Machine Learning & Data Mining

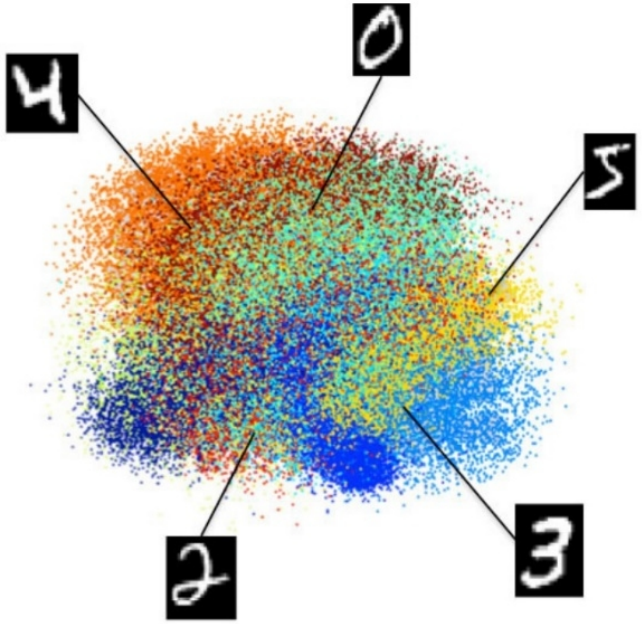
Ing. Julio Paciello

juliopaciello@cds.com.py

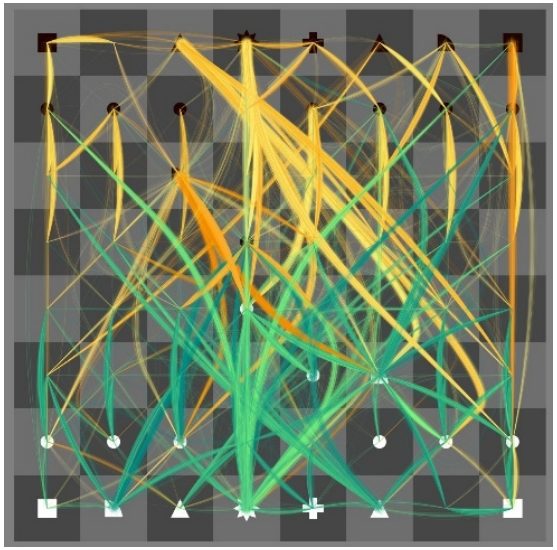
Machine Learning (ML)

Rama de la IA que trabaja con algoritmos que permiten a las máquinas aprender

Supervisado

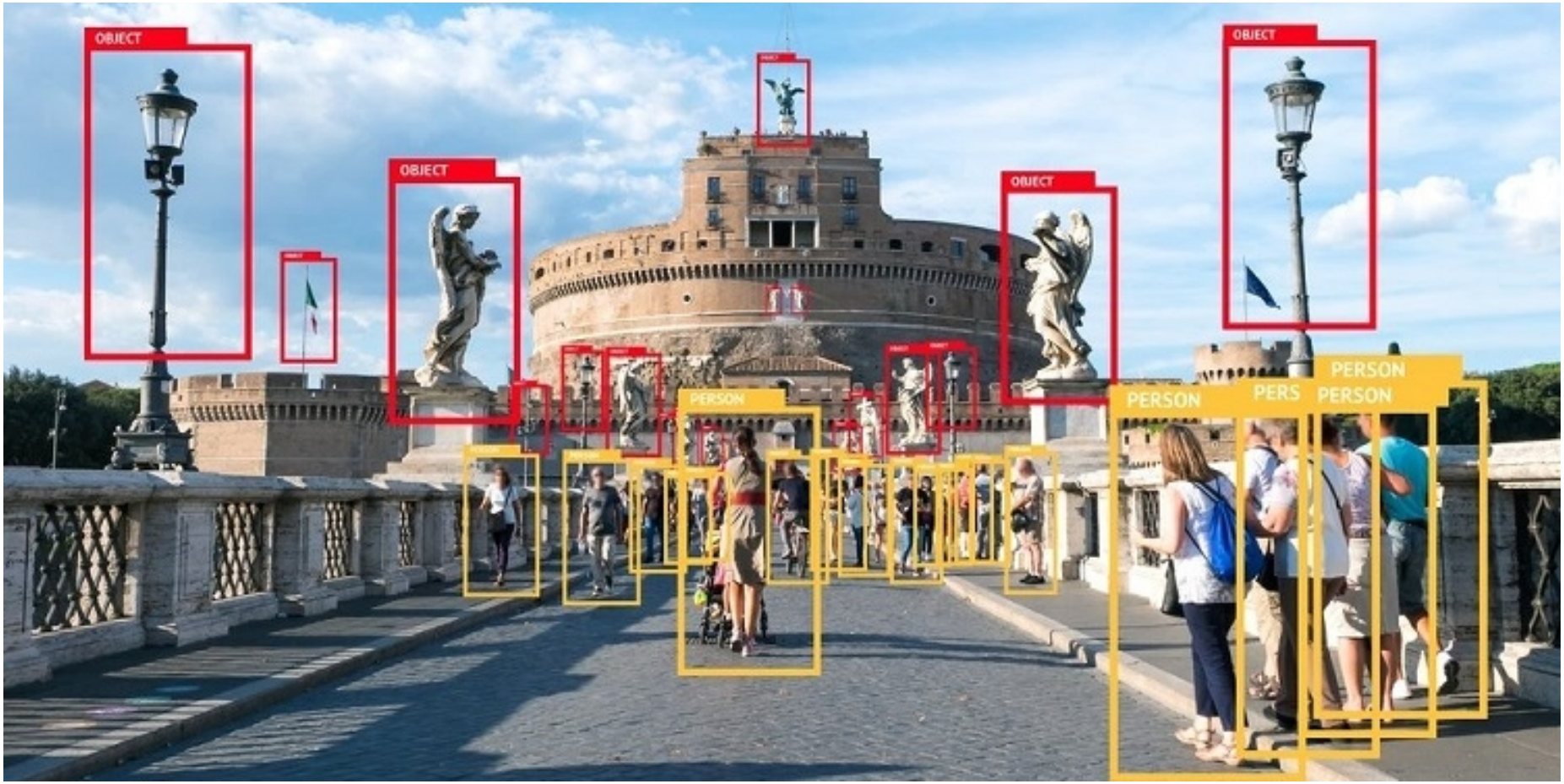


No Supervisado



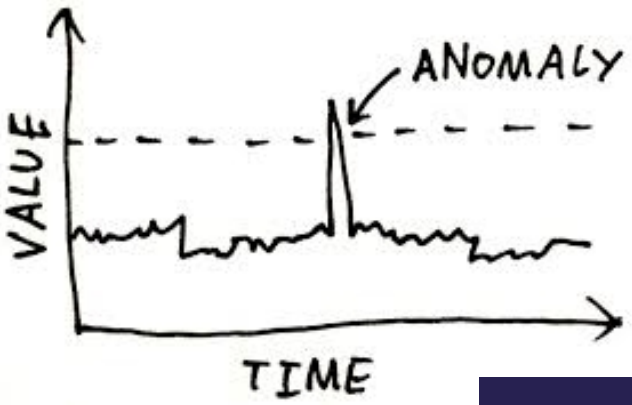
Por refuerzo

Aprendizaje Supervisado



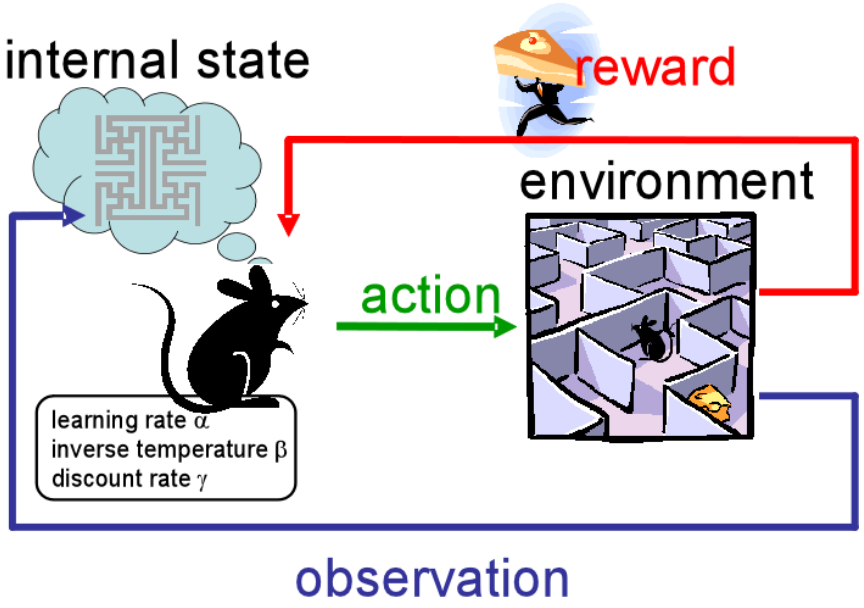
BMW Vision vs Mercedes F015 Self Driving Cars:
https://www.youtube.com/watch?v=CDX391WBwSY

Aprendizaje No Supervisado



Amazon SageMaker for Fraud Detection:
https://www.youtube.com/watch?v=wzwkLV9gDXk

Aprendizaje por Refuerzo



Premio Nacional de Ciencias 2018 en Paraguay:

"Ubicación de Máquinas Virtuales para Infraestructuras Elásticas en Centros de Datos de Computación en Nube bajo Incertidumbre".

Fabio López Pires, Benjamín Barán, Leonardo Benítez, Saúl Zalimben y Augusto Amarilla

Optimización de recursos:

- VMs en un Cloud
- Distribución del Tráfico
- Ruteo en redes
- muchas más!!

