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)

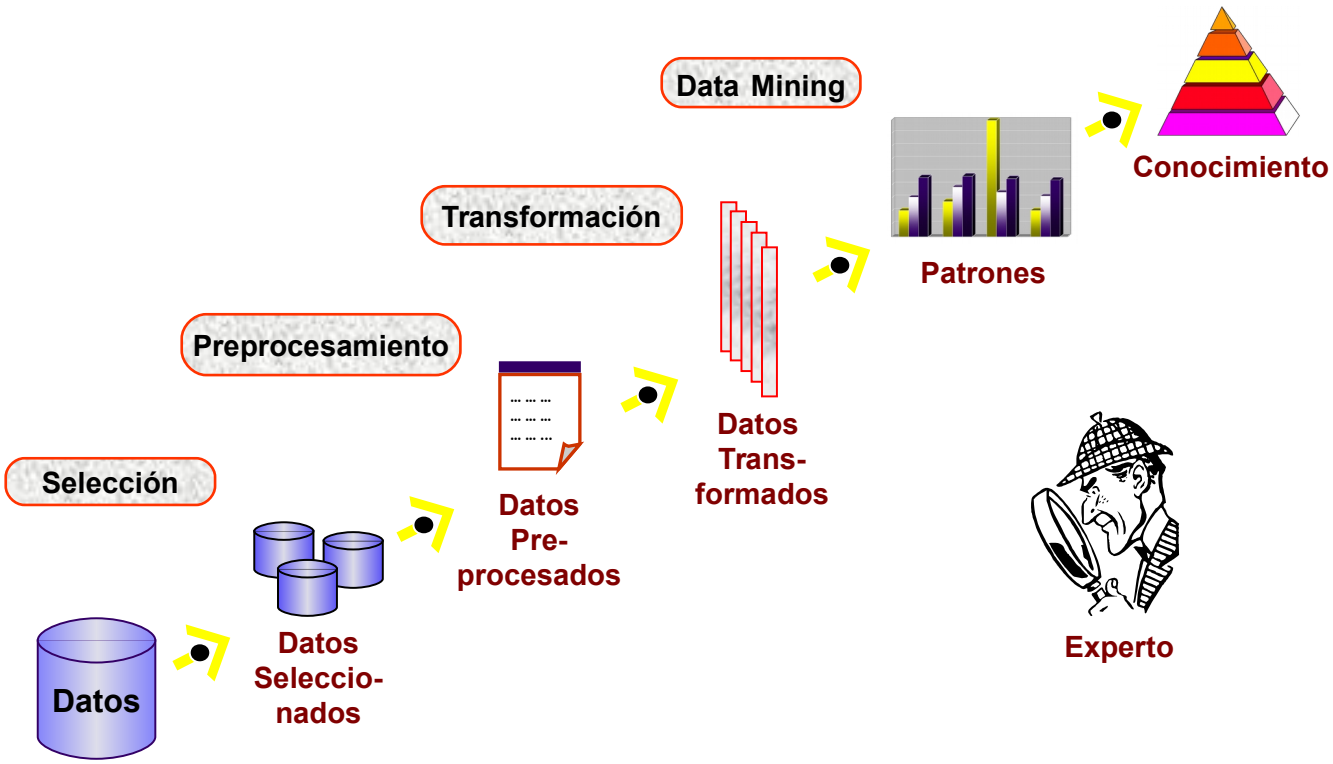
- *Novedoso*: los patrones deben ser novedosos. La novedad puede ser medida con respecto a los cambios en los datos (comparando los valores actuales, con los anteriores o con los esperados) o en el conocimiento (cómo un nuevo hallazgo se relaciona con los anteriores).

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

Data Mining

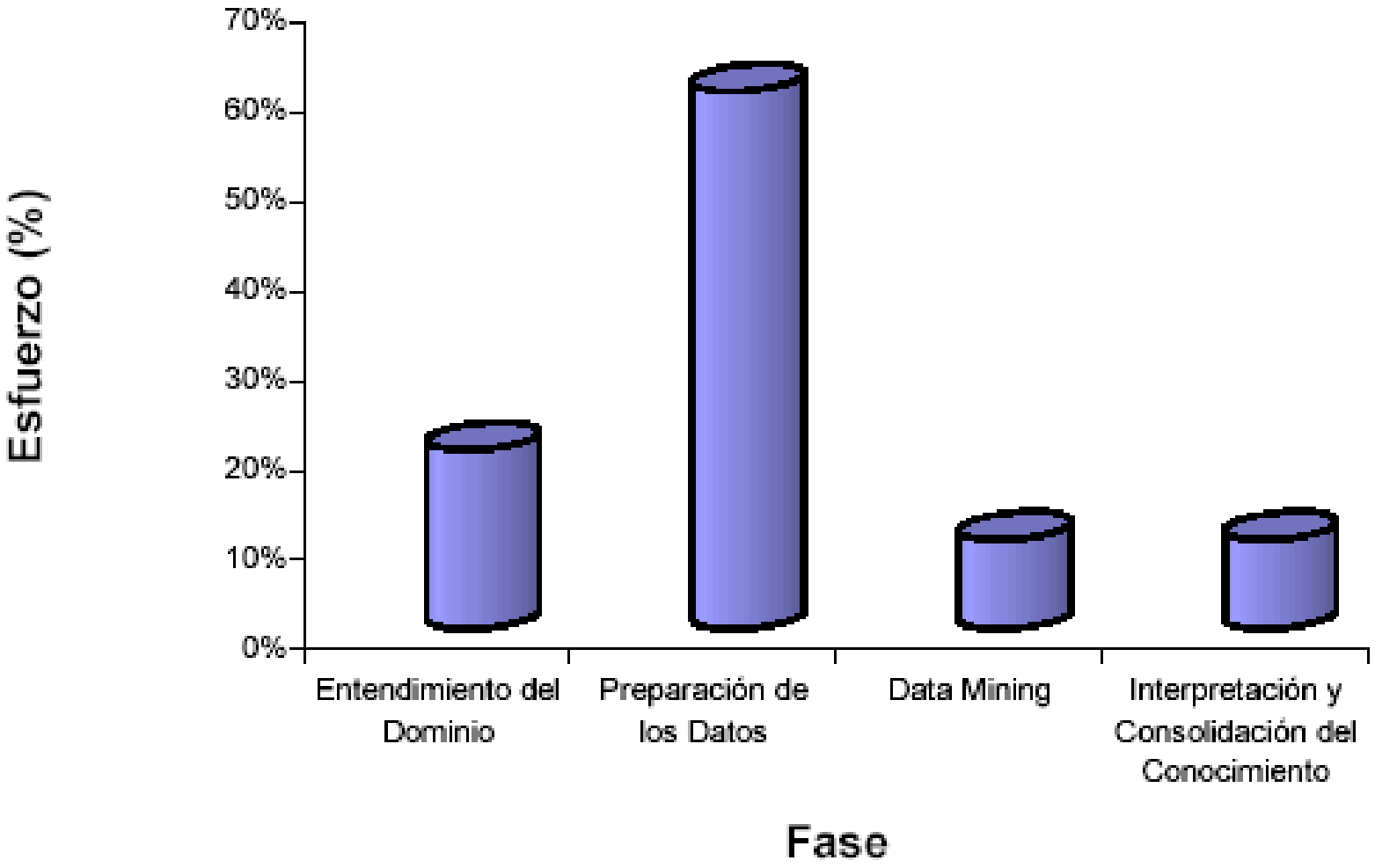
Data Mining es un paso en el proceso del KDD consistiendo de algoritmos particulares que, bajo algunas limitaciones aceptables de eficiencia computacional, produce una enumeración particular de patrones E_j sobre F

