DATA MINING

Ing. Julio Paciello
juliopaciello@cds.com.py

Prof. Ing. Wilfrido Inchaustti
winchaus@uca.edu.py

Knowledge Discovery (KDD)

KDD: Proceso no trivial de descrubrir conocimientos mediante la identificación de patrones en los datos, en forma válida, novedosa, potencialmente útil y entendible

Knowledge Discovery (KDD)

- *Datos*: es el conjunto de hechos F.

- *Patrón*: es una expresión E en un lenguaje L que describe los hechos en un subconjunto F(E) de F. E es denominado patrón si es más simple que la enumeración de todos los hechos en F(E). Ej: Se considera $f(x)=3x^2+x$ un patrón y $f(x)=\alpha x^2+\beta x$ un modelo.

Knowledge Discovery (KDD)

- *Proceso*: consiste en la preparación de los datos, búsqueda de patrones, evaluación del conocimiento y refinamiento. El proceso se asume como no trivial, en el sentido de que la búsqueda no es autónoma.

- *Válido*: el descubrimiento de patrones debe ser válido sobre los datos nuevos bajo un cierto grado de certeza.

- *Útil*: los patrones deben potencialmente conducir a alguna acción útil.

Knowledge Discovery (KDD)

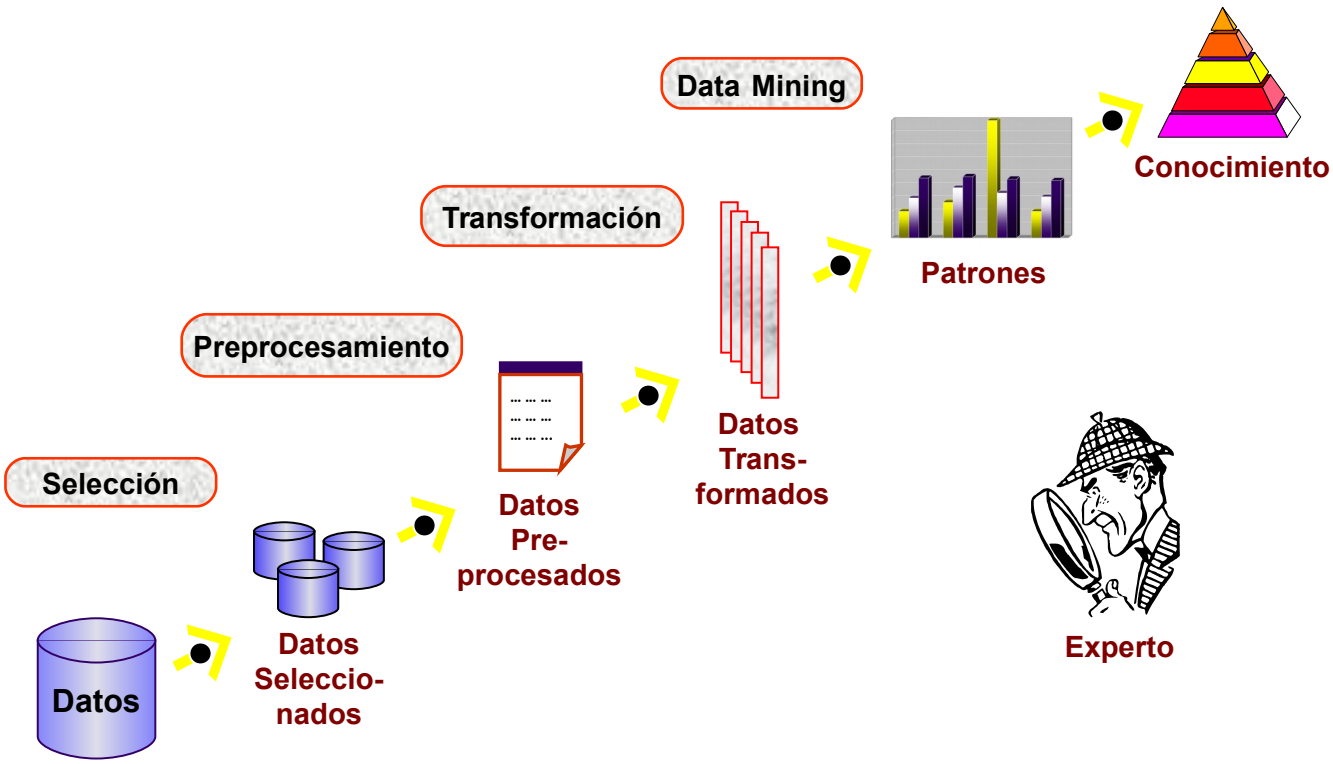
- *Novedoso*: los patrones deben ser novedosos. La novedad puede ser medida con respecto a los cambios en los datos (comparando los valores actuales, con los anteriores o con los esperados) o en el conocimiento (cómo un nuevo hallazgo se relaciona con los anteriores).

- *Entendible*: un objetivo del KDD es construir patrones entendibles para los humanos en orden a facilitar un mejor entendimiento de los datos.

Data Mining

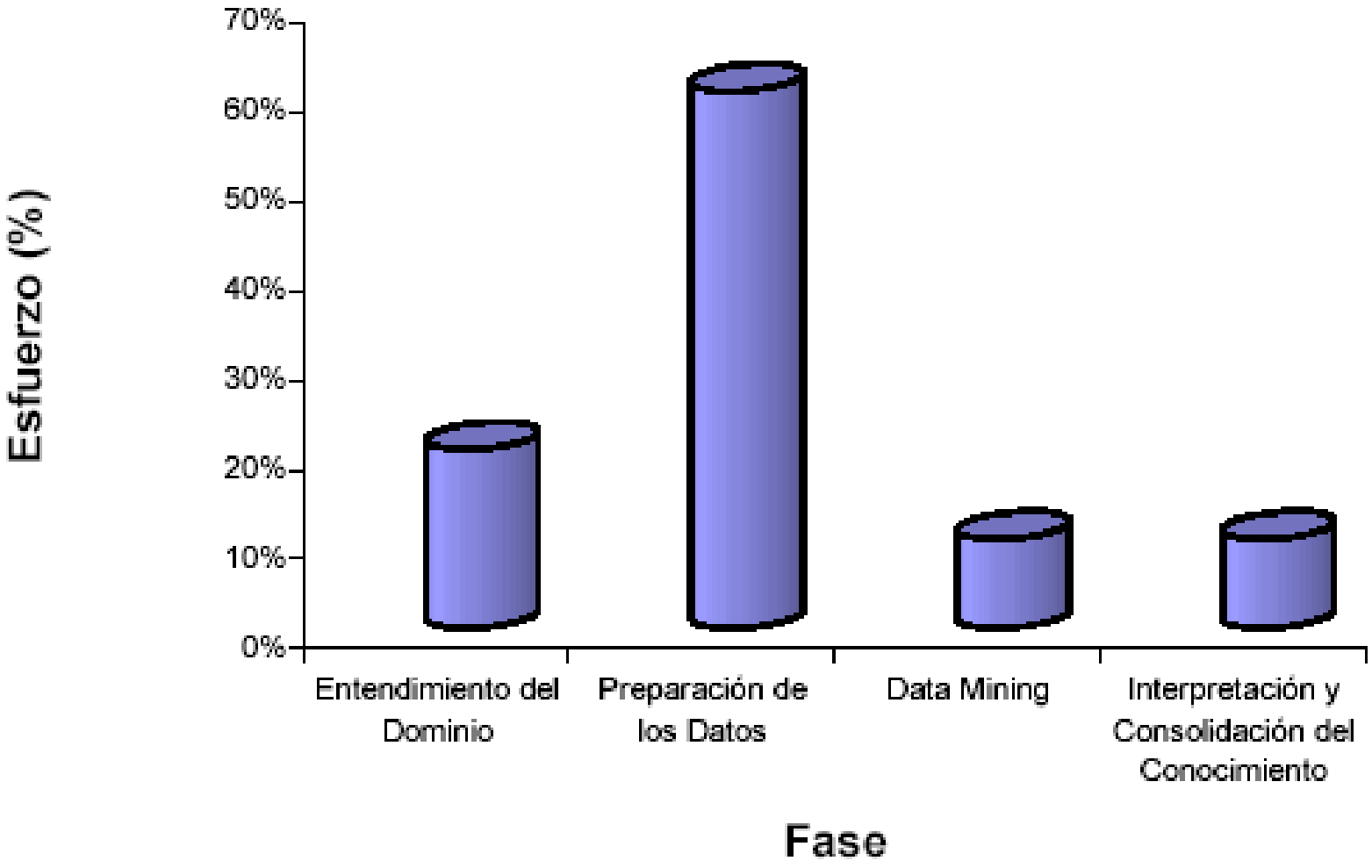
Data Mining es un paso en el proceso del KDD consistiendo de algoritmos particulares que, bajo algunas limitaciones aceptables de eficiencia computacional, produce una enumeración particular de patrones E_j sobre F