Knowledge Discovery (KDD) & Data Mining

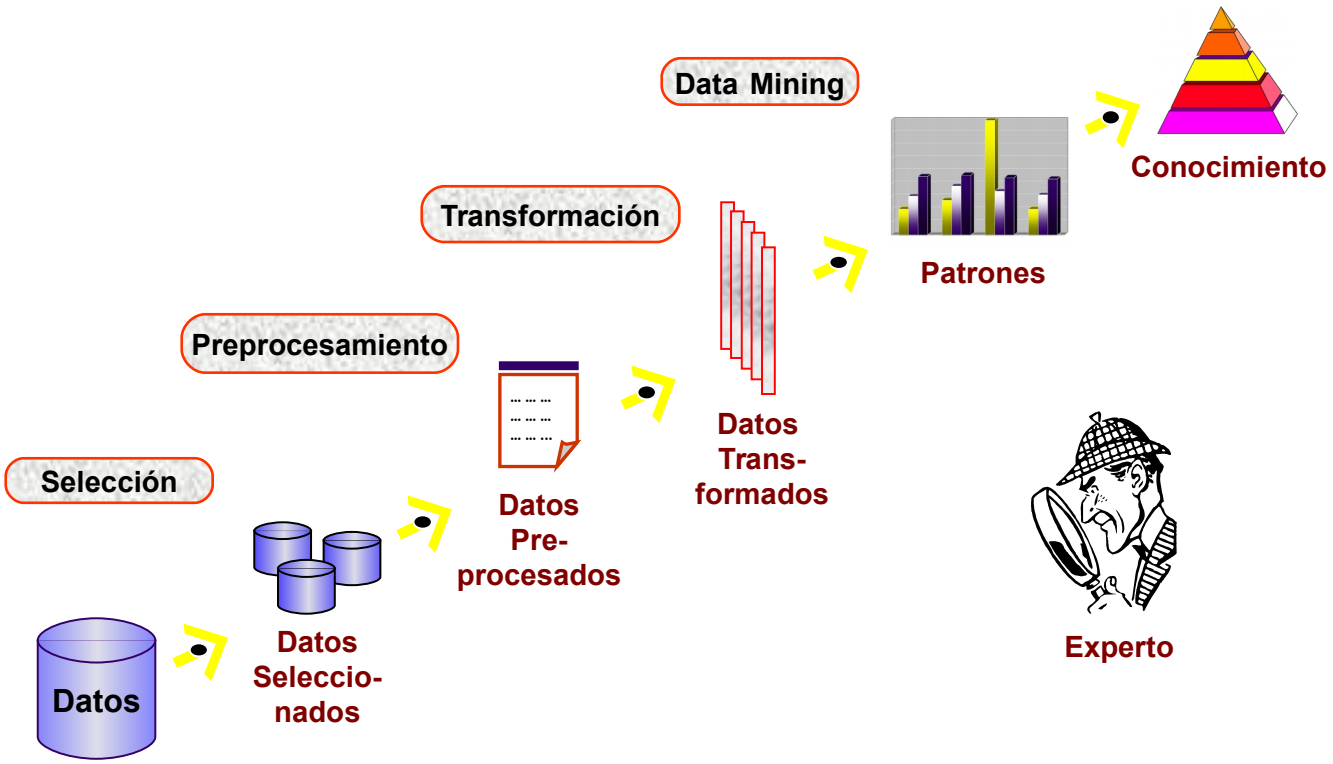
KDD: *descubrir conocimientos mediante la identificación de patrones en los datos*

Data Mining: *1 paso en KDD, que produce una enumeración particular de patrones sobre un conjunto de datos*

Knowledge Discovery (KDD) & Data Mining

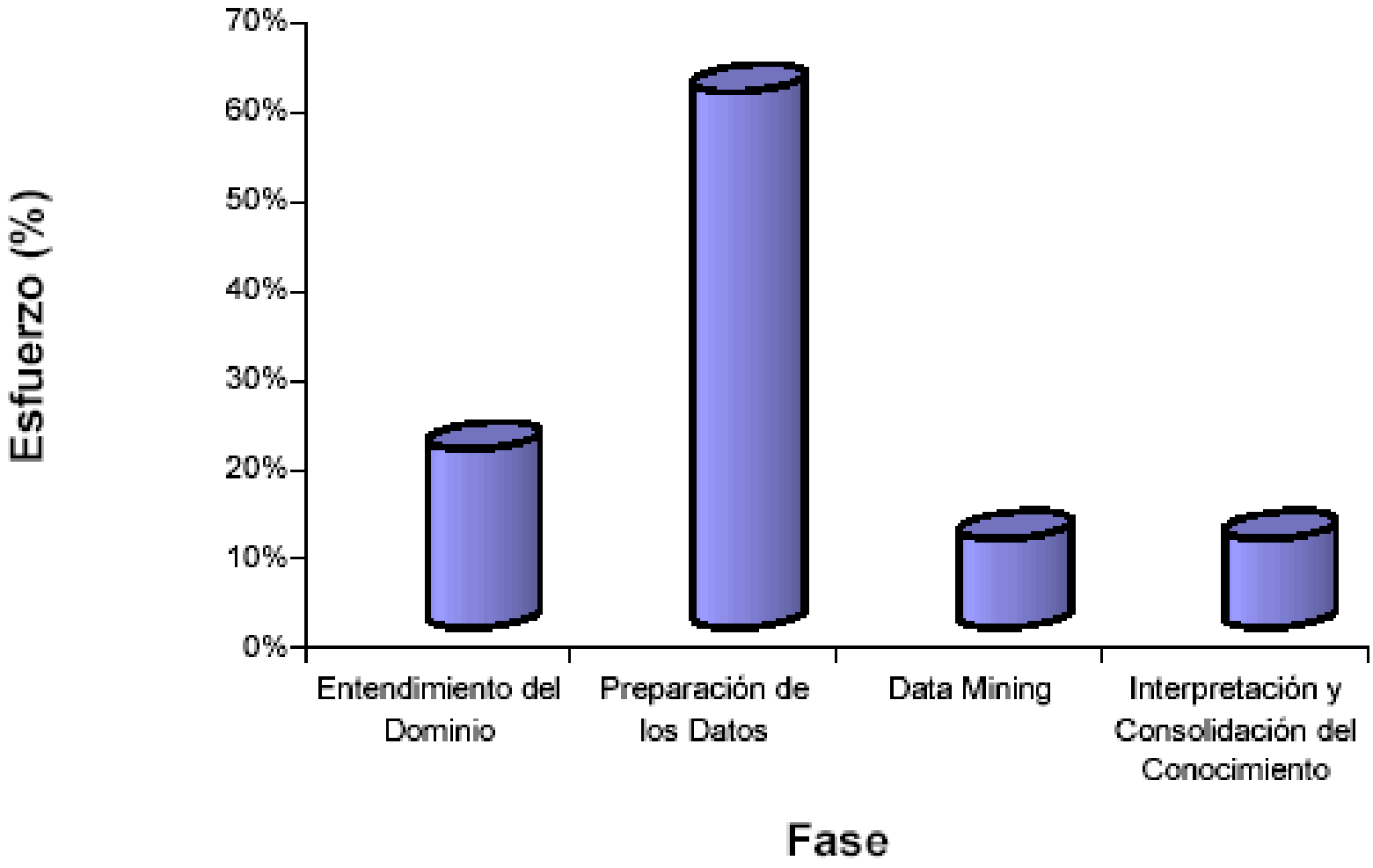
Data Mining: es un paso en el proceso del KDD *consistiendo de algoritmos particulares que, bajo algunas limitaciones aceptables de eficiencia computacional,* ***produce una enumeración particular de patrones sobre un conjunto de hechos***

Proceso de KDD

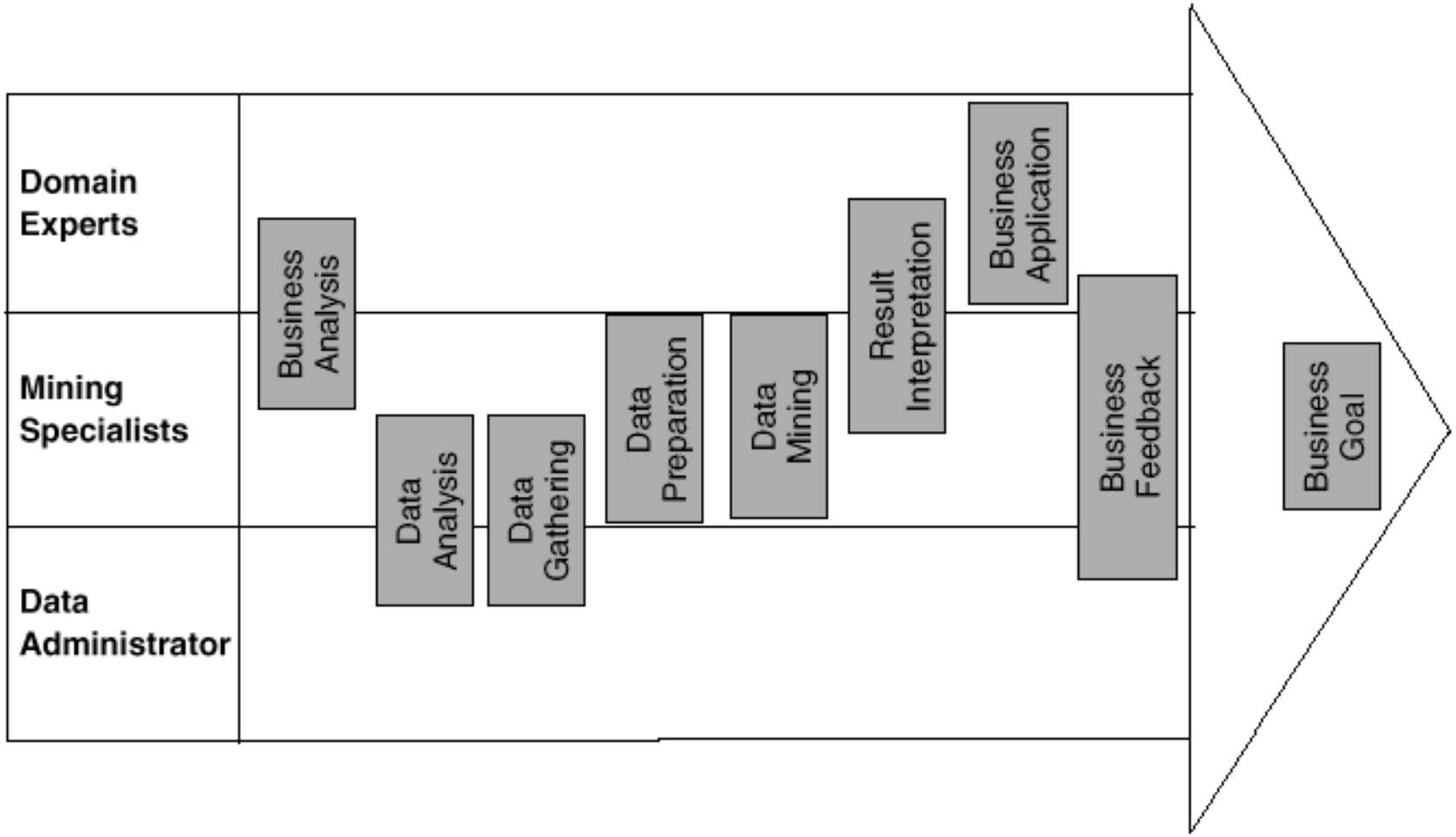


Proceso interactivo e iterativo que envuelve varios pasos y con decisiones a ser tomadas por el usuario