Proceso de KDD

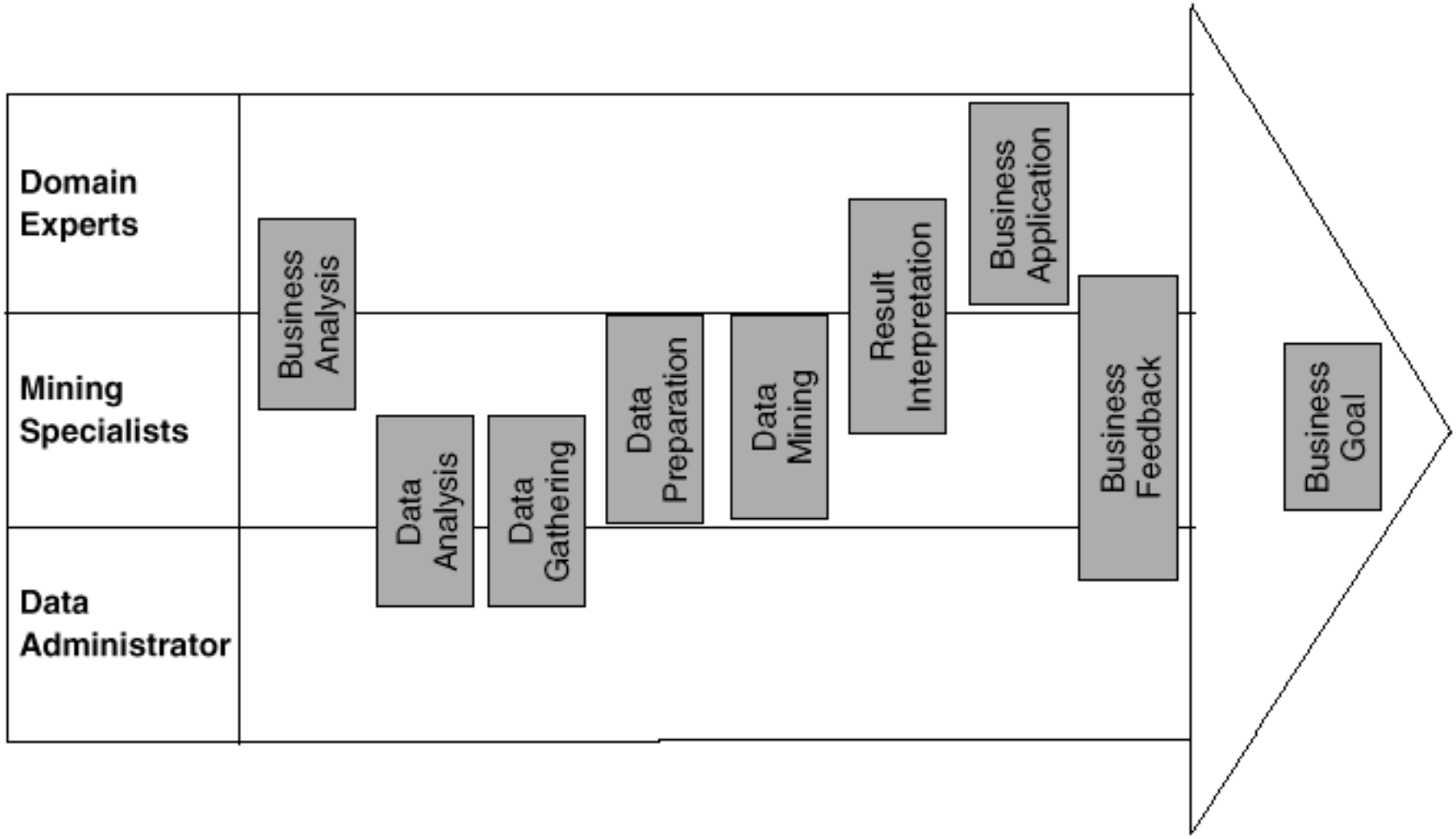


Proceso interactivo e iterativo que envuelve varios pasos y con decisiones a ser tomadas por el usuario

