- Detección de partículas
- Predicción meteorológica
- etc.

Aprendizaje supervisado

Clasificación (Cla)

Regresión/ predicción (Reg)

- Regresión logística (Cla)
- Regresión lineal (Reg)
- k-NN (Cla, Reg)
- Árboles de decisión (Cla, Reg)
- Métodos de ensembles: Random forests (Cla,Reg); AdaBoost (Cla); GBoost (Cla, Reg)
- Métodos de kernels (Cla, Reg)
- Máquinas de vector soporte (Cla, Reg)
- Redes neuronales (Cla, Reg)
- Redes probabilísticas (Cla, Reg)
- etc.

Aprendizaje no supervisado

Asociación (Aso)

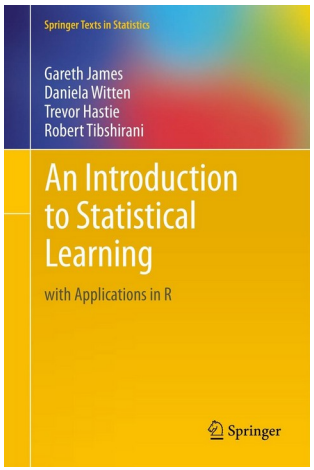
Clustering (Clu)

Reducción de la dimensionalidad (RDim)

- Reglas de asociación (Aso): Algoritmo Apriori, Algoritmo Eclat
- Clustering jerárquico (Clu): Dendograma
- Clustering no jerárquico (Clu): k-means
- Reducción de la dimensionalidad lineal (RDim): PCA, LDA
- Reducción de la dimensionalidad no lineal (RDim): MDS, MMF, Isomap, LLE, SNE
- Redes probabilísticas (Rdim)
- etc.

Minería de Datos Estadística Aprendizaje Automático I Aprendizaje Automático II

1




```
install.packages("ISLR")
library("ISLR")
library(help = "ISLR")
```

[An Introduction to Statistical Learning: With Applications in R]
James, G., Witten, D., Hastie, T., Tibshirani, R.
Springer (2013)

2

```
library(help = "datasets")
```

3

kaggle

https://www.kaggle.com/datasets

4

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

https://archive.ics.uci.edu

Incluye código (notebooks), discusiones, cursos, competiciones, etc.

Disponible en el **GitHub** de la asignatura

Iris Species

Classify iris plants into three species in this classic dataset

https://www.kaggle.com/uciml/iris

Mixto →

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 ...
```

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

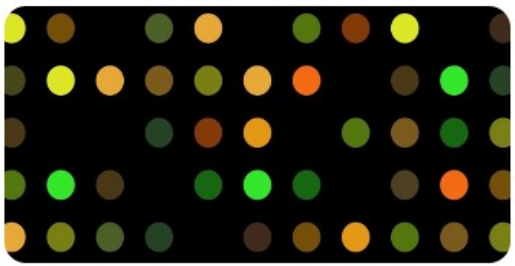
Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Disponible en el **GitHub** de la asignatura

Gene expression dataset (Golub et al.)

Molecular Classification of Cancer by Gene Expression Monitoring



https://www.kaggle.com/datasets/crawford/gene-expression?resource=download

Mixto →

```
'data.frame':   72 obs. of  7130 variables:
 $ X1   : num  -1.0172 -0.2498 0.3948 -0.2089 0.0878 ...
 $ X2   : num  0.0741 0.9063 1.1559 0.4798 0.3654 ...
 $ X3   : num  -0.4069 0.0576 -2.4363 2.2255 -0.5536 ...
 $ X4   : num  -0.906 0.837 1.069 -1.585 -0.191 ...
 $ X5   : num  -0.3412 -0.0875 -1.0042 -1.3561 0.1907 ...
 $ X6   : num  -1.0771 -0.0258 -1.6892 -1.2567 0.7461 ...
 $ X7   : num  0.927 -0.932 0.343 0.782 0.242 ...
 $ X8   : num  0.147 0.227 -1.765 -0.624 0.688 ...
 $ X9   : num  1.923 0.234 1.408 -0.348 -0.113 ...
 $ X10  : num  0.474 -0.106 -1.378 -0.296 0.677 ...
                          ...
 $ X7128: num  -0.389 -0.573 -0.329 -0.493 -0.605 ...
 $ X7129: num  -0.16 0.412 -0.26 -1.504 0.139 ...
 $ label: chr  "ALL" "ALL" "ALL" "ALL" ...
```

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Disponible en el **GitHub** de la asignatura

Play tennis

Simple dataset with decisions about playing tennis

https://www.kaggle.com/datasets/fredericobreno/play-tennis/code

Categórico →