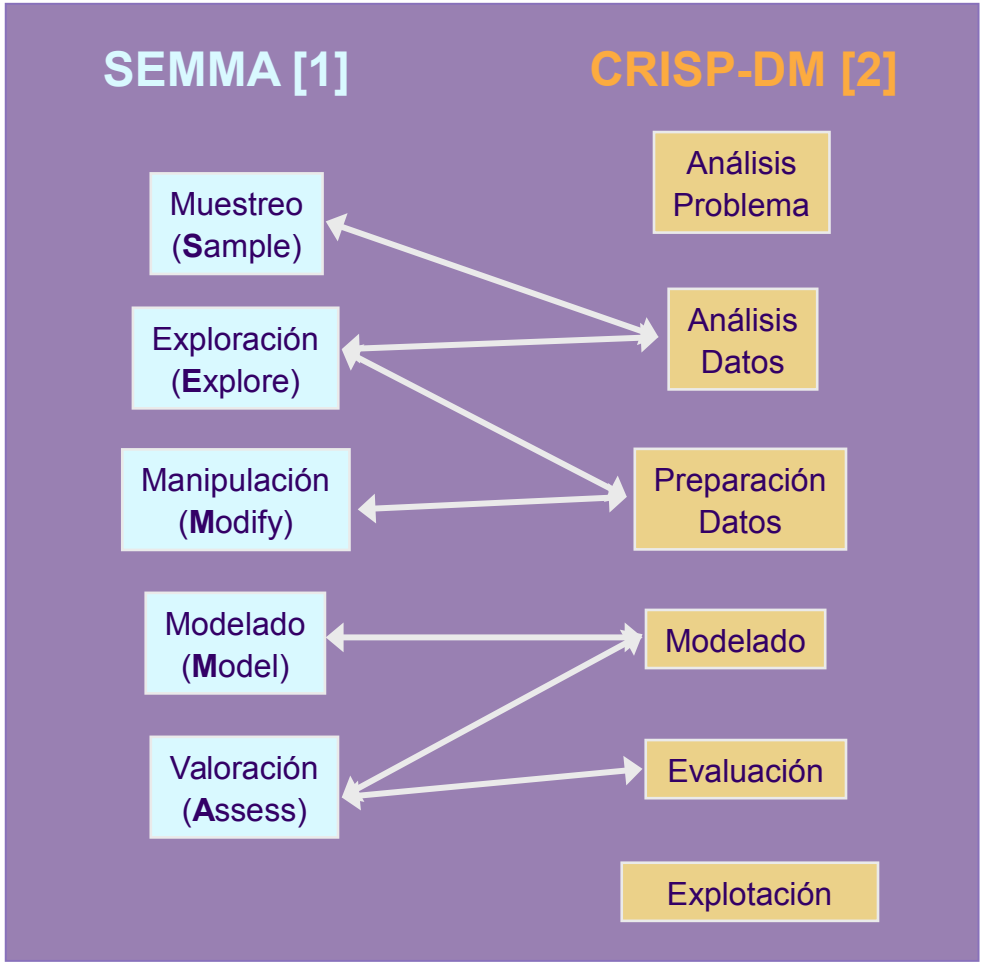
Esfuerzo requerido KDD



Roles en KDD

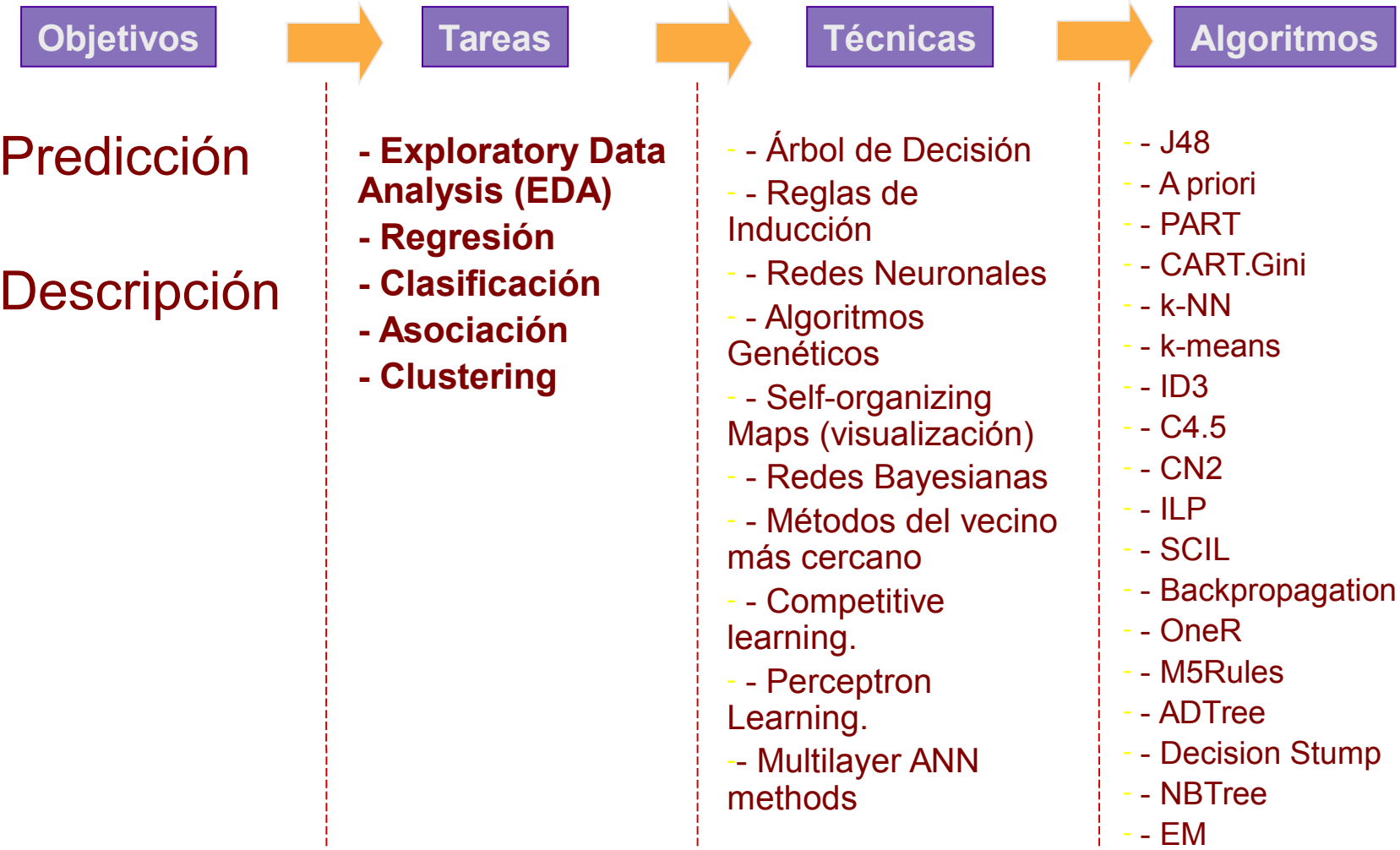

Domain Experts
Mining Specialists
Data Administrator
Business Analysis
Data Analysis
Data Gathering
Data Preparation
Data Mining
Result Interpretation
Business Application
Business Feedback
Business Goal

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)**:

¿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)**:

- Los clientes que compran pañales suelen comprar cerveza
- El tabaco y el alcohol son los factores más importantes en X enfermedad
- Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto

Exploratory Data Analysis (EDA)

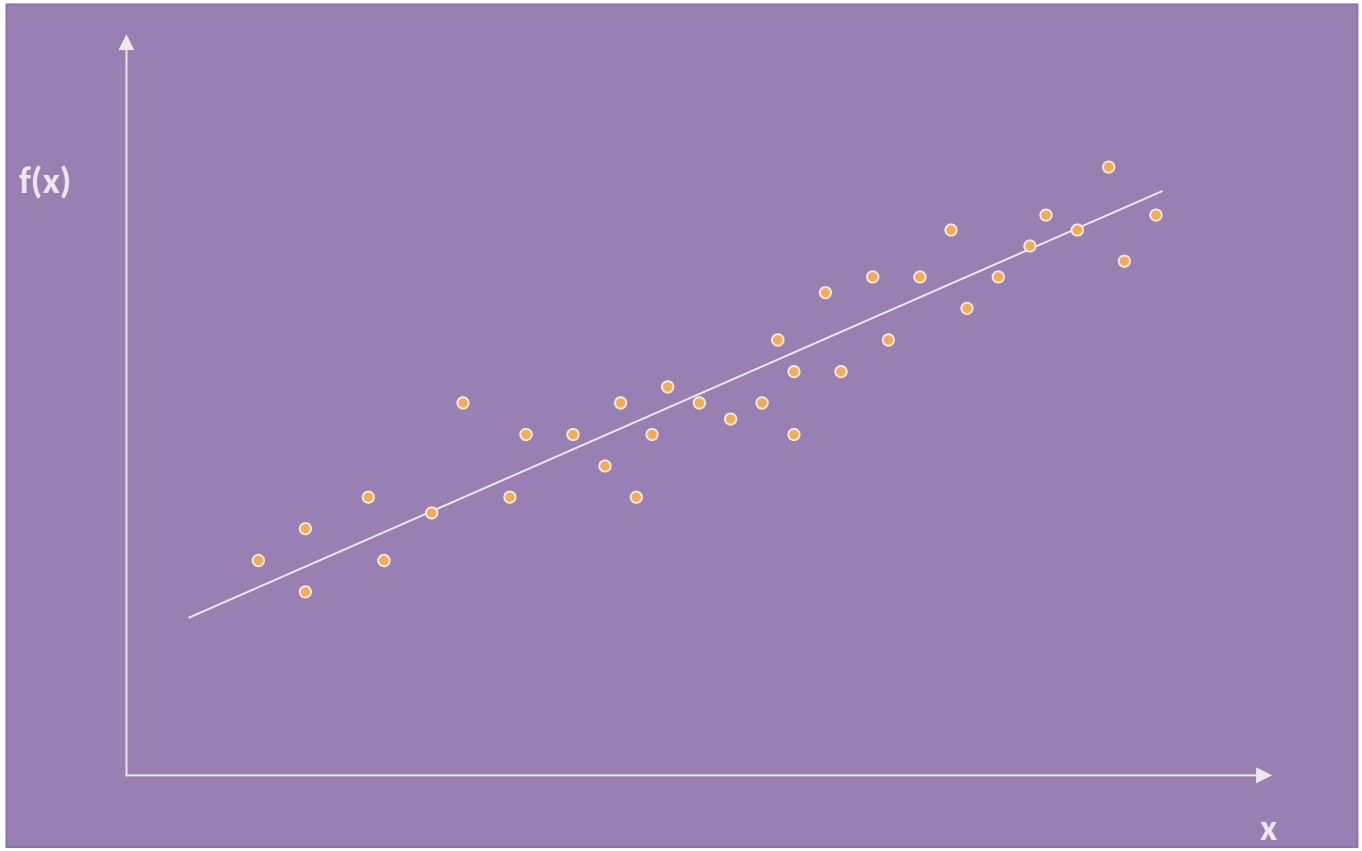
- **Visualizaciones:** consisten en generar modelos visuales que permitan al usuario sacar meta-conocimientos de los mismos