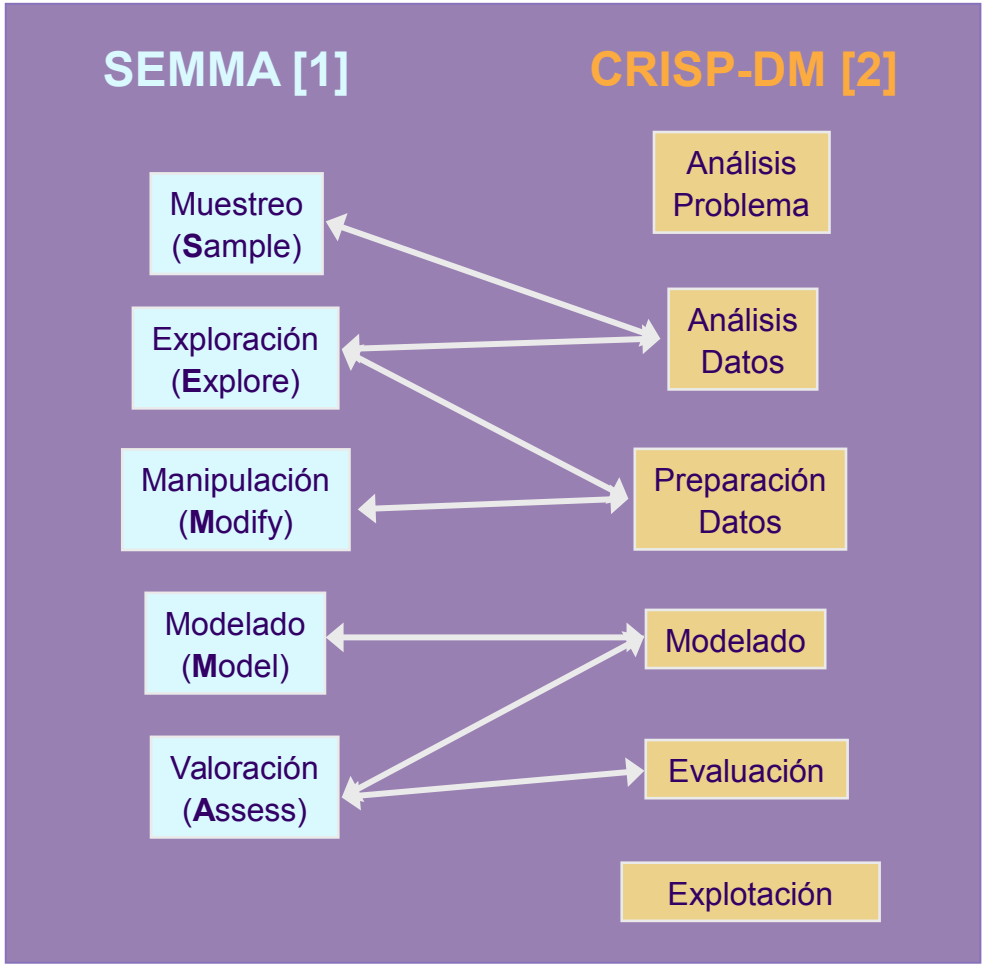
Esfuerzo requerido KDD



Roles en KDD

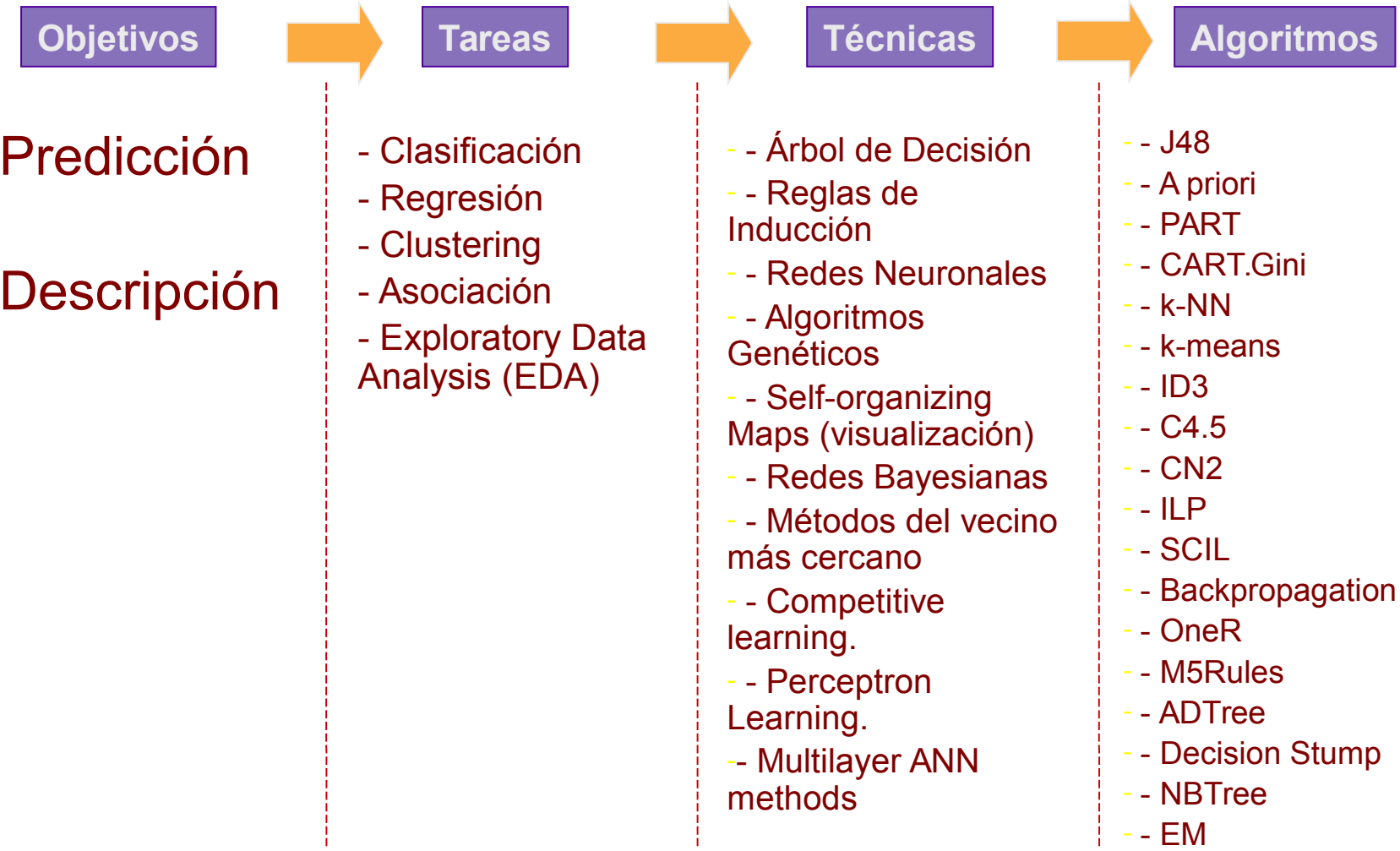

Domain Experts
Mining Specialists
Data Administrator
Business Analysis
Data Analysis
Data Gathering
Data Preparation
Data Mining
Result Interpretation
Business Application
Business Feedback
Business Goal

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)

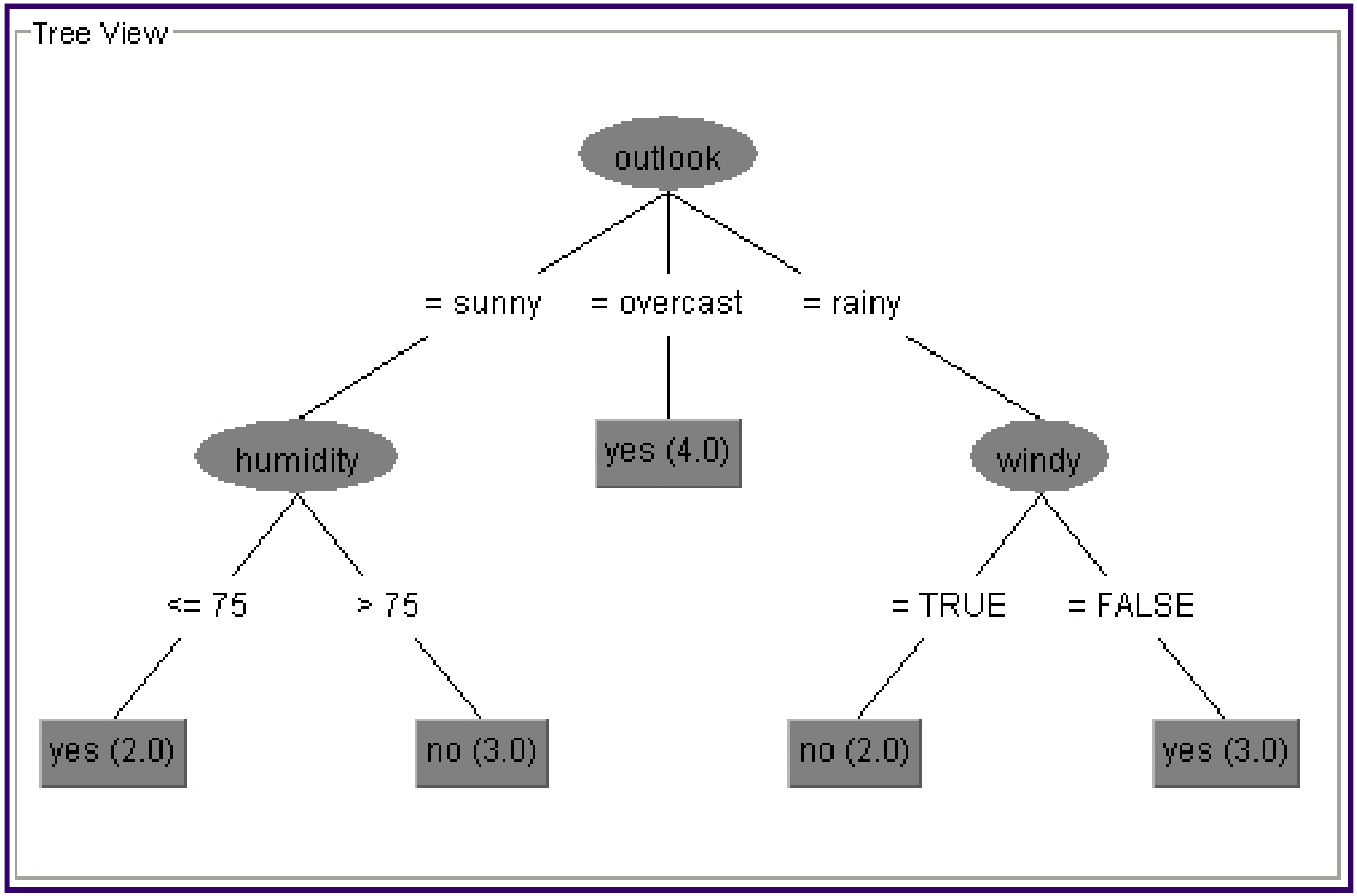
IF windy = TRUE AND
outlook = rainy
THEN play = no (2.0)