```
'data.frame':   14 obs. of  5 variables:
 $ outlook : chr  "sunny" "sunny" "overcast" "rainy" ...
 $ temp    : chr  "hot" "hot" "hot" "mild" ...
 $ humidity: chr  "high" "high" "high" "high" ...
 $ windy   : chr  "false" "true" "false" "false" ...
 $ play    : chr  "no" "no" "yes" "yes" ...
```

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Disponible en el **GitHub** de la asignatura

Instacart Market Basket Analysis

Which products will an Instacart consumer purchase again?

https://www.kaggle.com/c/instacart-market-basket-analysis

Utilizaremos un dataset más pequeño, “Groceries”, disponible en el paquete de R **arulesViz**

```
install.packages("arulesViz")
data(“Groceries”)
Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
```

← **Categórico**

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Disponible en el **GitHub** de la asignatura

Mushroom Classification

Safe to eat or deadly poison?

https://www.kaggle.com/uciml/mushroom-classification

Categórico

```
'data.frame':   8124 obs. of  23 variables:
 $ class                   : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
 $ cap.shape               : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...
 $ cap.surface             : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...
 $ cap.color               : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10 ...
 $ bruises                 : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
 $ odor                    : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
                                        etc.
```

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Disponible en el **GitHub** de la asignatura

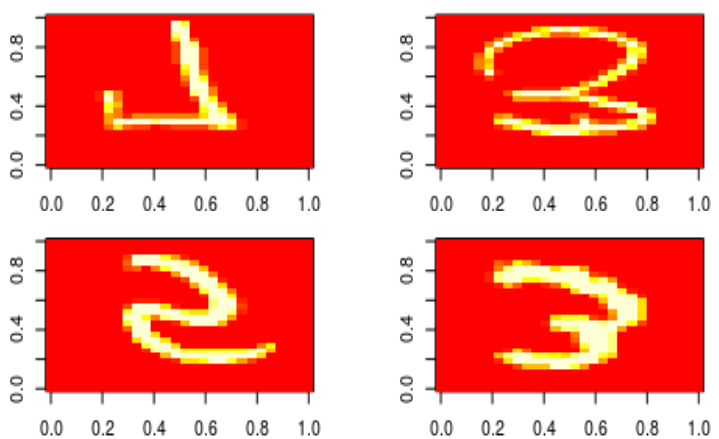
Digit Recognizer

Learn computer vision fundamentals with the famous MNIST data

https://www.kaggle.com/c/digit-recognizer/data

Mixto →