Análisis Descriptivo + Análisis Exploratorio

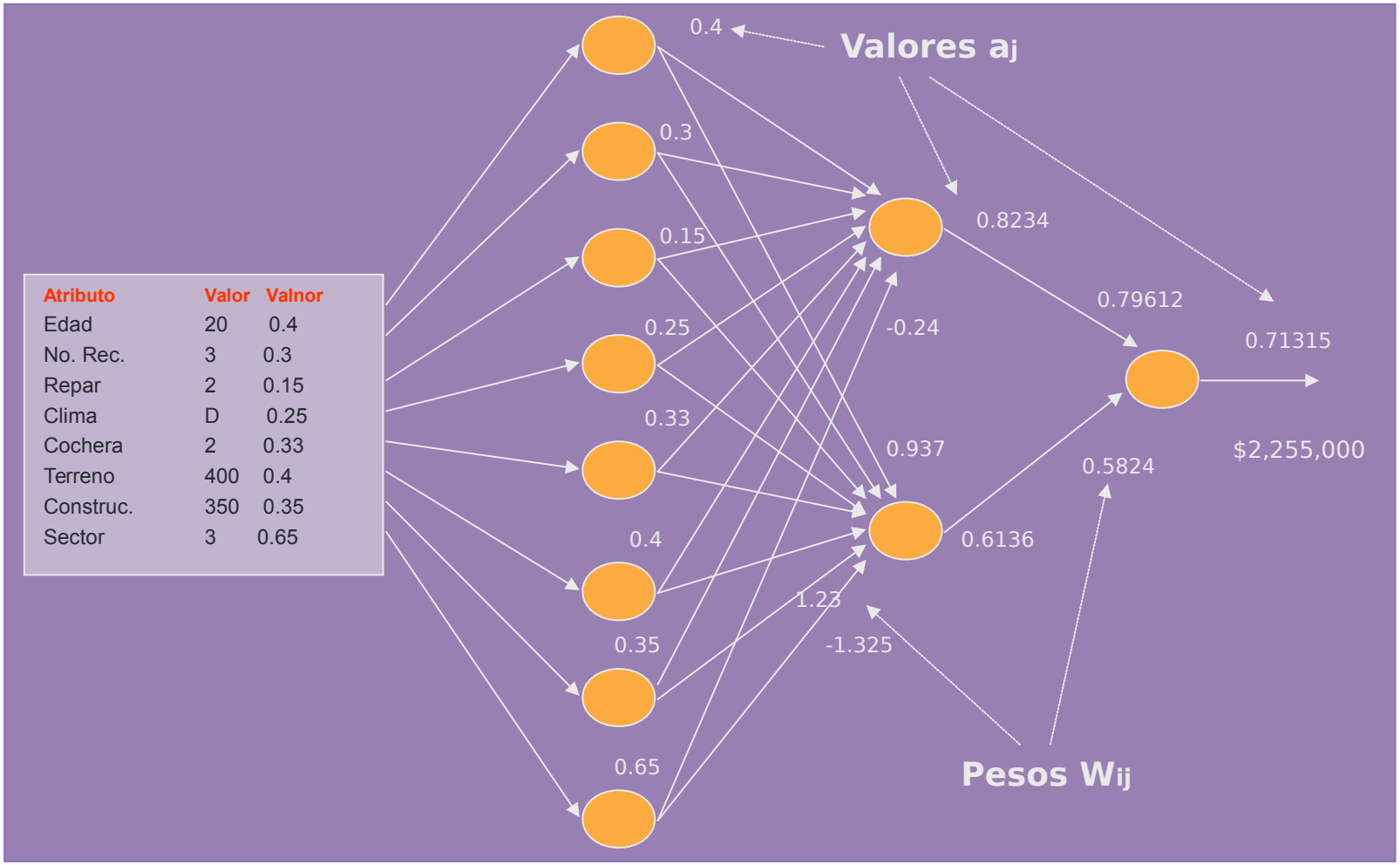
- Pie charts, Donut charts, Histograms, KPI, Maps, Heatmaps, Scatter plot, Box plot, etc

Regresión (Forecasting)

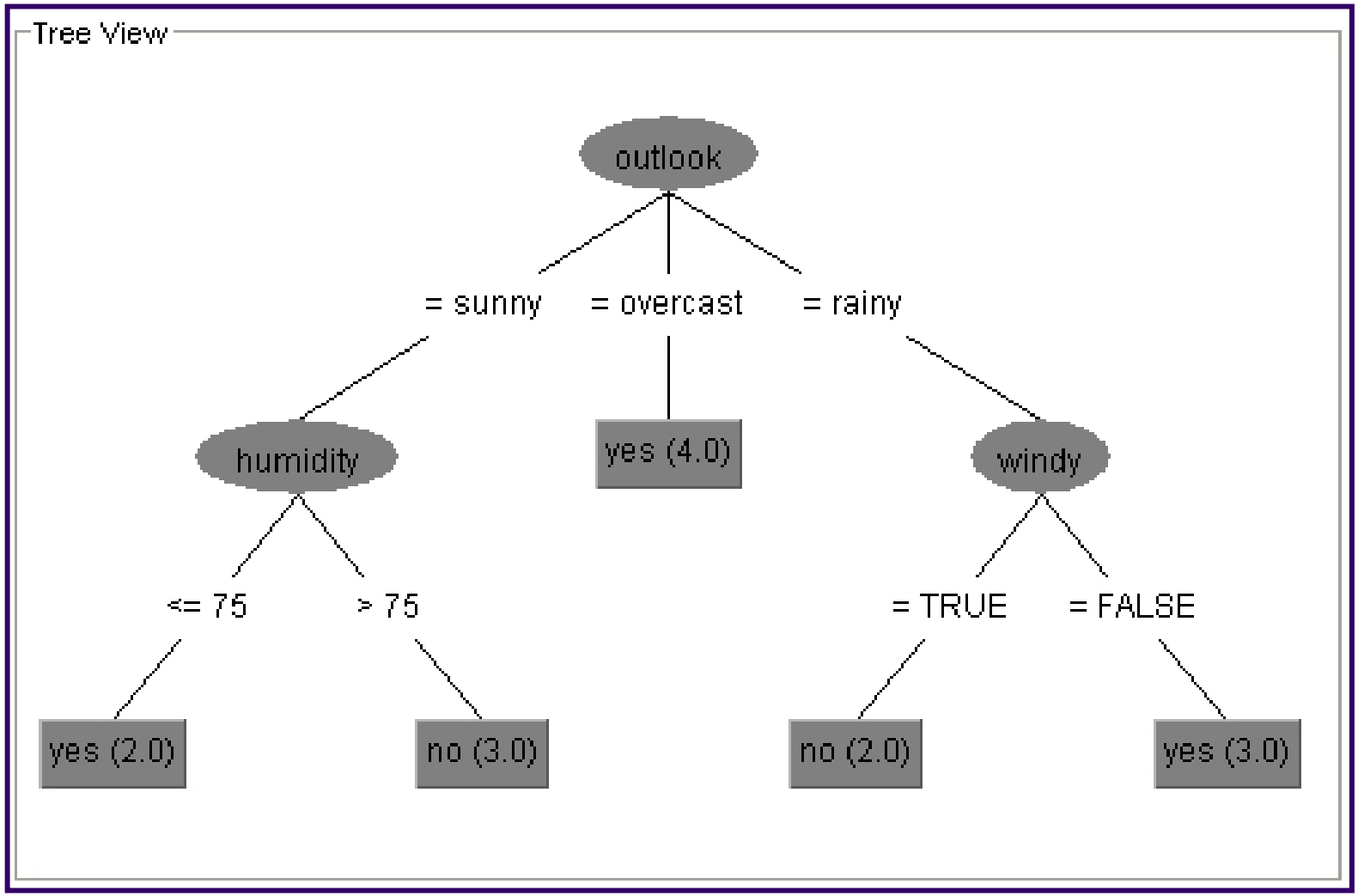
- Ej. se intenta predecir el número de clientes, los ingresos, ganancias, costes, a partir de los resultados de semanas, meses o años anteriores



Clasificación / Predicción – Redes Neuronales



Clasificación – Decision Tree



Modelos Asociativos

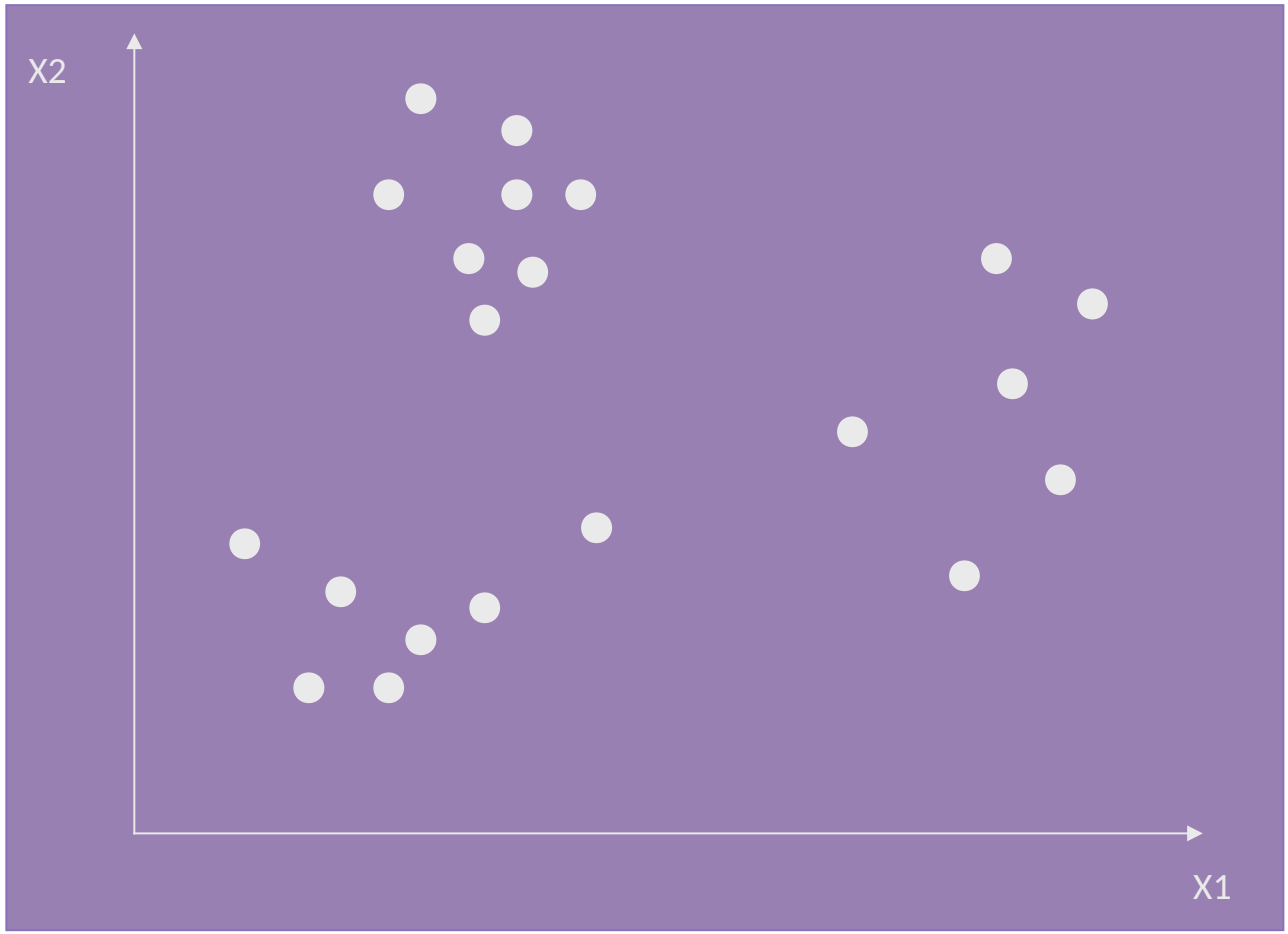
- Ej.: el 70% de los clientes que consumen el producto A y B, también consumen el producto C, D y E.