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

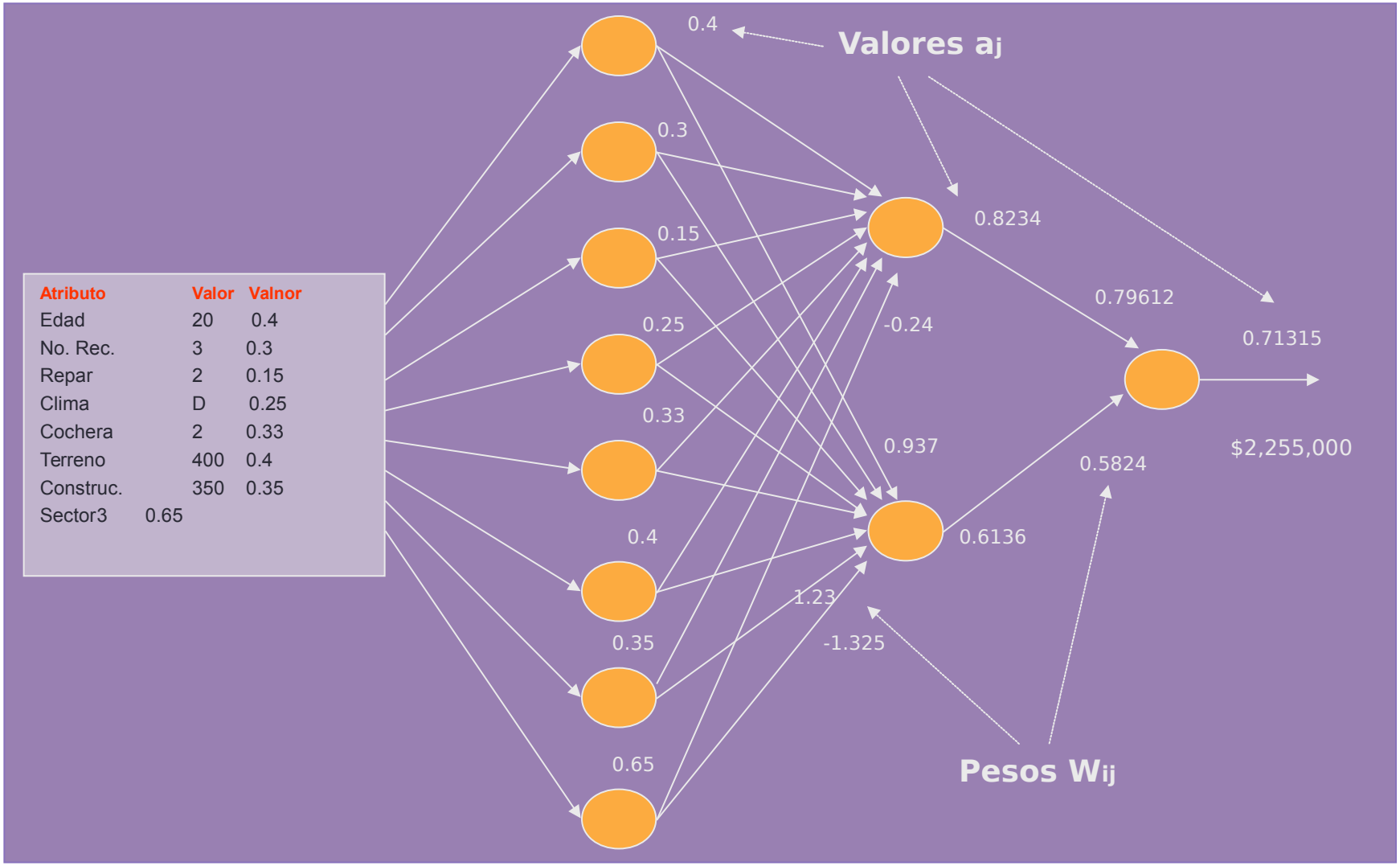
Clasificación

- **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



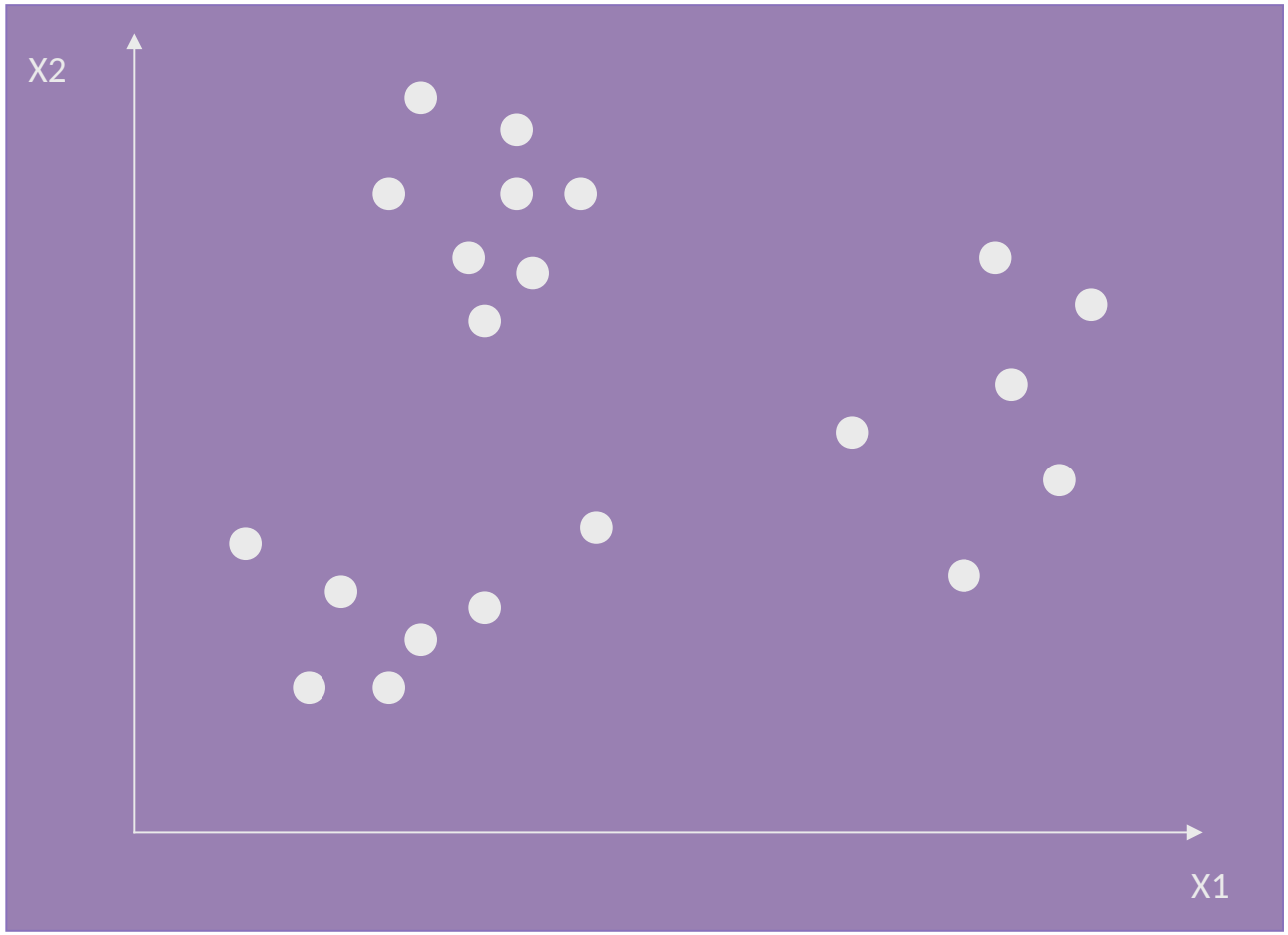
Clasificación – Redes Neuronales



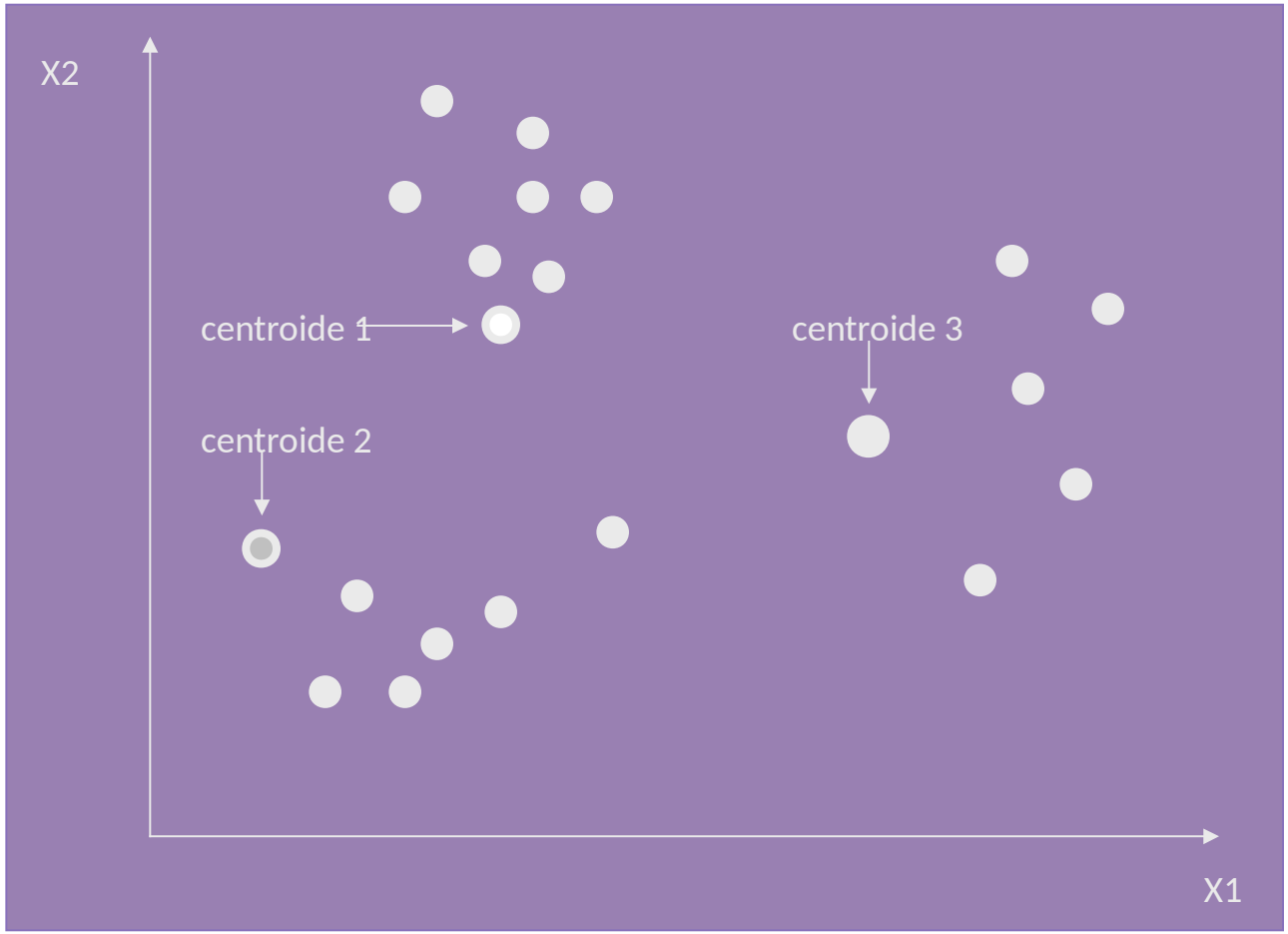
Agrupamiento o Clustering

- **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

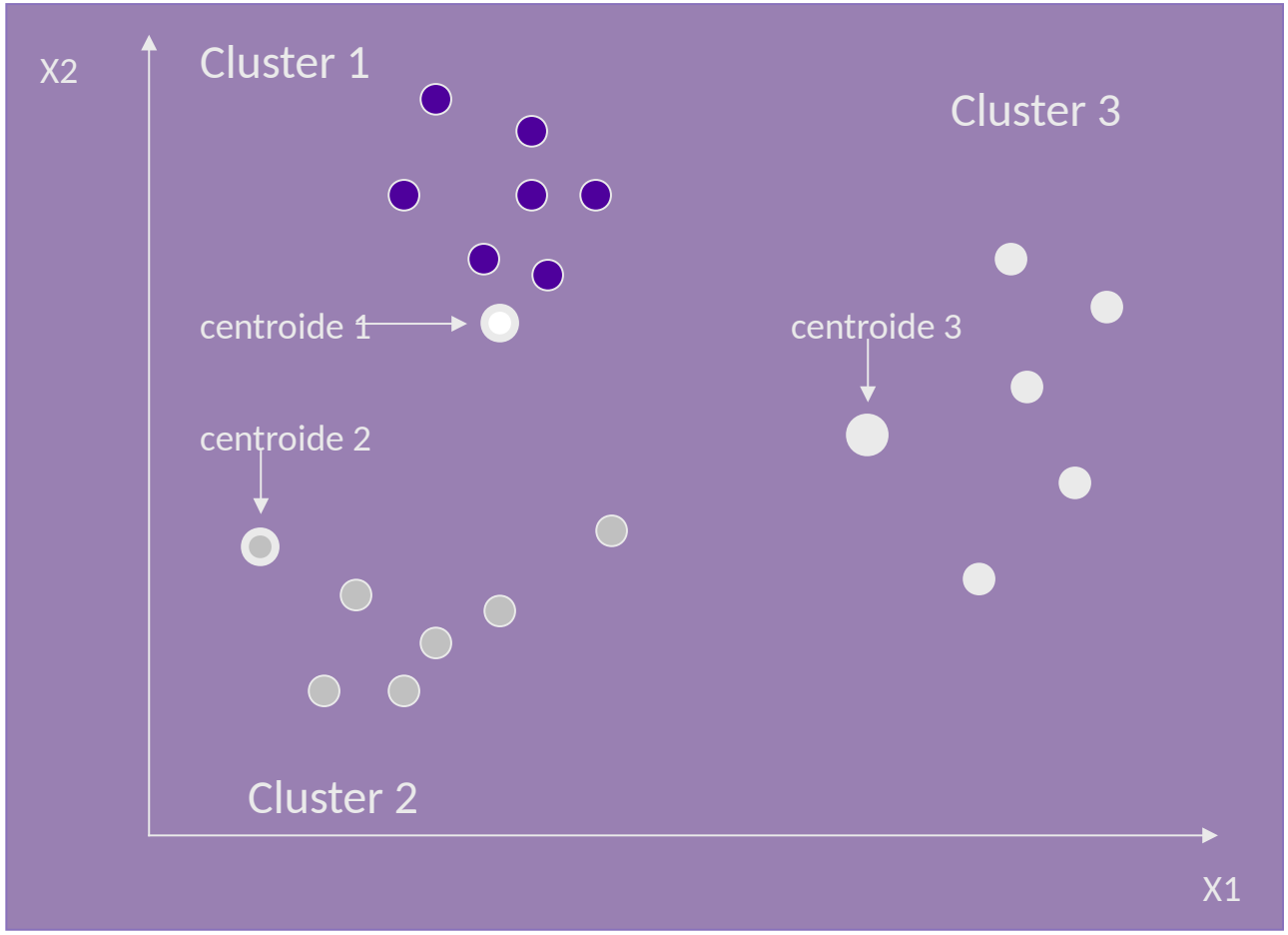
Clustering kMeans



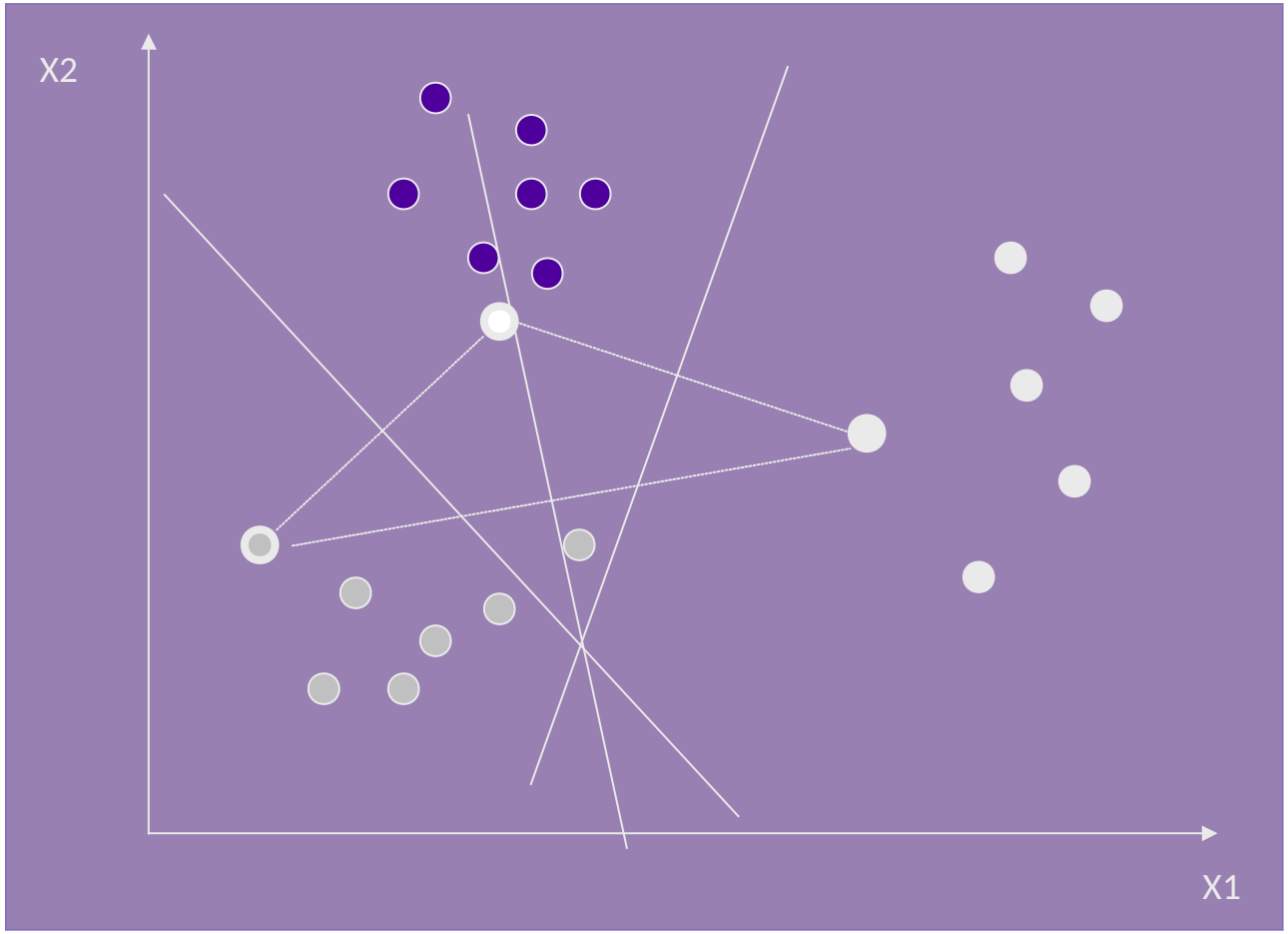
Clustering kMeans



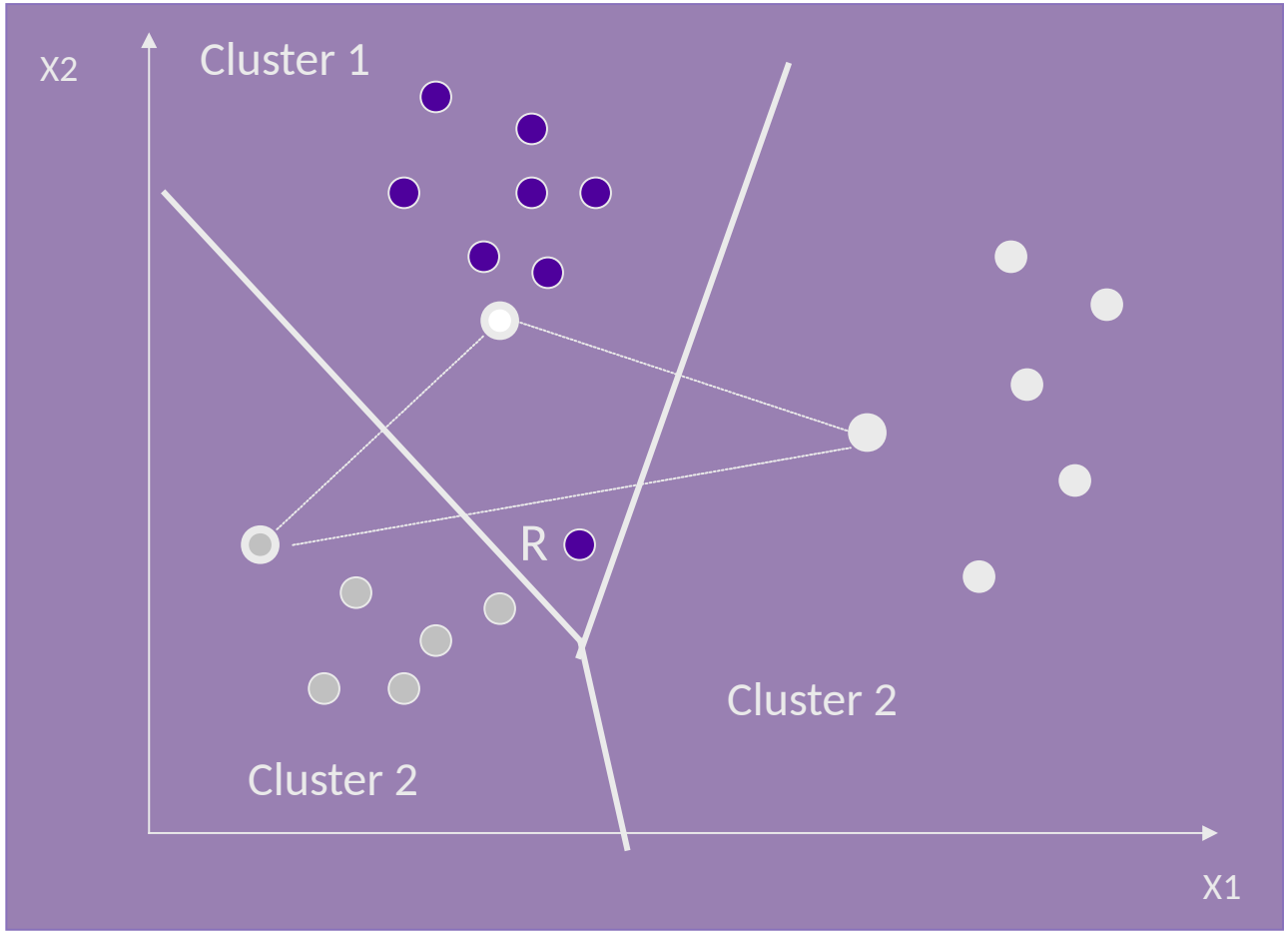
Clustering kMeans



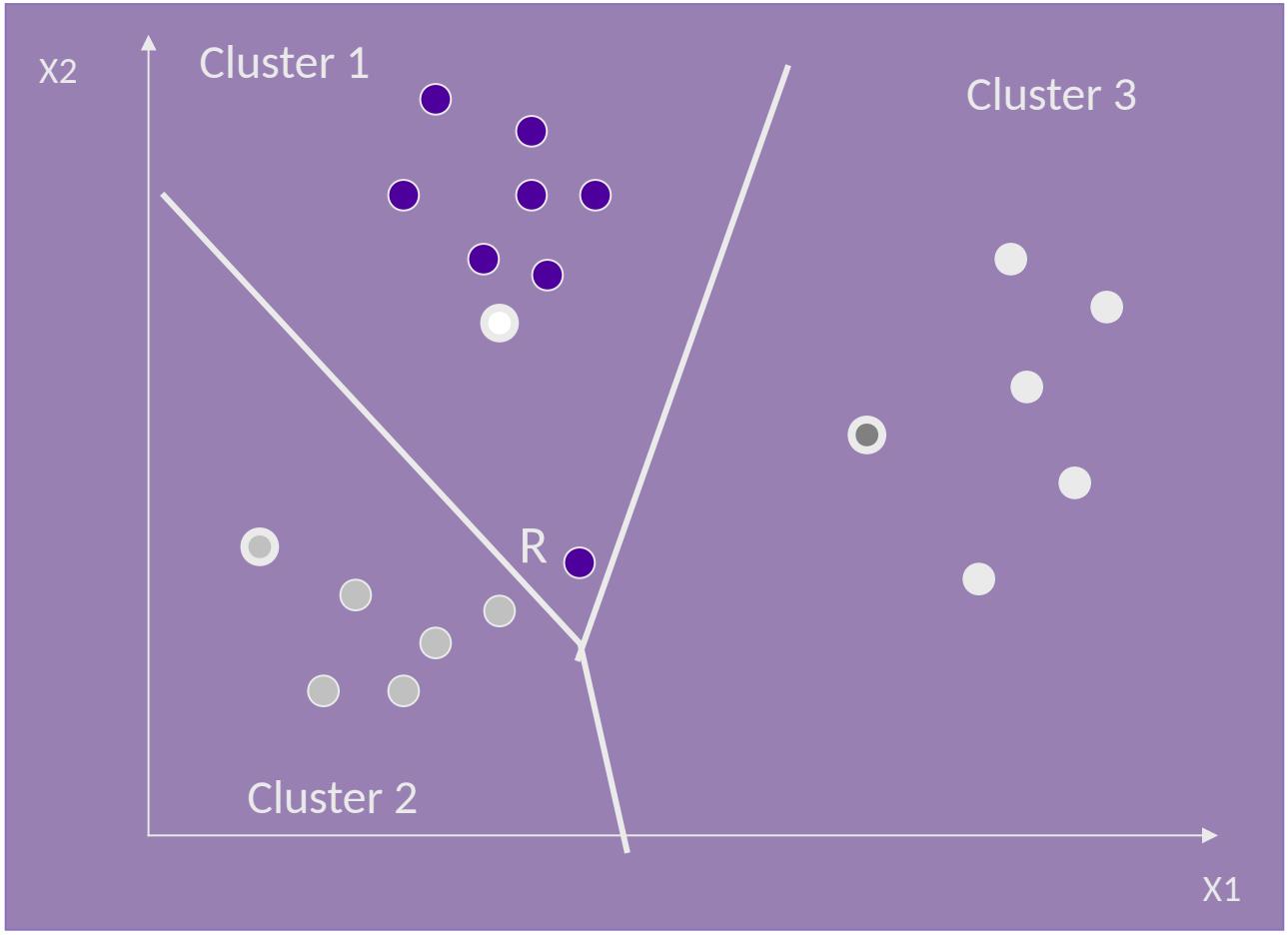
Clustering kMeans



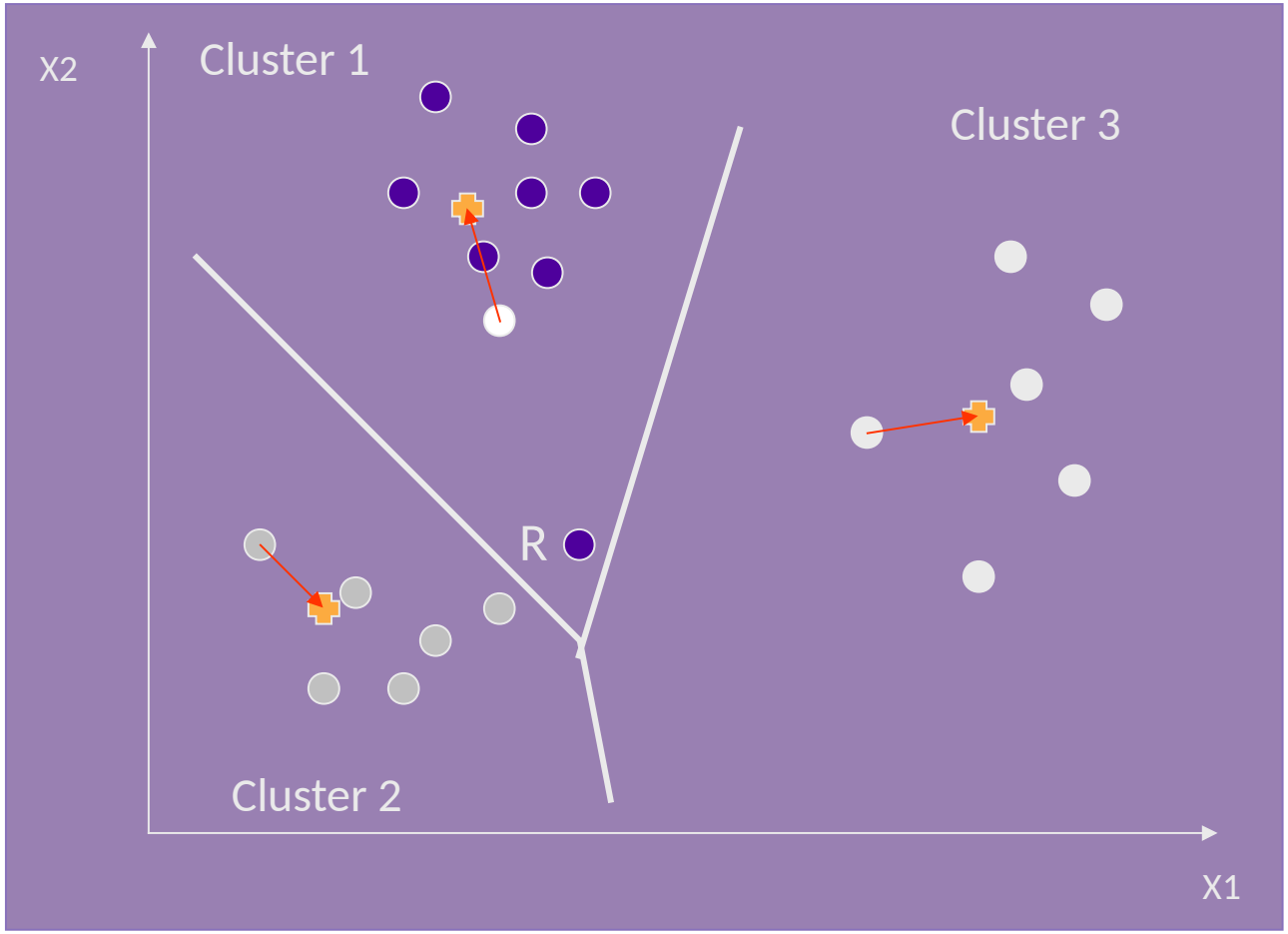
Clustering kMeans



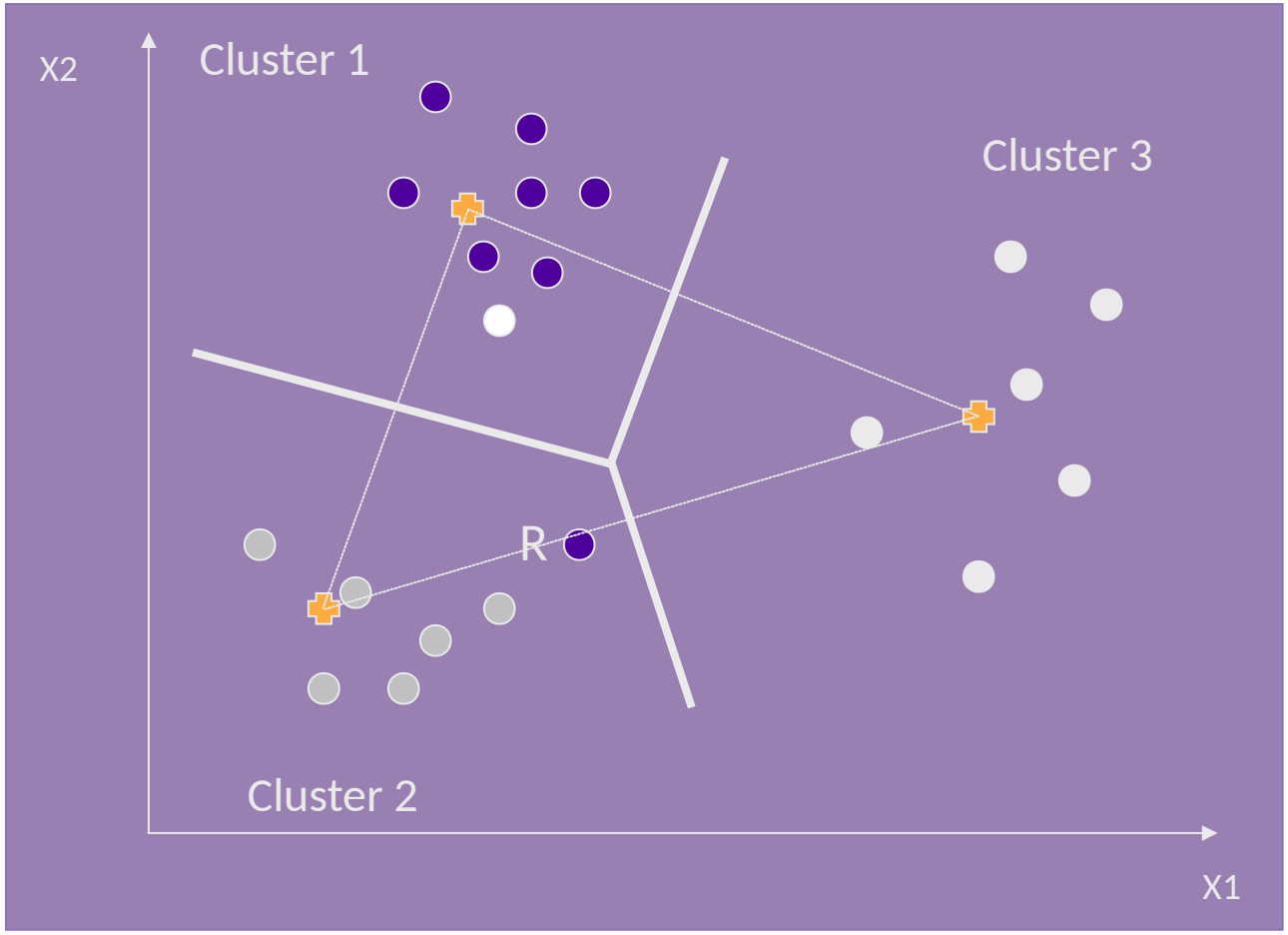
Clustering kMeans



Clustering kMeans

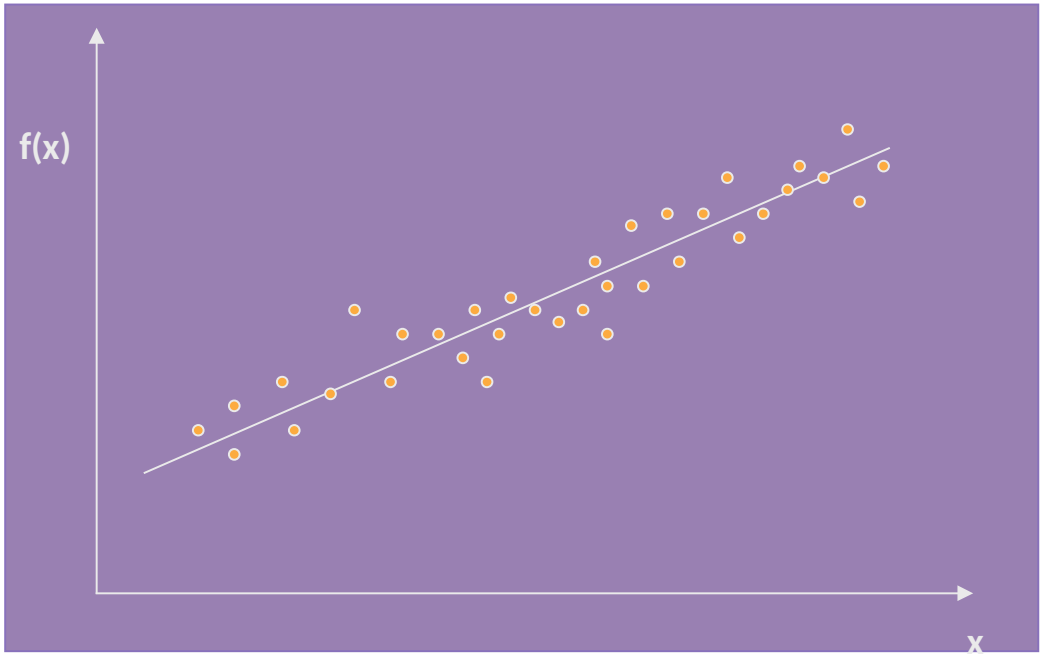


Clustering kMeans



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

Tipos de Machine Learning (ML)

- **Aprendizaje Supervisado (supervised learning)**

Encontrar una función de aprendizaje que mapea inputs a outputs, utilizando como entrenamiento ejemplos conocidos de inputs y sus correspondientes outputs. Ej: Regresión en series temporales

- **Aprendizaje No Supervisado (unsupervised learning)**

Encontrar una función que permita describir el input (categorizar) a partir de inputs no categorizados. No es posible tener una medida de la efectividad del modelo ya que no es conocido el output esperado. Ej: Anomaly Detection

Tipos de Machine Learning (ML)

- **Aprendizaje por refuerzo (reinforcement learning)**