```
'data.frame':       42000 obs. of  785 variables:
 $ label  : int  1 0 1 4 0 0 7 3 5 3 ...
 $ pixel0 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pixel1 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pixel2 : int  0 0 0 0 0 0 0 0 0 0 ...
                 etc.
```



Aprendizaje supervisado

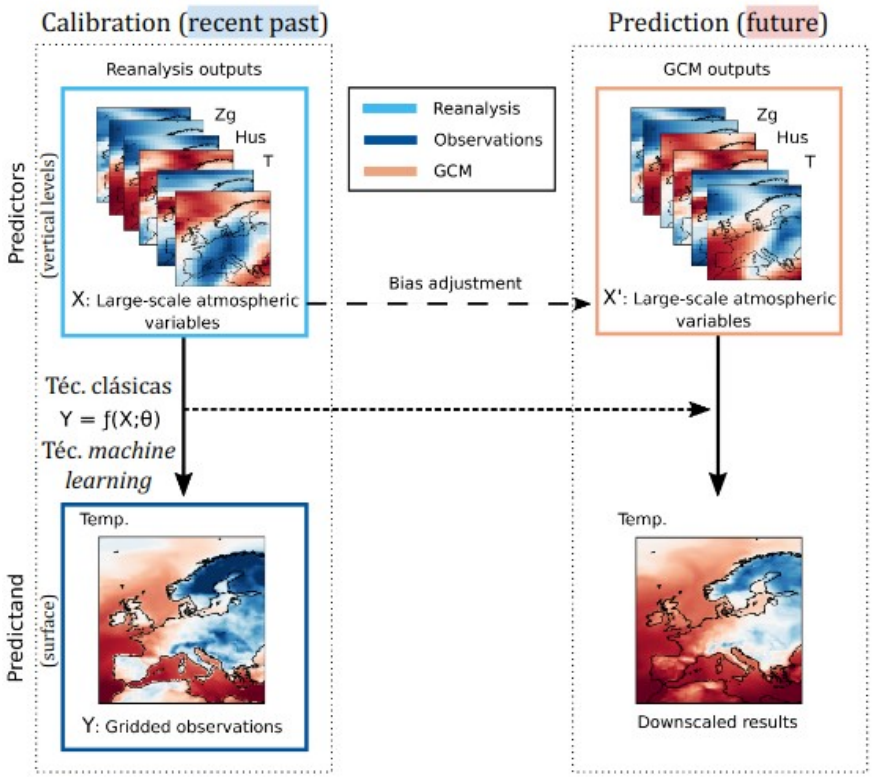
- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Datasets

Disponible en el **GitHub** de la asignatura



y: Precipitación diaria en Lisboa para el período 1979-2008

X: 8 variables atmosféricas de larga escala, definidas sobre una malla de 40 puntos que cubre la Península Ibérica (8 x 40 = 320 predictores):

- Altura geopotencial en 500 hPa
- Temperatura del aire en 850 hPa, 700 hPa y 500 hPa
- Temperatura del aire en superficie
- Humedad específica del aire en 850 hPa
- Presión a nivel del mar

Continuo

```
'data.frame':   10958 obs. of  321 variables:
 $ y   : num  10.9 0.6 13 0 0 1.2 1.1 0 0 0.7 ...
 $ X1  : num  57043 56963 56523 54628 53584 ...
 $ X2  : num  56535 56493 55971 53980 53391 ...
 $ X3  : num  55884 55931 55304 53494 53310 ...
 $ X4  : num  55176 55340 54498 53073 53293 …
                         etc.
```

Aprendizaje supervisado

- Clasificación (Cla)
- Regresión/ predicción (Reg)

Aprendizaje no supervisado

- Asociación (Aso)
- Clustering (Clu)
- Reducción de la dimensionalidad (RDim)

Ejercicio 1 (dataset **iris**)
Dibuja un scatterplot en el que visualices un predictor frente a otro, para todos los pares que se puedan formar. En cada caso, identifica con distintos colores las diferentes clases de iris presentes en el dataset. ¿Puedes sacar alguna conclusión?

Ejercicio 2 (dataset **mushrooms**)
1) Crea un gráfico de barras que muestre el porcentaje de setas de cada color que hay en el dataset.
2) Crea un p-color que muestre el número de setas que hay para cada combinación posible de color y forma (puedes utilizar la función image.plot del paquete fields).

Ejercicio 3 (dataset **digit recognizer**)
1) Crea un gráfico de barras que muestre el porcentaje de observaciones que hay para cada dígito en el dataset. ¿Dirías que está balanceado?
2) Dibuja los 9 primeros "8" (puedes utilizar la función image.plot del paquete fields).

Ejercicio 4 (dataset **meteo**)
1) Dibuja la serie temporal de precipitación diaria en Lisboa
2) Dibuja el valor medio de cada variable predictora y su rango de variabilidad (diferencia entre el valor máximo y el mínimo). ¿Qué puedes concluir?
3) Calcula la correlación (función cor) existente entre todos los pares posibles de predictores y dibújala en un p-color. ¿Qué puedes concluir?
4) Calcula la correlación existente entre cada predictor y la precipitación diaria en Lisboa (considera la correlación de Spearman en este caso). Dibuja los resultados obtenidos en un gráfico y marca todas las correlaciones cuyo valor absoluto sea superior a 0.4