IF outlook = overcast
THEN play = yes (4.0)

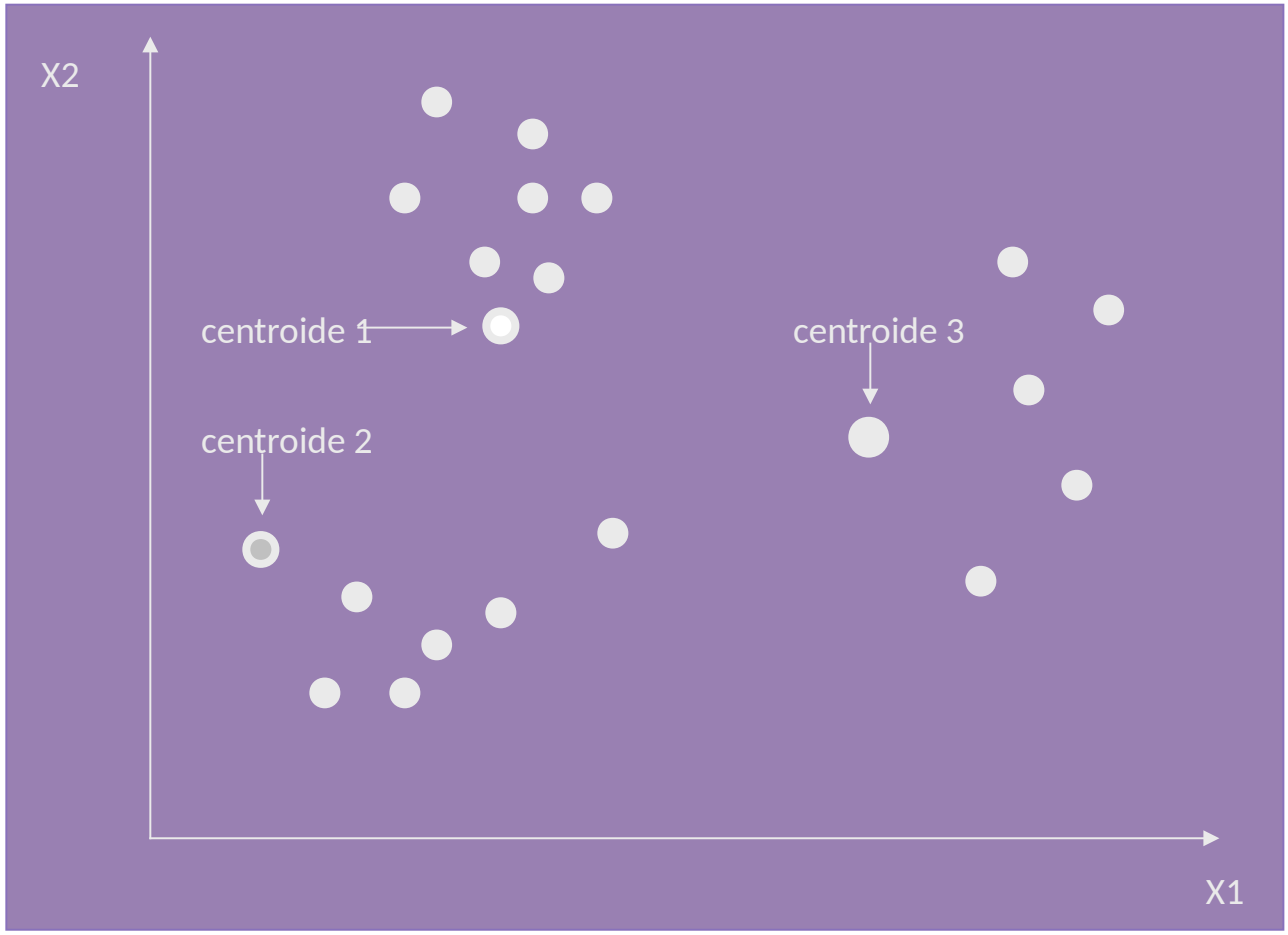
IF windy = TRUE AND
outlook = rainy
THEN play = no (2.0)