El modelo cuenta con una tabla de aprendizaje donde se encuentran los pares inputs/condiciones posibles asociados a sus correspondientes outputs/acciones con una probabilidad de que tan deseable es para el modelo utilizar el par. Se requiere contar con una función de evaluación que pueda determinar el grado de preferencia de un par sobre otro y una regla de actualización de probabilidades en la tabla. De esta forma el modelo puede iniciar con probabilidades iniciales uniformemente distribuidas, va observando inputs y modificando sus probabilidades de acuerdo a su función de evaluación (aprendiendo de la experiencia luego de observar los inputs, fase de entrenamiento)

Ej: Bots de juegos

Tipos de variables

Según su influencia:

- ***Variables Independientes (x)*:** sus valores no están en dependencia de otras variables y creemos influyen en el fenómeno estudiado. Ej: clima, precios de canasta básica
- ***Variables dependientes (y)*:** sus valores están determinados por otras variables generalmente independientes. Ej: número de casos de dengue, IPC,

Según la Parametrización:

- ***Variable Input*:** Típicamente las variables independientes que influyen o creemos influyen en el modelo
- ***Variable Target*:** La variable de estudio que es influida por las variables independientes

Análisis Predictivo

- **Regresión Lineal (polinomio de grado 1)**

  $a_0 + a_1.x$

- **Regresión Polinomial (polinomio de grado n)**

  Ej. Grado 5: $a_0 + a_1.x + a_2.x^2 + a_3.x^3 + a_4.x^4 + a_5.x^5$

- **Multi Layer Perceptron (MLP – Red Neuronal)**
  - *Input Layer*: 1 neurona por variable independiente