Proceso de KDD

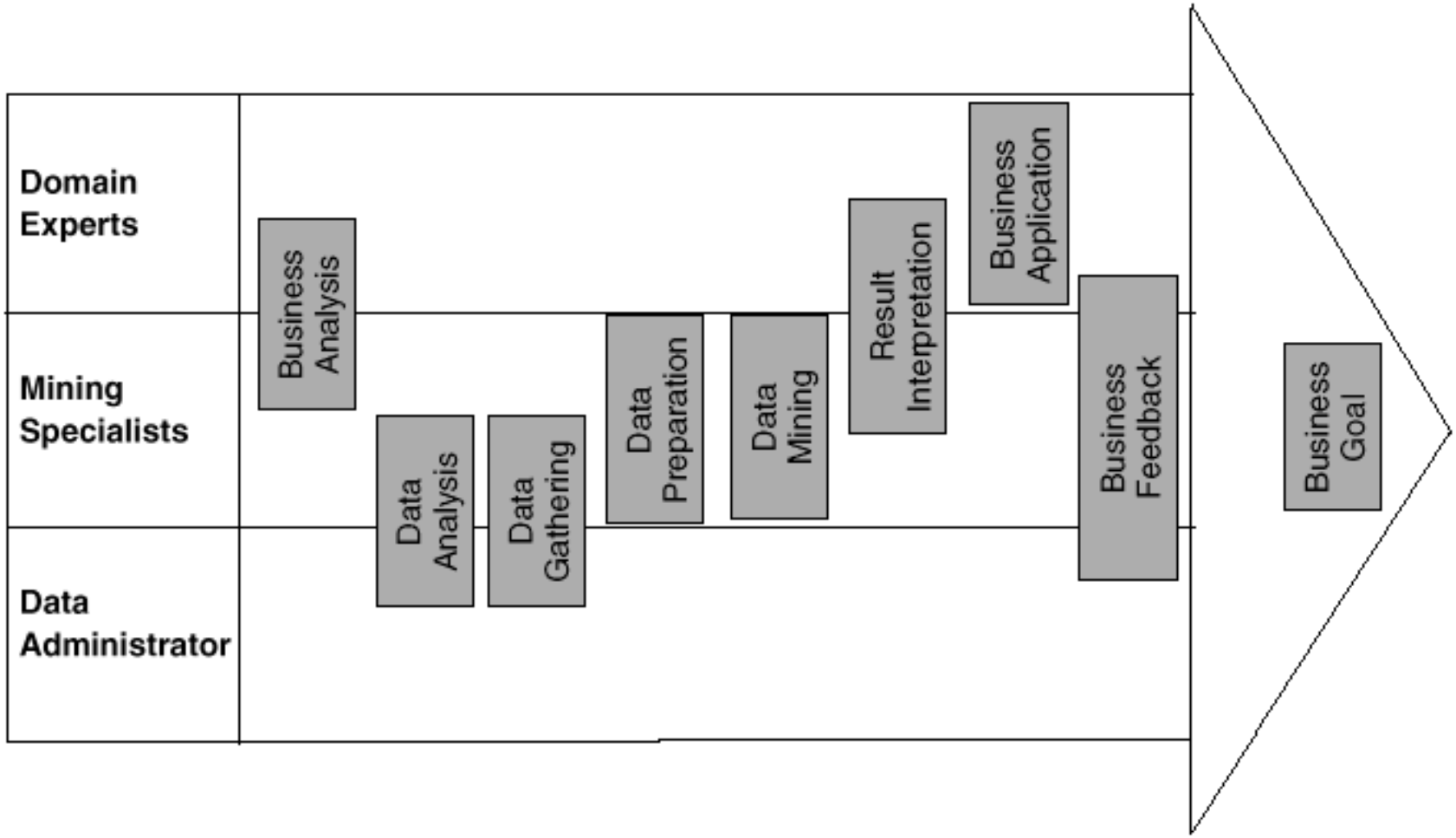


Proceso interactivo e iterativo que envuelve varios pasos y con decisiones a ser tomadas por el usuario