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

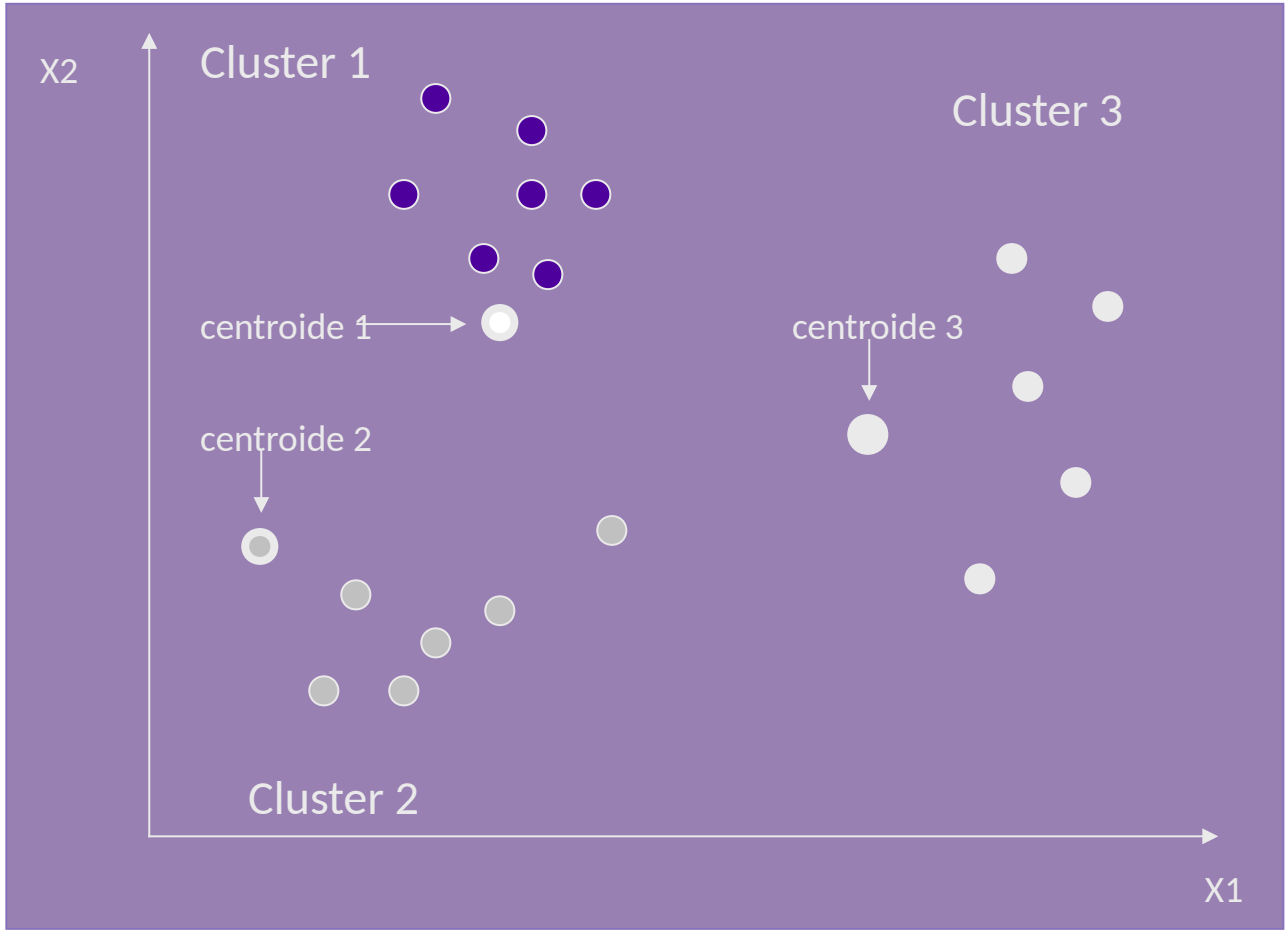
Clustering kMeans



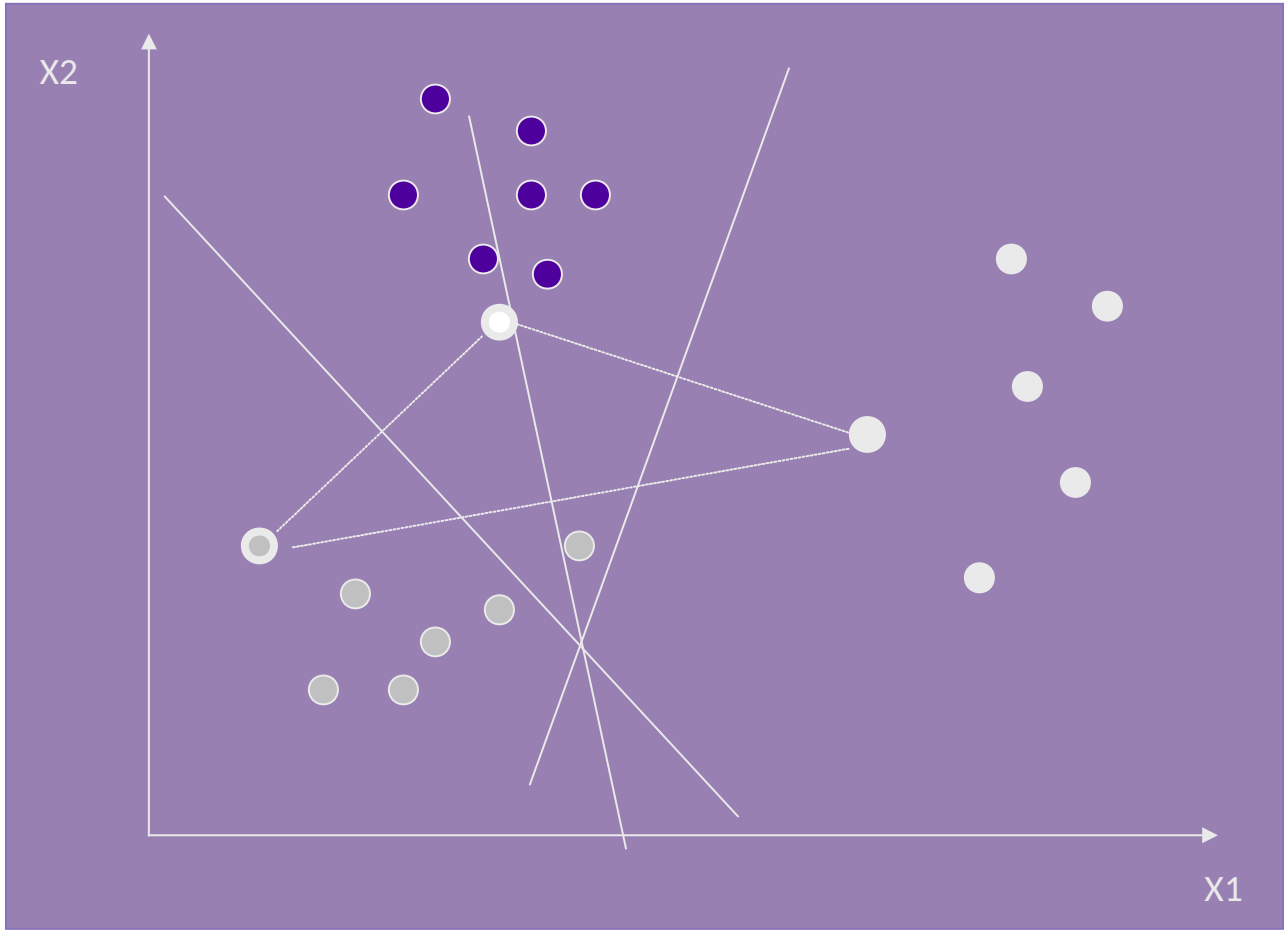
Clustering kMeans



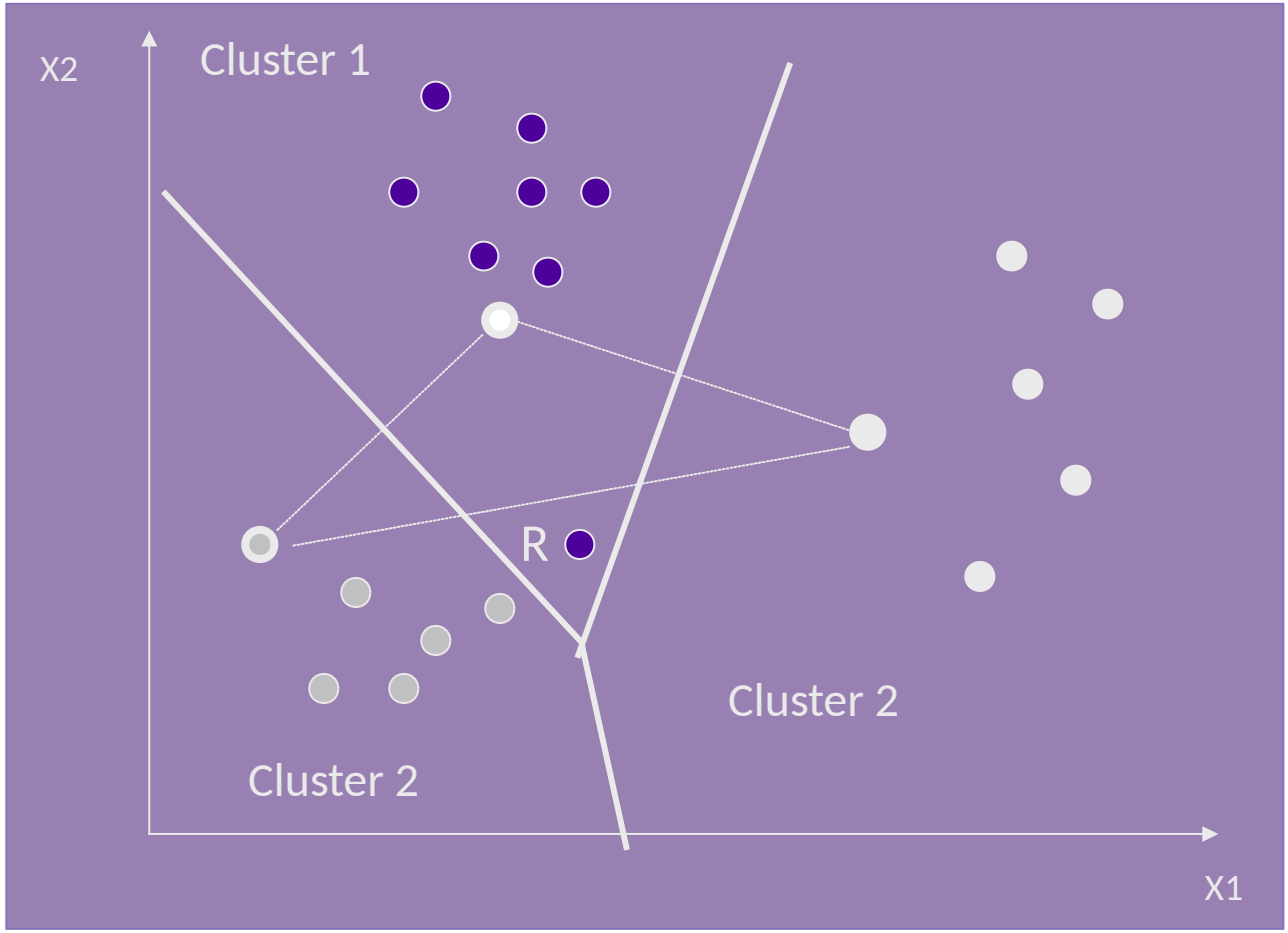
Clustering kMeans



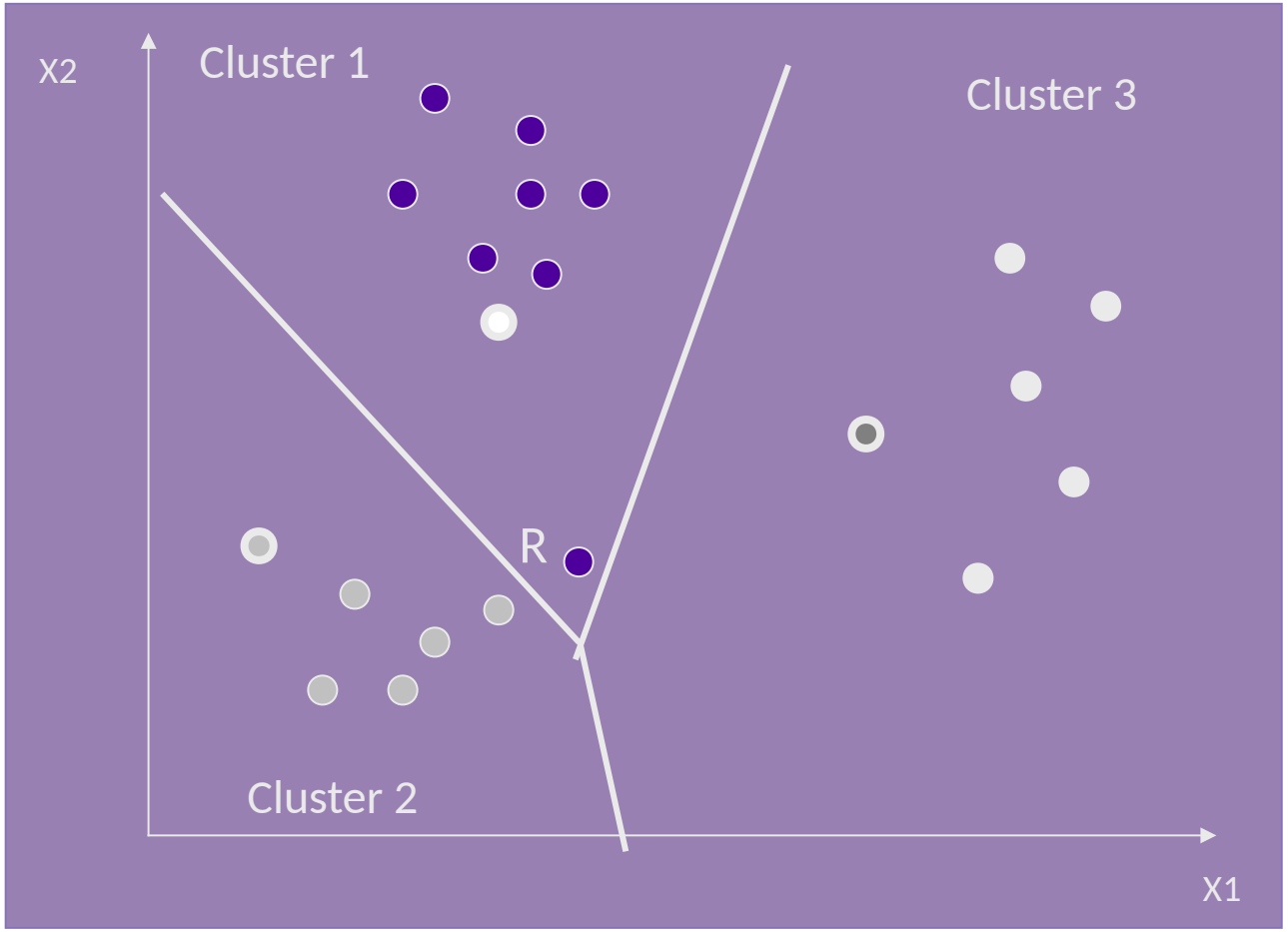
Clustering kMeans



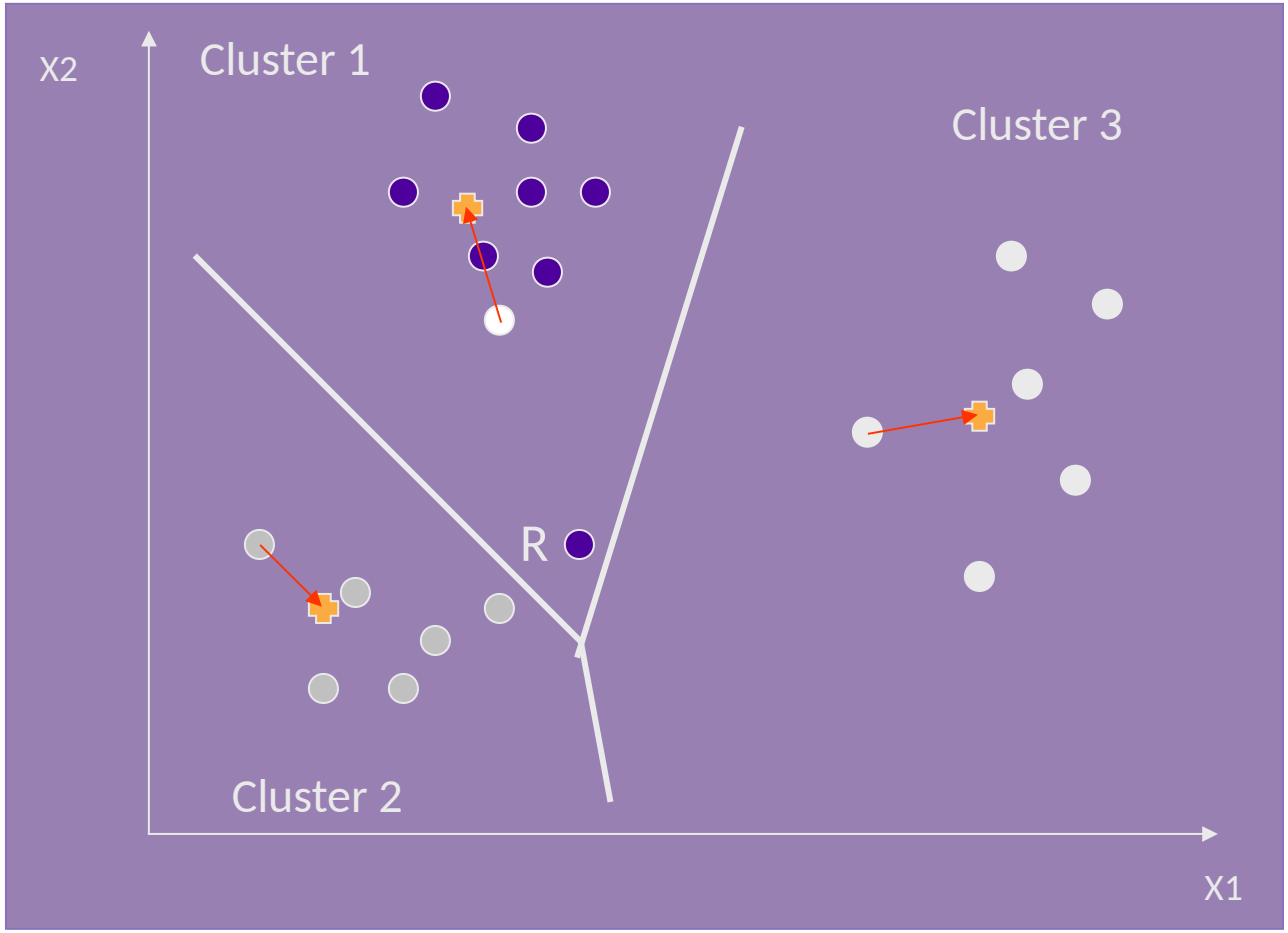
Clustering kMeans



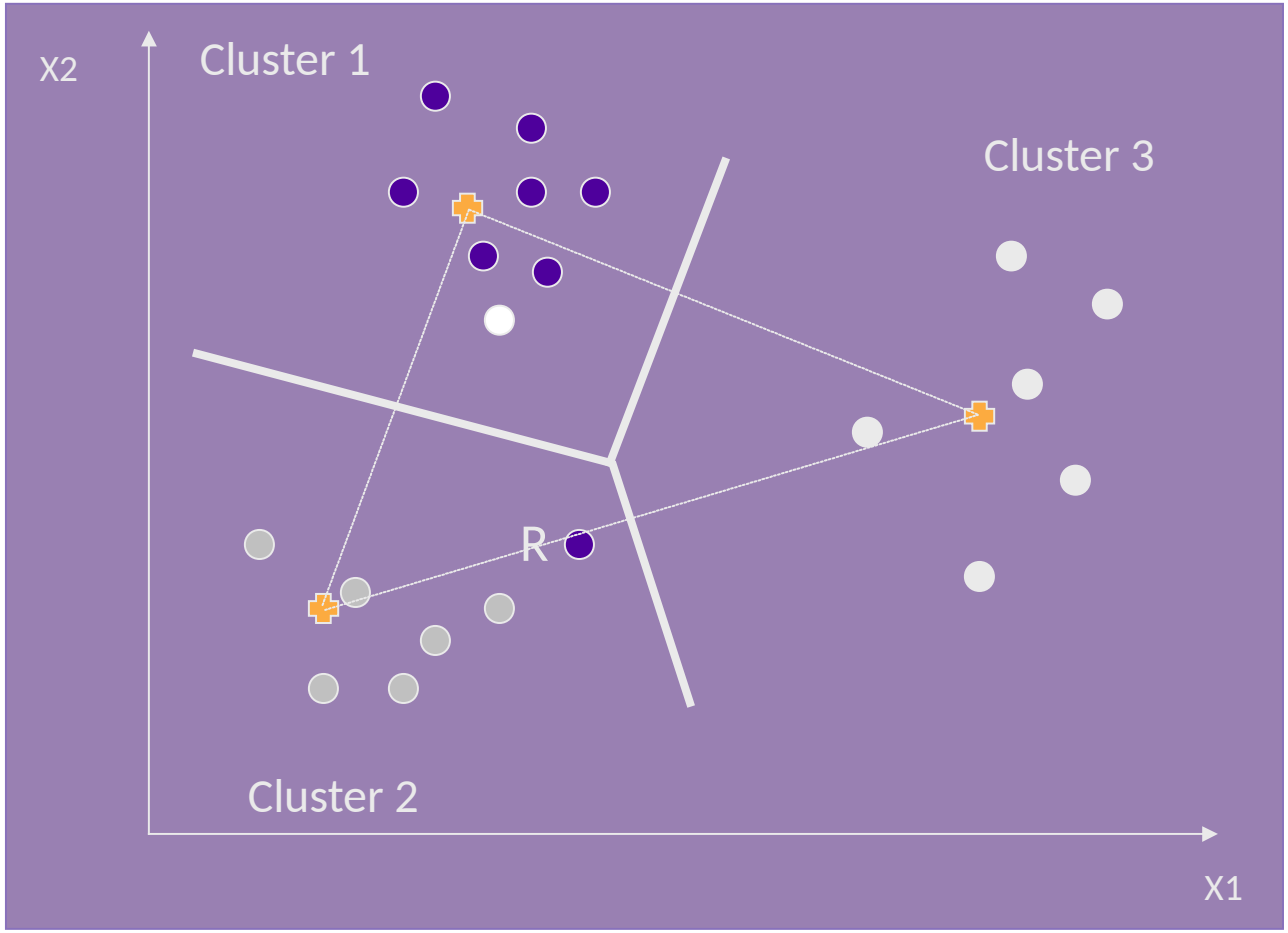
Clustering kMeans



Clustering kMeans



Clustering kMeans



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