  - *Hidden Layers*: Parametrizable cantidad de hidden layers y neuronas
  - *Ouput Layer*: 1 neurona con el resultado de la variable target

Estadística básica

- **Varianza (X = {x1,x2,...,xn})**

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

- **Desviación estándar (X = {x1,x2,...,xn})**

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- **Covarianza (XY, con X = {x1,x2,...,xn} e Y = {y1,y2,...,yn})**

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Numerical Scoring

- **Coeficiente de correlación de Pearson (*ρ*)**

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- **Coeficiente R^2**

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$$

- **Mean Square Error (MSE)**

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **Root Mean Square Error (RMSE o RMSD)**

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}$$

- **Mean Absolute Error (MAE)**

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Training / Validation Sets

- *Training Dataset*: datos utilizados para el entrenamiento del modelo. Se conocen los valores de la variable target y estos son proveídos a la técnica para el cálculo del Scoring y ajuste de parámetros de la técnica

- *Validation Dataset*: datos utilizados para la validación de la efectividad del modelo. Se conocen los valores de la variable target, pero estos no son proveidos al modelo, solo son proveidos para el cálculo final del Scoring y evaluación así de la efectividad del modelo y ajuste de hiperparámetros de la técnica

Selección del validation set

- *Método Holdout:* parte de los datos de la muestra los apartamos y utilizamos como set de validación. Ej: 70% / 30%, 2011-2016 / 2017

- *Método k-fold Cross-validation:* la muestra es particionada en k submuestras de igual tamaño, 1 subgrupo se utiliza para validación y k-1 para entrenamiento. El experimento se repito k veces de manera a que cada subgrupo al menos 1 vez pertenezca el conjunto de validación

Entropía de la información

- *Entropía de Shannon*: cantidad de información promedio que contienen los símbolos utilizados. Aquellos símbolos con menor probabilidad de aparición aportan más información. Ej: en un texto en español las palabras “que” “el” que tienen alta probabilidad de ocurrencia y por ende aportan baja información, si las eliminamos igual podría ser entendible el texto.

- En un lenguaje binario={0,1}, cada valor posible posee un 1/2 (50%) de probabilidad de ocurrencia, aportan la misma cantidad de información.

Entropía de la información

$$c_i = \log_2(k) = \log_2[1/(1/k)] = \log_2(1/p) = \underbrace{\log_2(1)}_{=0} - \log_2(p) = -\log_2(p)$$

$$H = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - \ldots - p_k \log_2(p_k) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Es la suma de la cantidad de información aportada por cada símbolo ponderada por su probabilidad de ocurrencia

Ejemplo de Entropía

- Asumamos una variable X con 3 valores posibles ALTO, MEDIO, BAJO con probabilidades de ocurrencia de 1/3 (33%), 1/2 (50%) y 1/6 (17%) respectivamente, la entropía H de la variable sería:

  $H(X) = 1/3 \log_2(3) + 1/2 \log_2(2) + 1/6 \log_2(6) =$ **1,46**