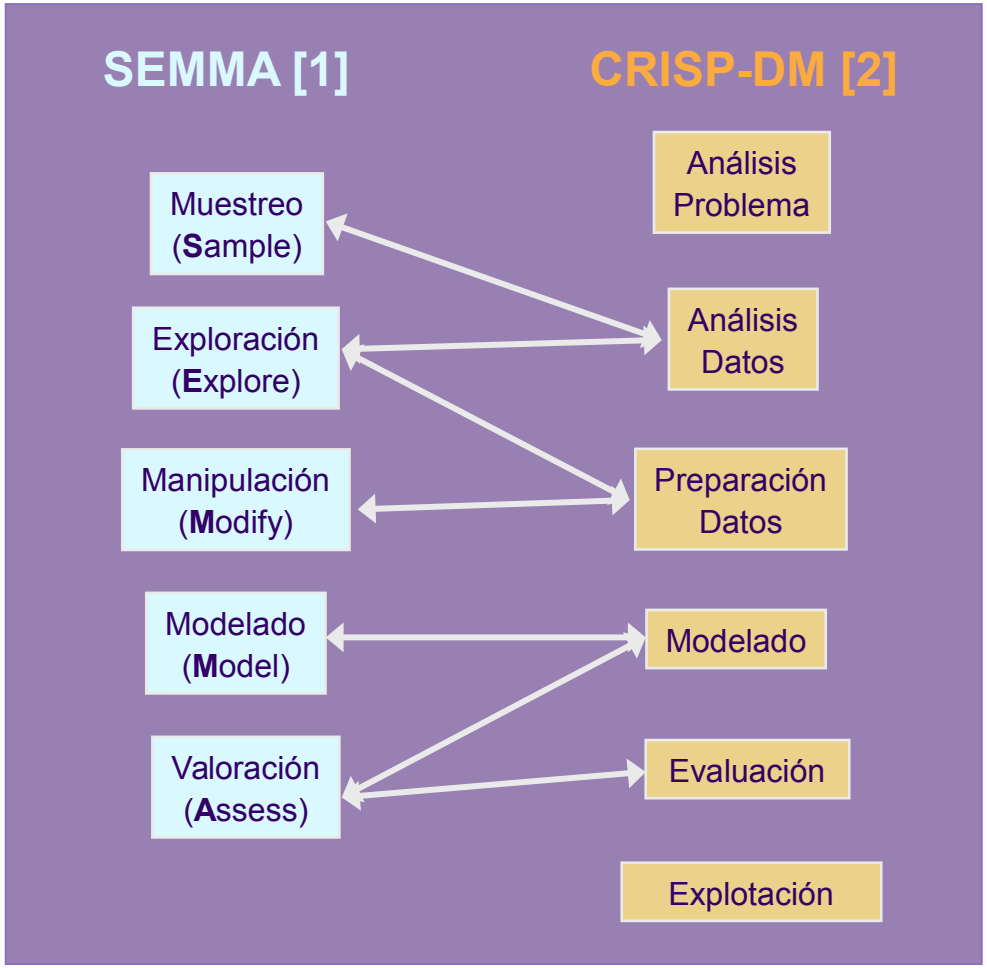
Esfuerzo requerido KDD



Roles en KDD


Domain Experts
Mining Specialists
Data Administrator
Business Analysis
Data Analysis
Data Gathering
Data Preparation
Data Mining
Result Interpretation
Business Application
Business Feedback
Business Goal

Metodologías de KDD



[1] http://www.sas.com/technologies/analytics/datamining/miner/semma.html
[2] http://www.crisp-dm.org/

Objetivos, Tareas y Técnicas

Objetivos

- **La Predicción (Directed data mining)**: consiste en utilizar algunas variables o campos de la Base de Datos para predecir valores desconocidos o futuros de otras variables de interés. Un modelo predictivo responde preguntas sobre datos futuros. Ej. ¿Cuáles serán las ventas el año próximo?, ¿Es esta transacción fraudulenta?, ¿Qué tipo de seguro es más probable que contrate el cliente X?

- **La Descripción (Undirected data mining)**: se centra en encontrar patrones interpretables por el ser humano, a partir de la descripción de los datos. Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características. Ej. a) Los clientes que compran pañales suelen comprar cerveza. b) El tabaco y el alcohol son los factores más importantes en la enfermedad Y. c) Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto.

Asociación

- **Modelo de Dependencias (o Asociación):** consiste en encontrar un modelo el cual describa las dependencias significantes entre las variables. De otra manera, dado un conjunto de datos, identificar las relaciones entre atributos, de forma tal a identificar que la ocurrencia de cierto/s patrón/es implica la ocurrencia de otro/s. Ej.: el 70% de los clientes que consumen el producto A y B, también consumen el producto C, D y E.

IF outlook = overcast
THEN play = yes (4.0)

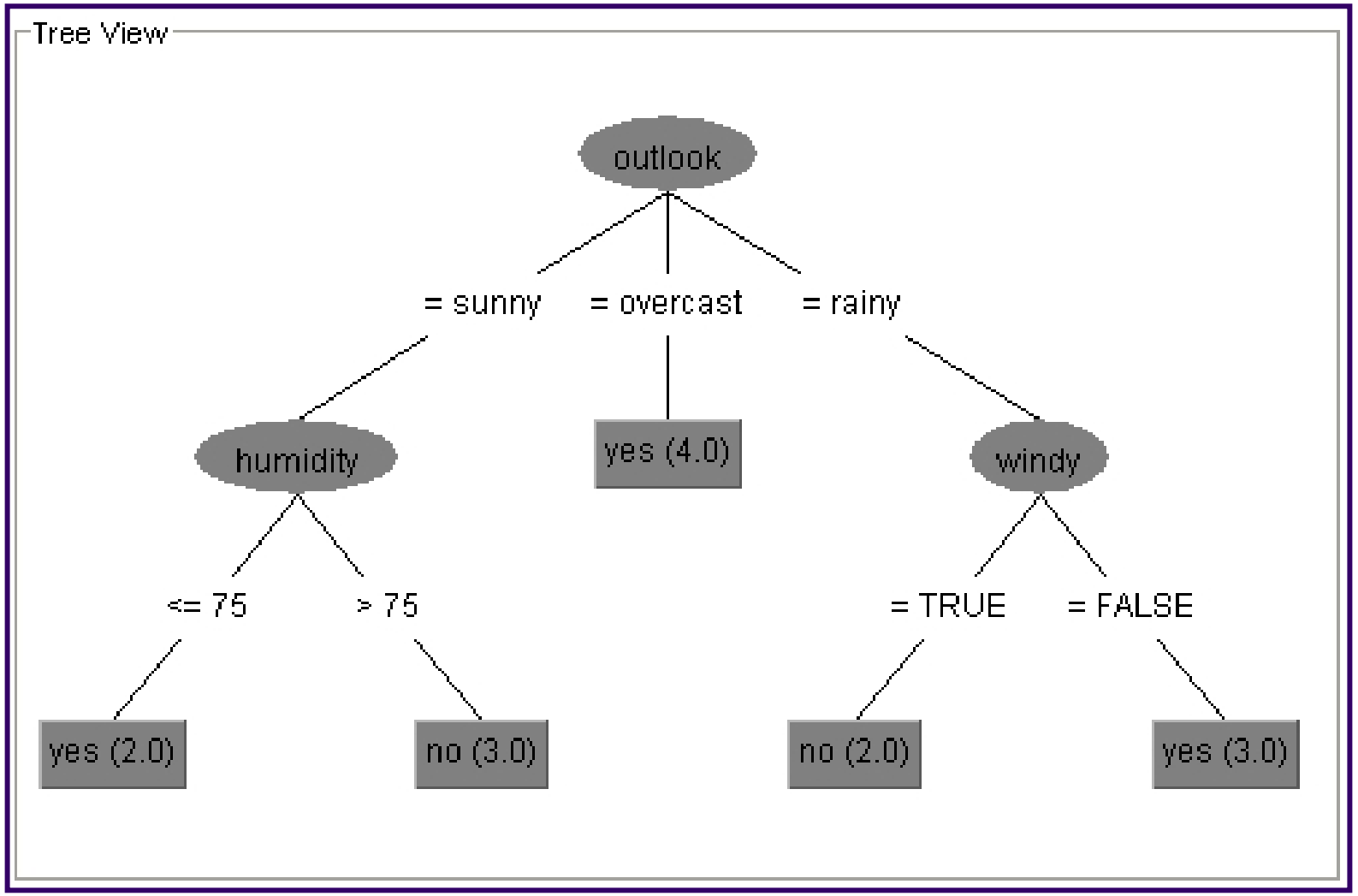
IF windy = TRUE AND
outlook = rainy
THEN play = no (2.0)

IF outlook = sunny AND
humidity > 75
THEN play = no (3.0)