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**

Scoring en clasificación

Binary Classification Test:

- *True Positive (TP)*: valores positivos clasificados correctamente por el predictor

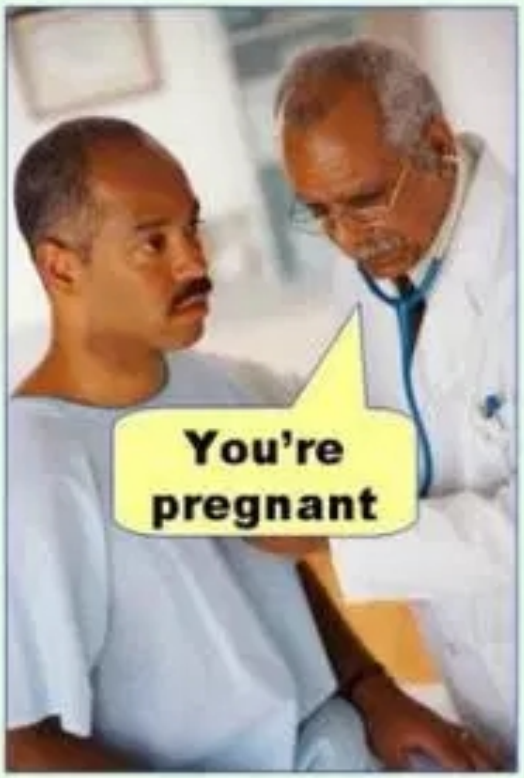
- *True Negative (TN)*: valores negativos clasificados correctamente por el predictor

- *False Positive (FP)*: valores negativos clasificados como positivos. Error del Tipo I

- *False Negative (FN)*: valores positivos clasificados como negativos. Error del Tipo II