- *Valor Máximo de H*, distribución uniforme, 1/3 (33%), 1/3 (33%), 1/3 (33%):

  $H(X) = 1/3 \log_2(3) + 1/3 \log_2(3) + 1/3 \log_2(3) =$ **1,58**

- *Valor Mínimo de H*, distribución sesgada, 0%, 100%, 0%:

  $H(X) = 0 + 1 \log_2(1) + 0 =$ **0**

Métricas de particionamiento Decision Tree's

- *Gain Ratio:* Utiliza la entropía *H* como métrica, seleccionando una variable que favorezca particiones de con baja entropía

- *Gini Index*: $$I_G(p) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} {p_i}^2$$

  Gini Impurity, mide la probabilidad de un valor de la variable de ser elegida ponderado por la probabilidad de ser incorrectamente clasificado. Selecciona una variable que favorezca particiones con baja impureza Gini

Métricas de particionamiento Decision Tree's

- *Gain Ratio:* Utiliza la entropía *H* como métrica, seleccionando una variable que favorezca particiones de con baja entropía

- *Gini Index*: $$I_G(p) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} {p_i}^2$$

  Gini Impurity, mide la probabilidad de un valor de la variable de ser elegida ponderado por la probabilidad de ser incorrectamente clasificado. Selecciona una variable que favorezca particiones con baja impureza Gini

Ejemplo de Gini Index

- X={ALTO (33%), MEDIA (50%), BAJA (17%)}
  - GINI = 1 – $(0,33^2+0,5^2+0,17^2)$ = **0,6122**

- *Valor máximo*: **X={ALTO (33%), MEDIA (33%), BAJA (33%)}**

  GINI = 1 – $(0,33^2+0,33^2+0,33^2)$ = **0,6666**

- *Valor Mínimo*: **X={ALTO (0%), MEDIA (100%), BAJA (0%)}**

  GINI = 1 – $(0^2+1^2+0^2)$ = **0**

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