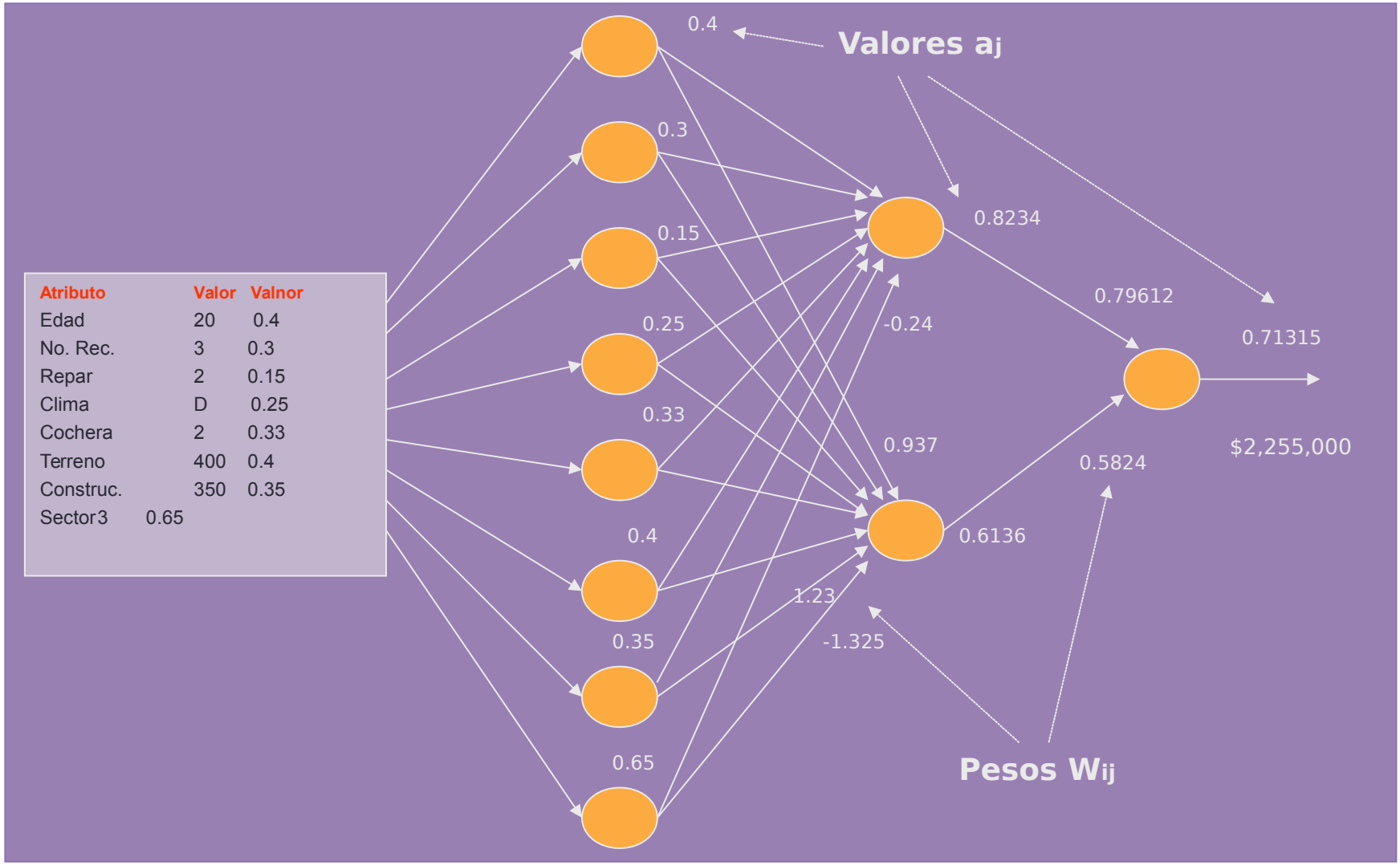
Clasificación

- **Clasificación:** se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta, dicho de otra manera, se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas. Ej.: se sabe (por un estudio de dependencias) que los atributos edad, grado de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria. Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

Clasificación – Decision Tree



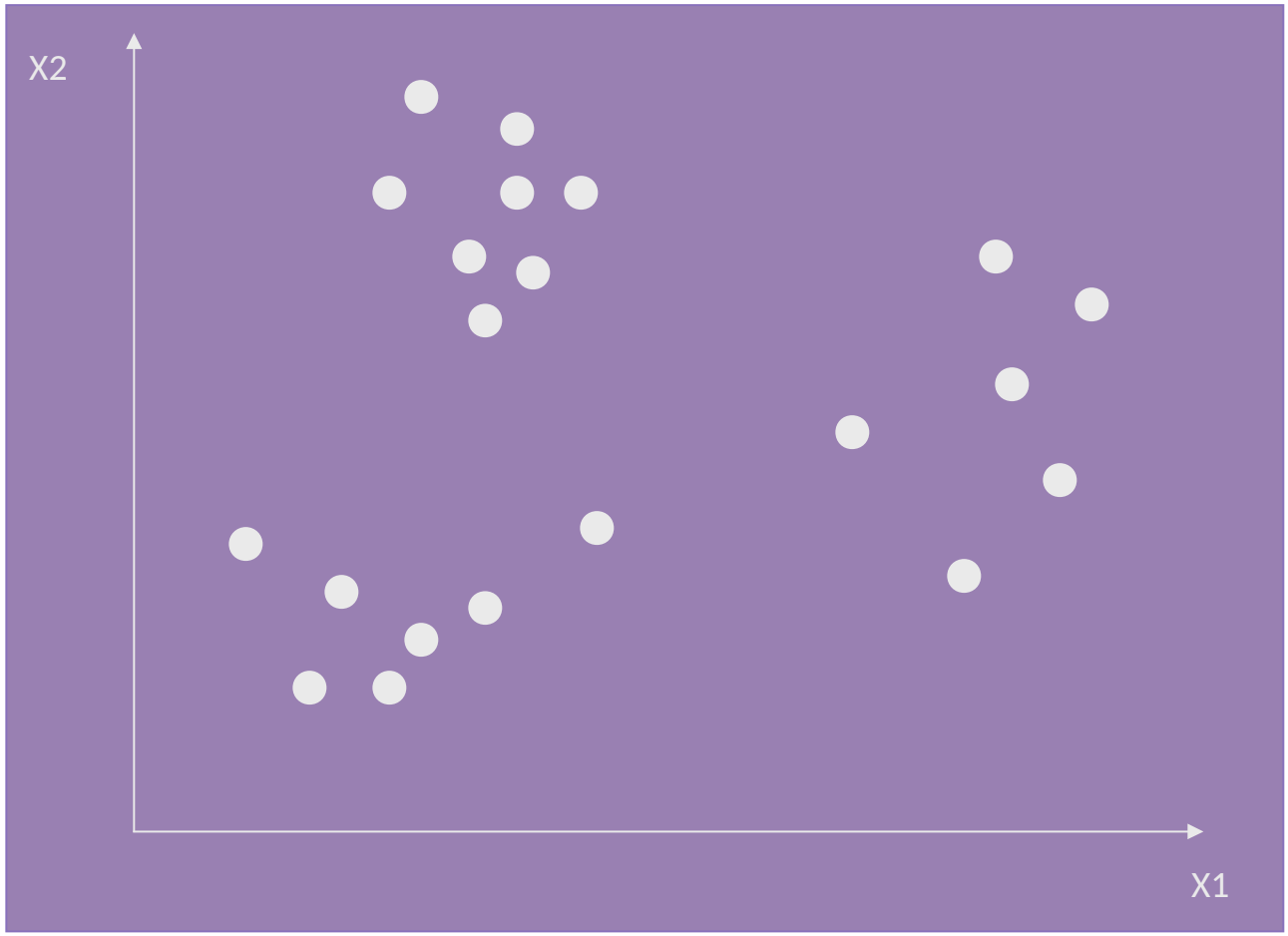
Clasificación – Redes Neuronales



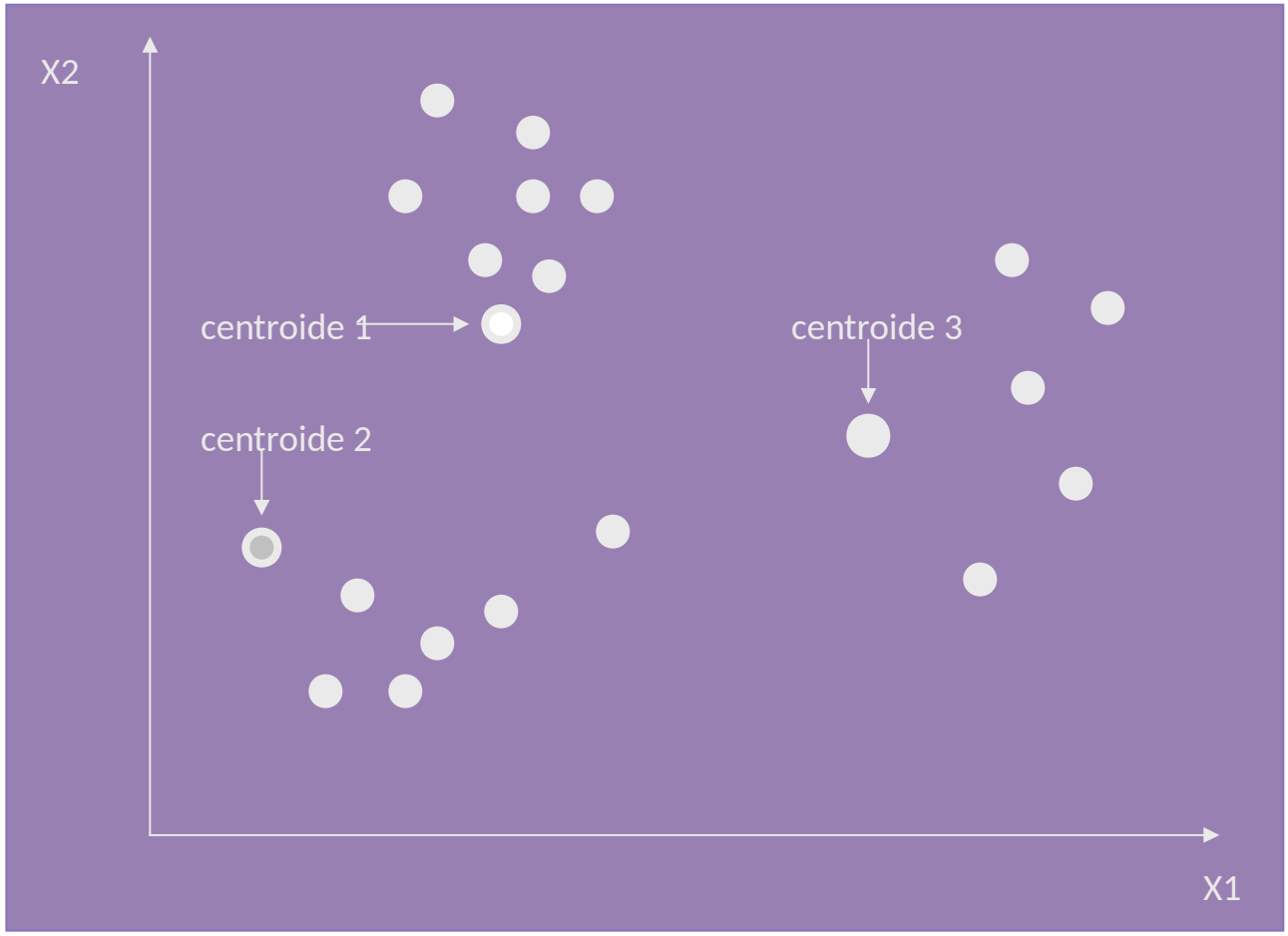
Agrupamiento o Clustering

- **Agrupamiento (Clustering) o Segmentación:** divide a los datos en diferentes grupos, el objetivo es encontrar una agrupación de datos de forma que los datos de un mismo grupo sean muy similares y muy diferentes entre grupos distintos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto

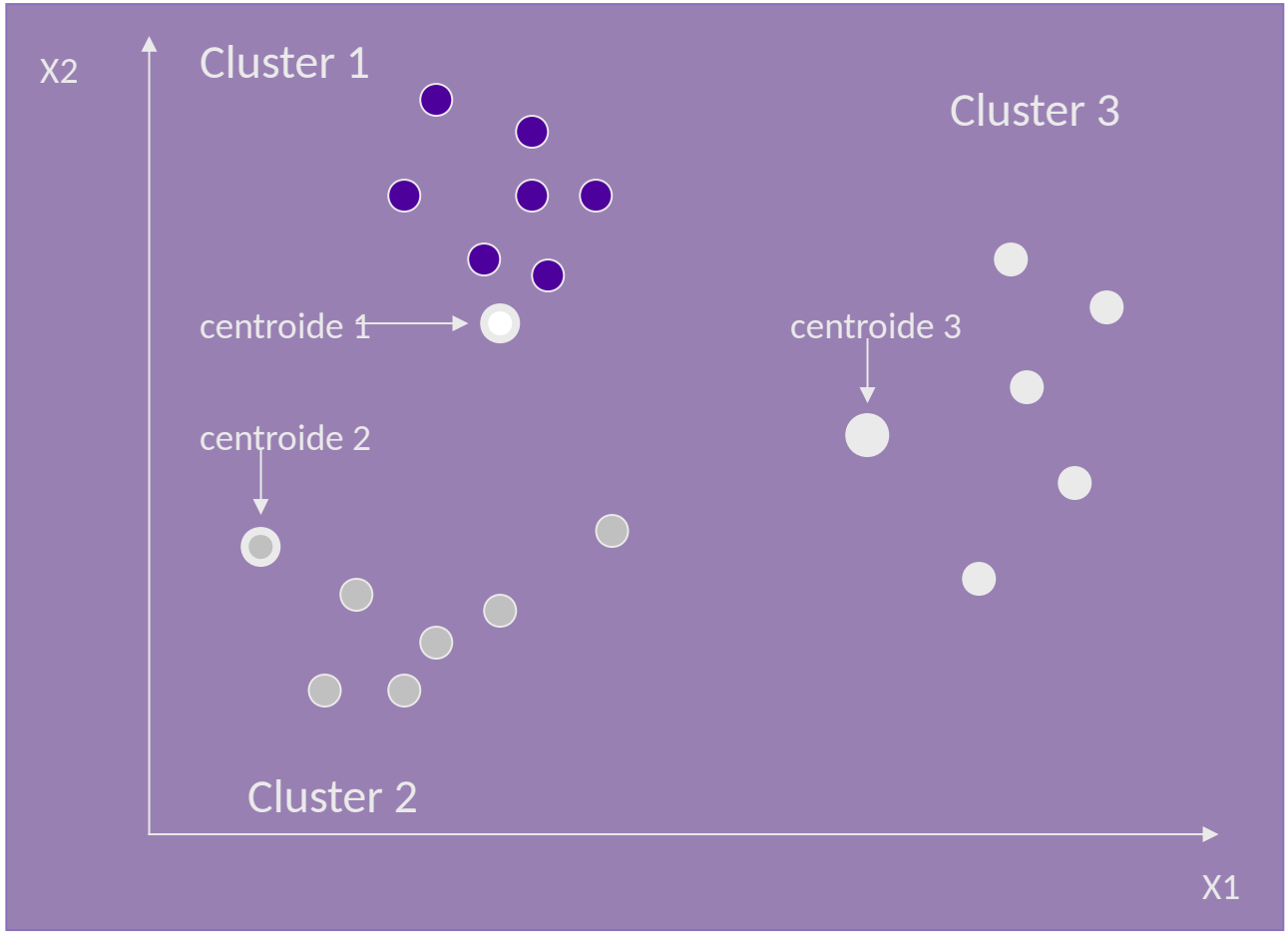
Clustering kMeans



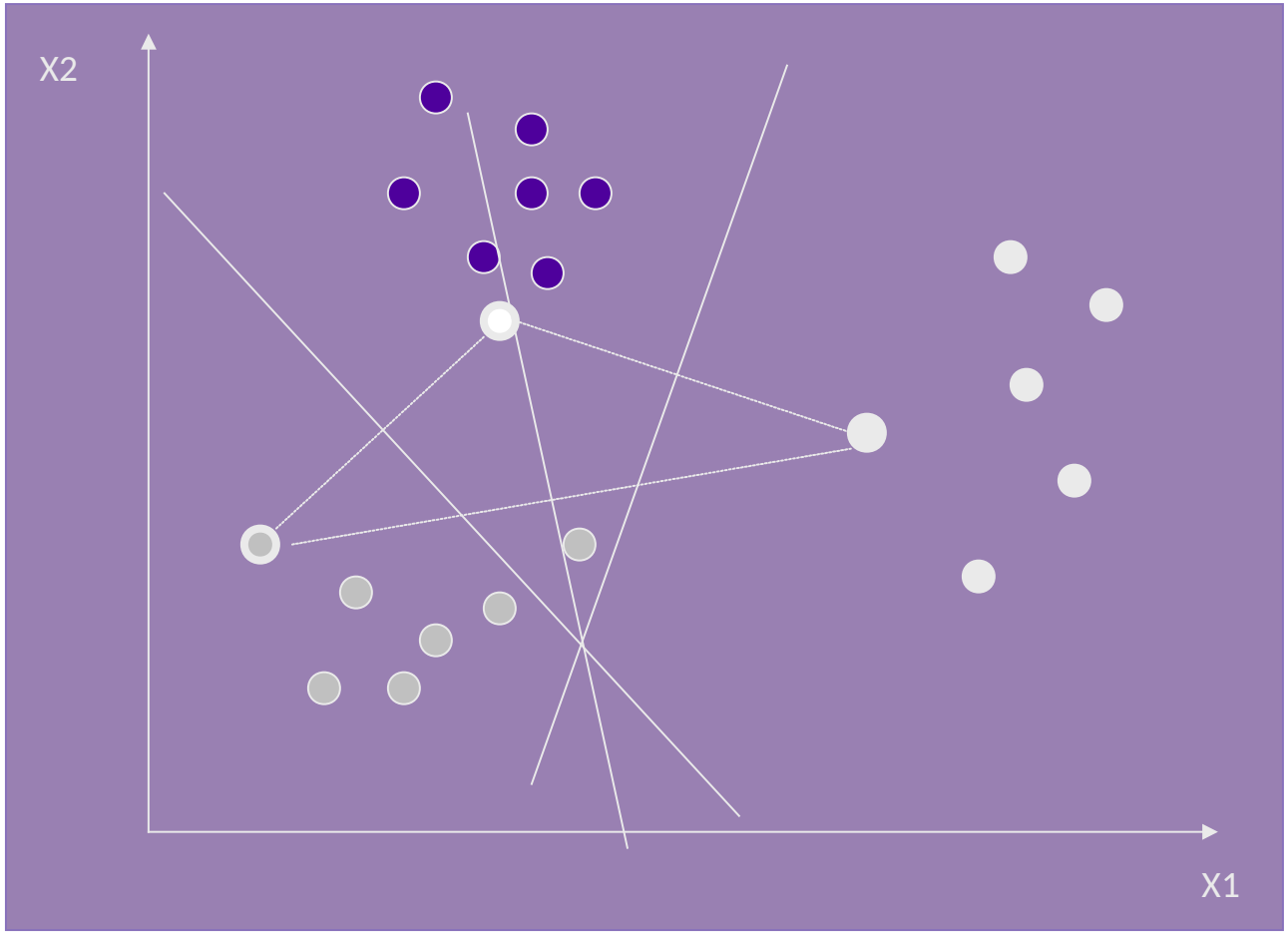
Clustering kMeans



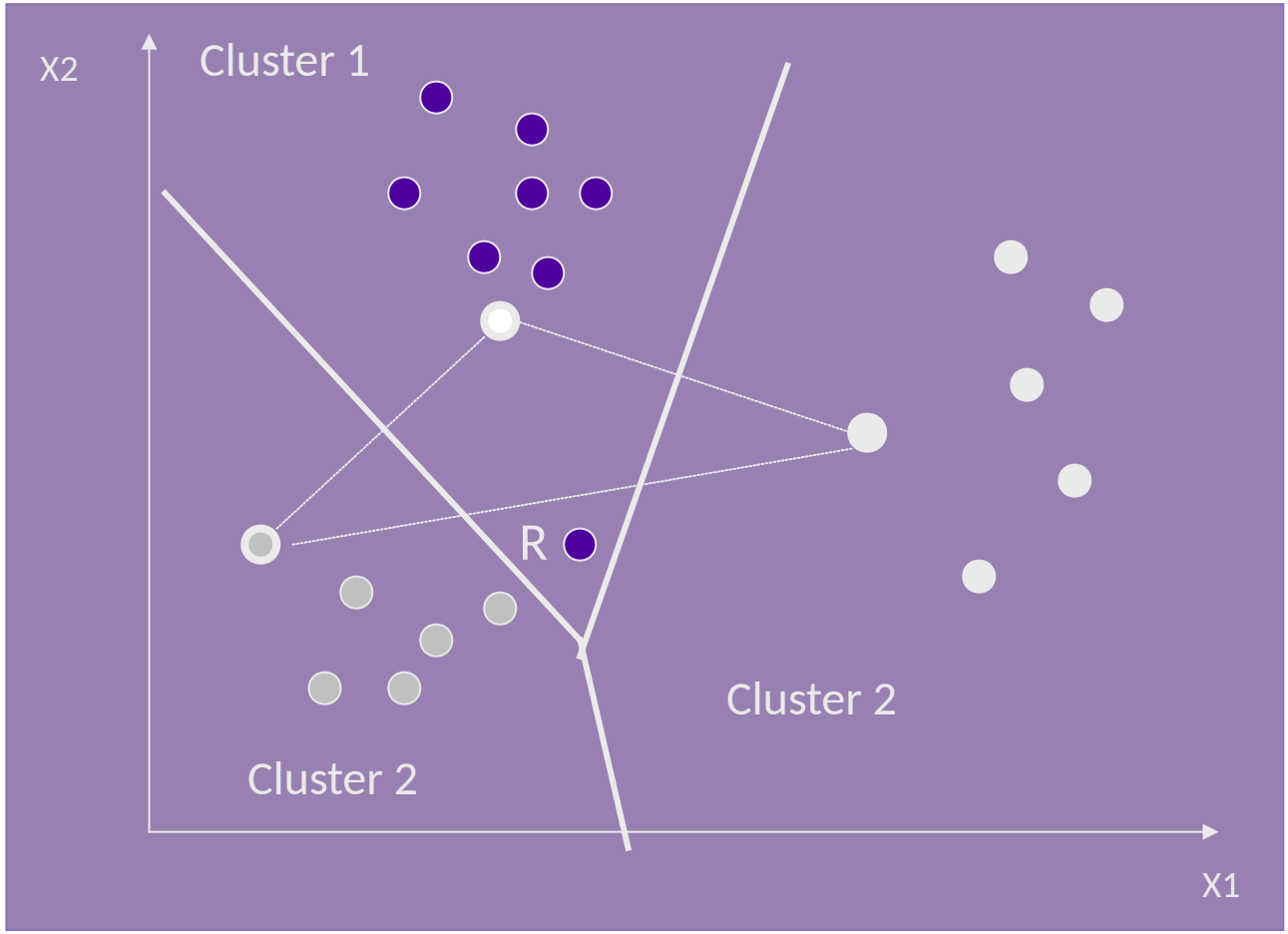
Clustering kMeans



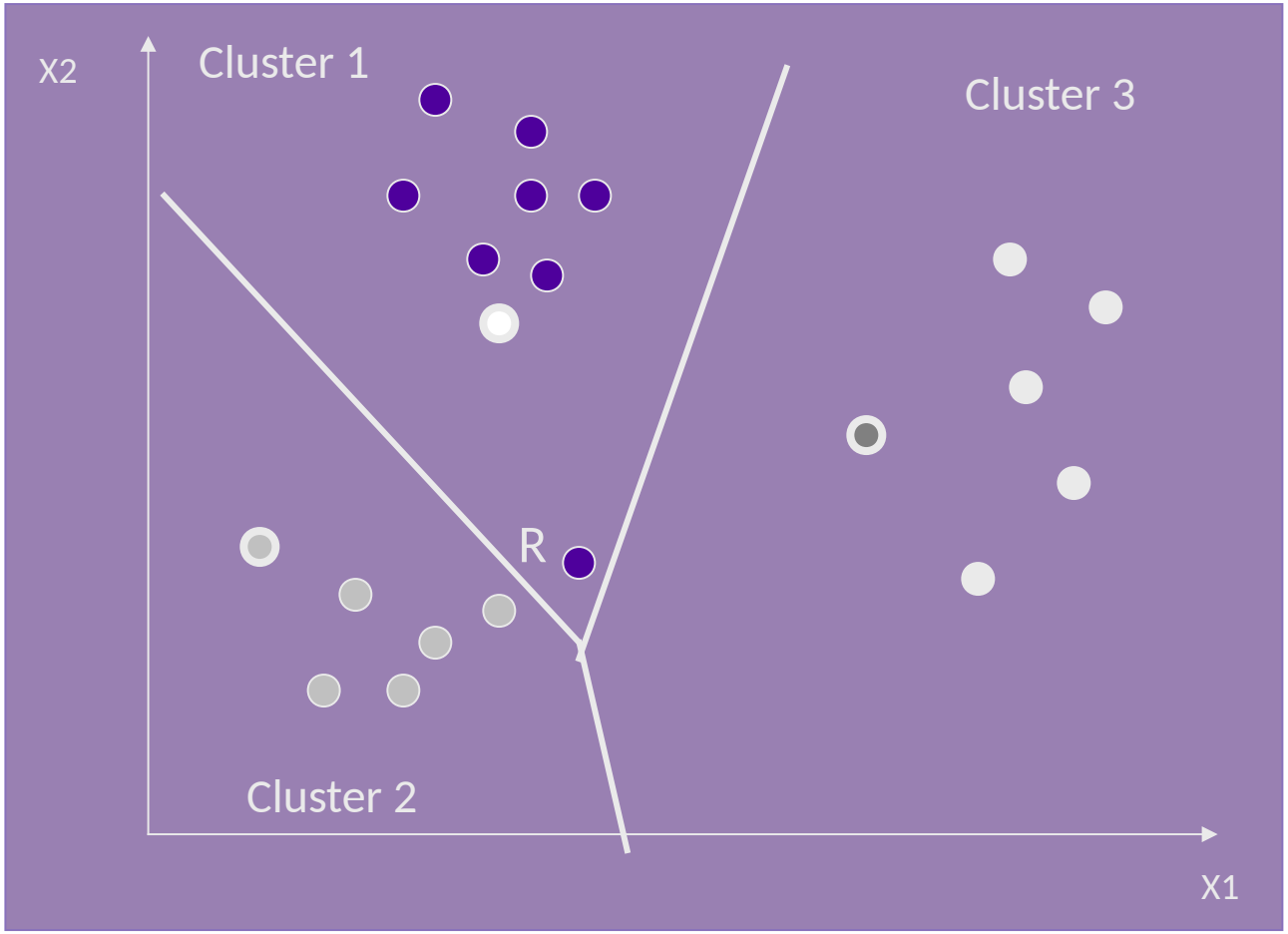
Clustering kMeans



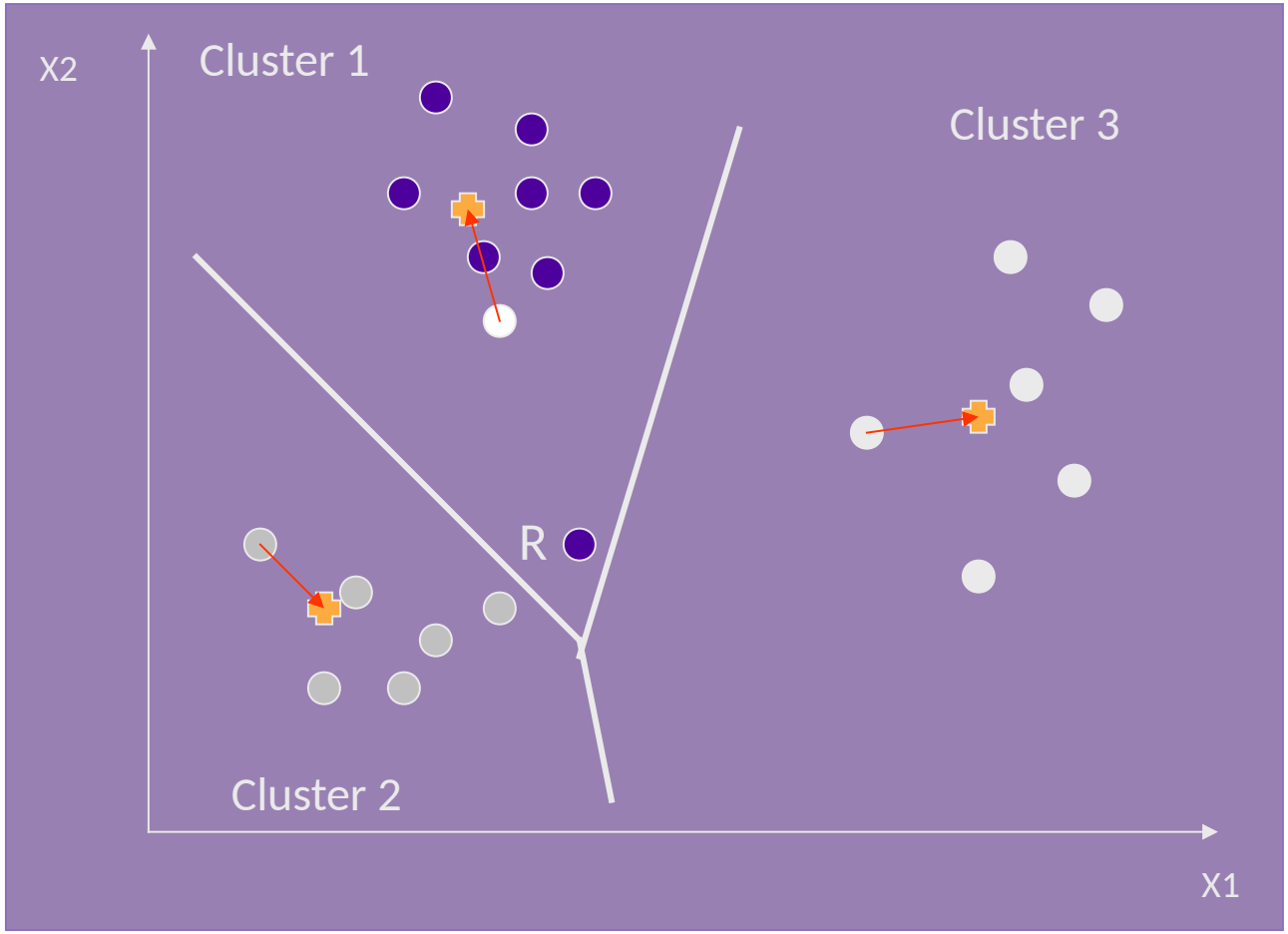
Clustering kMeans



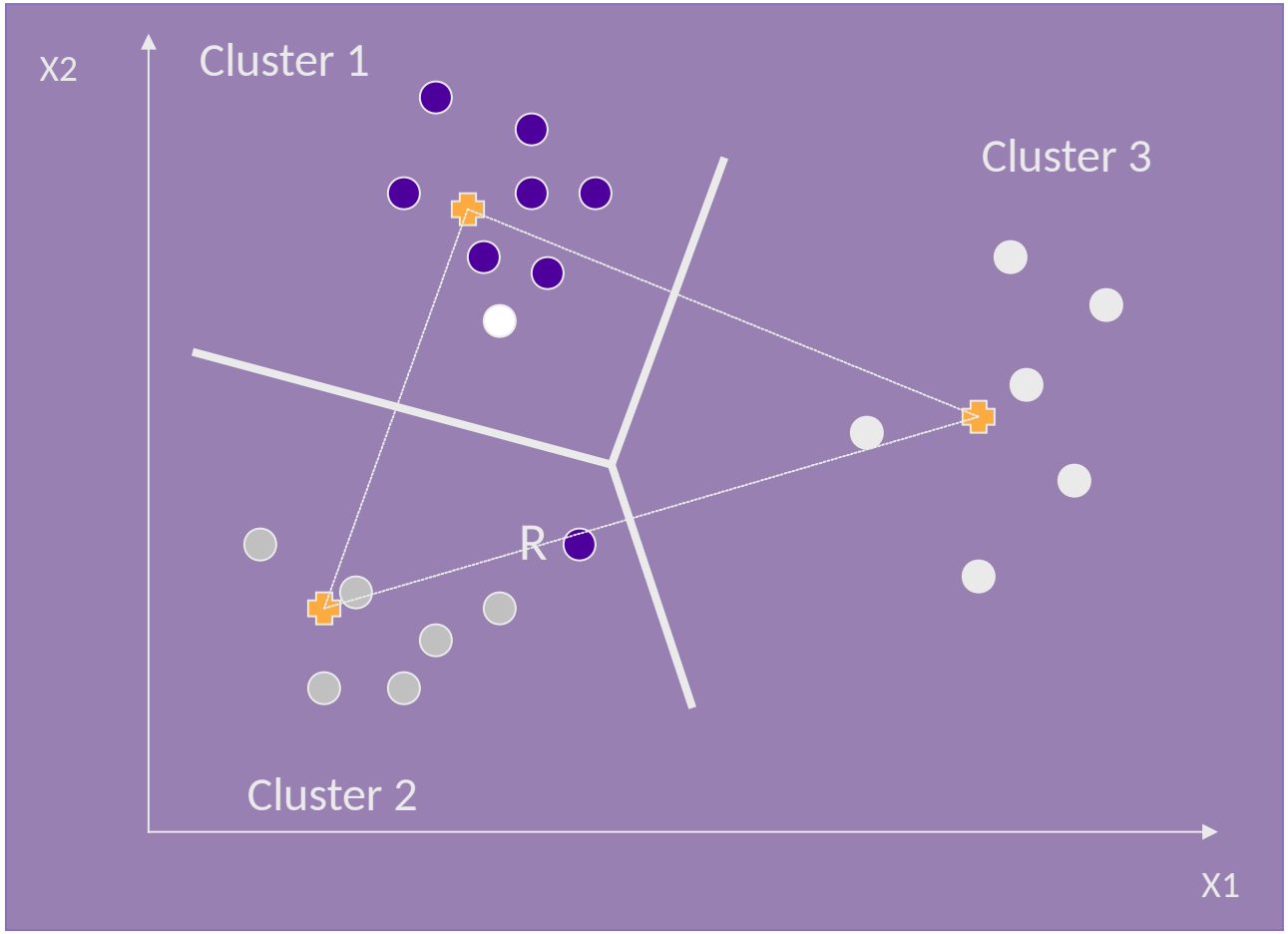
Clustering kMeans



Clustering kMeans

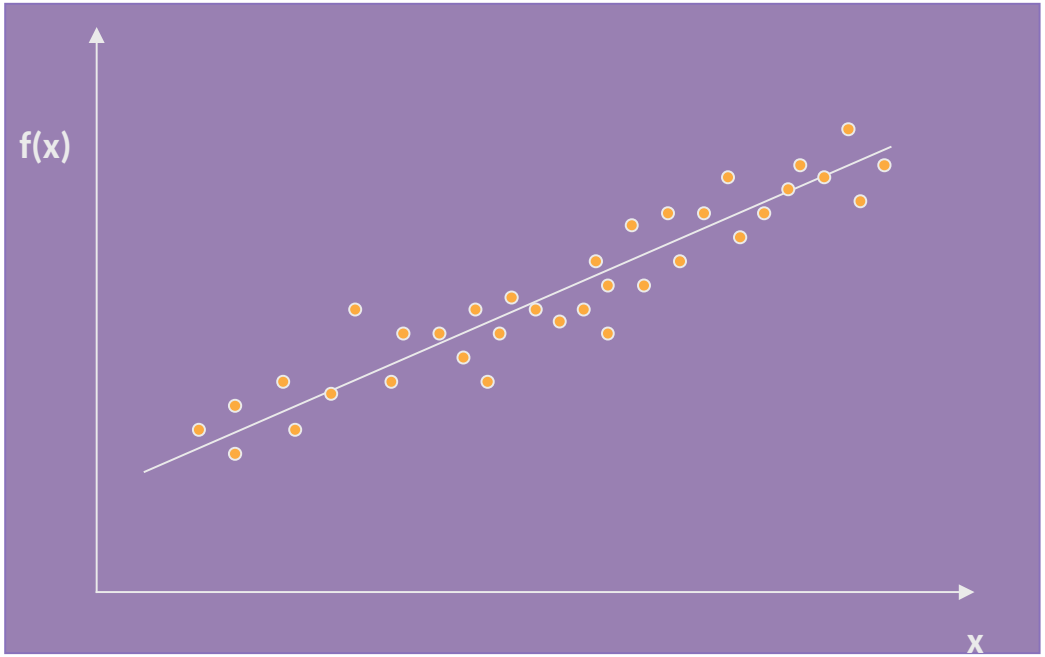


Clustering kMeans



Regresión

- **Tendencias / Regresión:** consiste en adquirir una función que mapee un elemento de dato a una variable de predicción de valor real. Dicho de otro modo, se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable. Ej. se intenta predecir el número de clientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores



Exploratory Data Analysis (EDA)

- **Visualizaciones:** consisten en generar modelos visuales que permitan al usuario sacar meta-conocimientos de los mismos

Análisis Descriptivo + Análisis Exploratorio

- Pie charts, Donut charts, Histograms, KPI, Maps, Heatmaps, Scatter plot, Box plot, etc
  Obs: Analytics que realizamos con el PowerBI

Referencias

- Larose, D. Discovering Knowledge in Data: An introduction to Data Mining. 1st Ed, Wiley. 2005
- Han, J., Kamber, M. Data Mining: Concepts and Techniques. 2nd Ed, Morgan Kaufmann. 2006