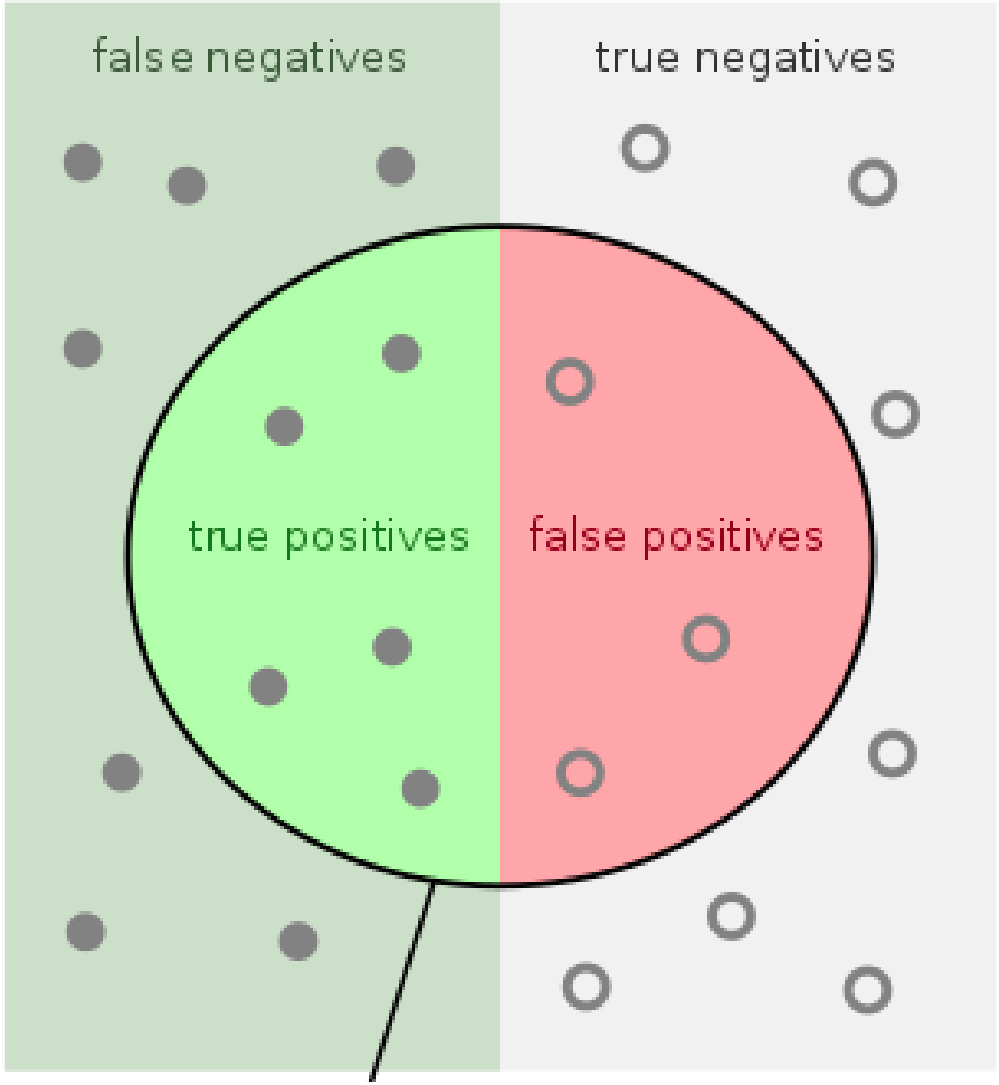
Error Types I, II

Type I error
(false positive)



Type II error
(false negative)

Scoring en clasificación



Scoring en clasificación

- ***Sensitivity o Recall o True Positive Rate (TPR):*** TP / P = TP / (TP + FN)
- ***Specificity (SPC) o True Negative Rate (TNR):*** TN / N = TN / (TN + FP)
- ***False Positive Rate (FPR):*** FP / N = 1 - SPC
- ***False Negative Rate (FNR):*** FN / (TP + FN) = 1 – TPR
- ***Precision:*** TP / (TP + FP)
- ***Accuracy:*** (TP + TN) / (TP + TN + FP + FN)
- ***F-Score o F-Measure***, media armónica entre precision y recall:

  2 . (precision . recall) / (precision + recall)

Scoring en clasificación

- **Confusion Matrix o Error Matrix**

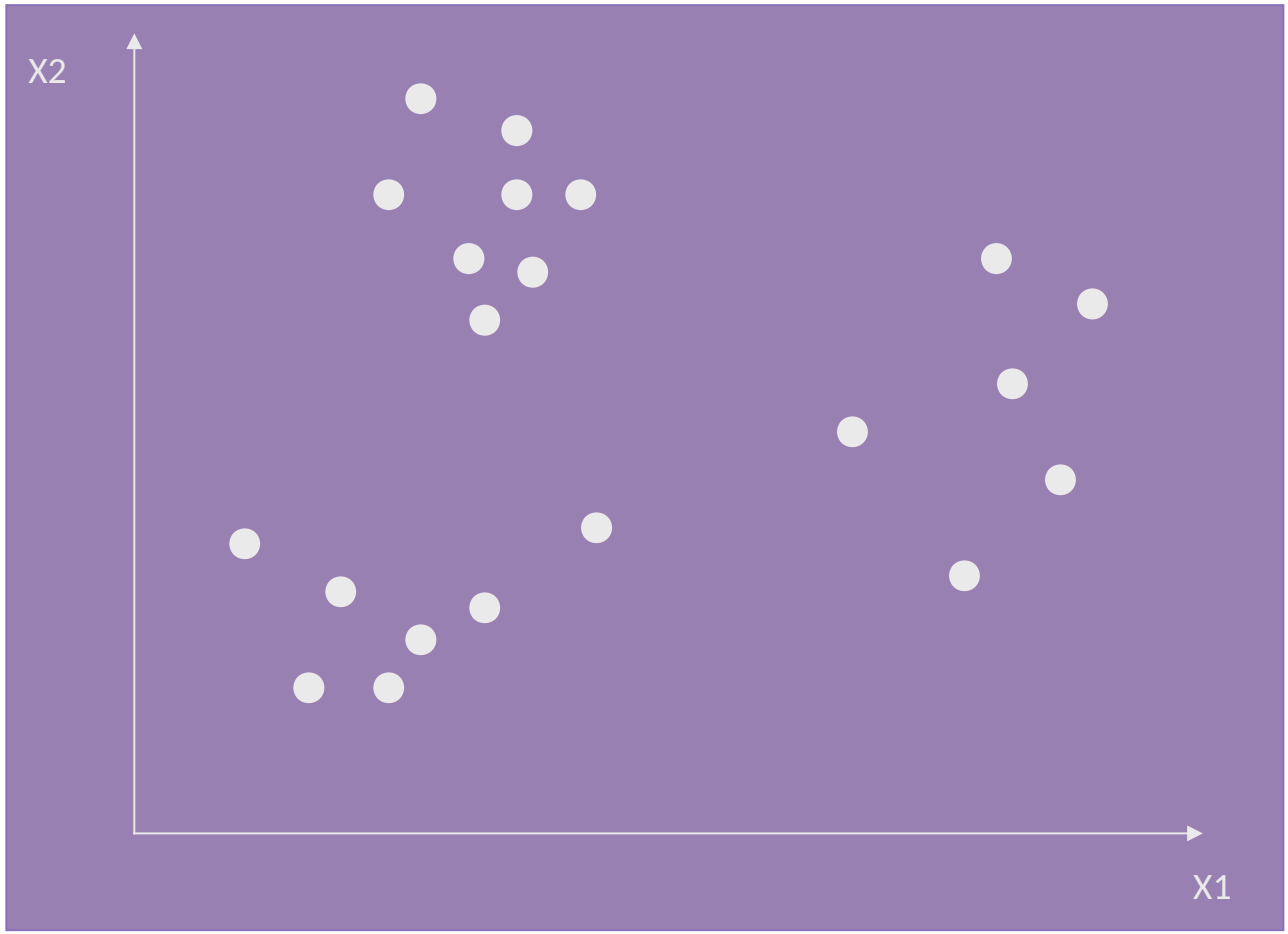
		Actual class		
		Cat	**Dog**	**Rabbit**
Predicted class	**Cat**	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

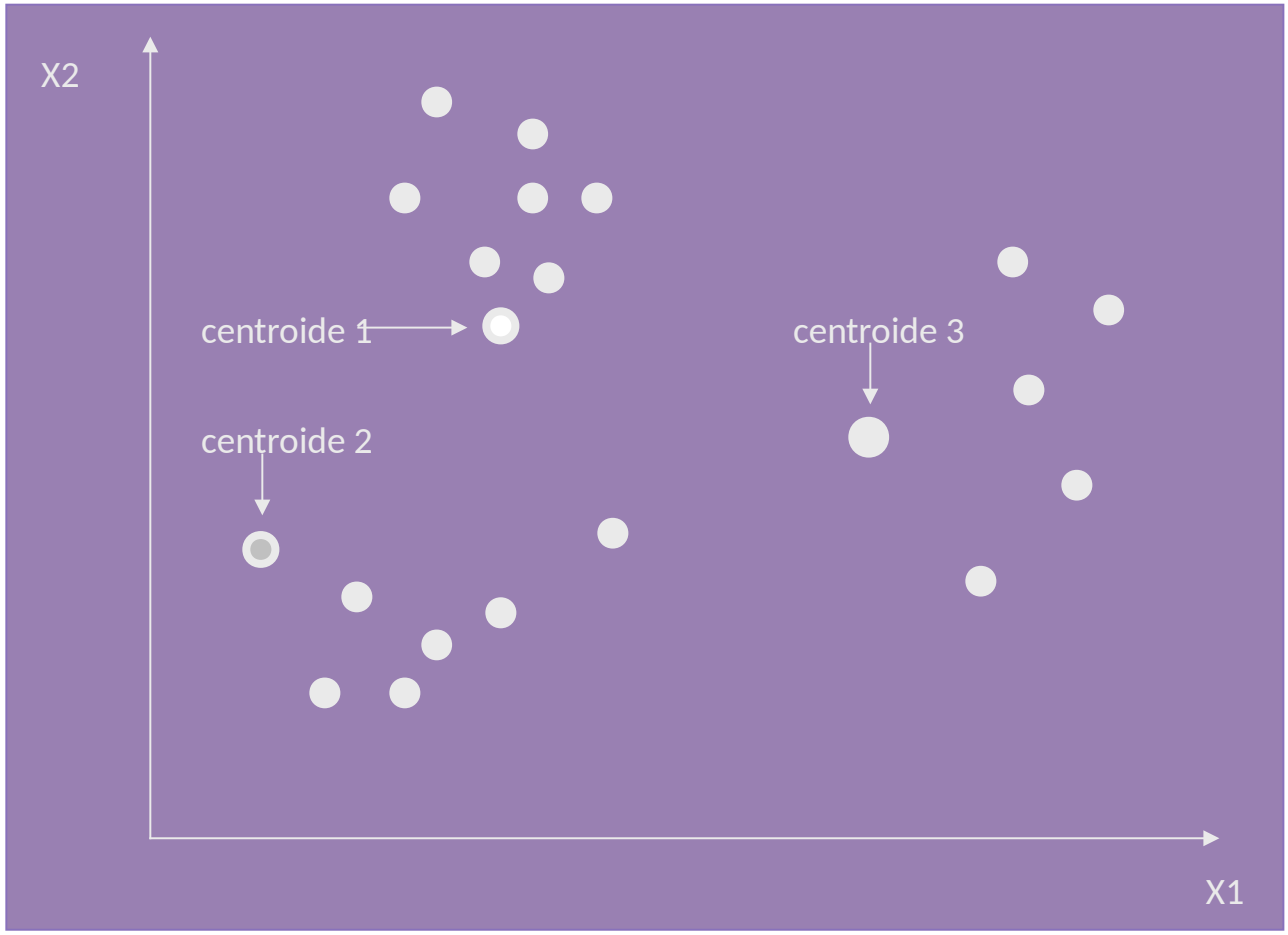
Clustering o Agrupamiento

- Aprendizaje no supervisado:
  - No tenemos una variable target
  - No contamos con un dataset clasificado con los resultados correctos (dataset marcado)
- De acuerdo al tipo de variables:
  - Clustering de variables numéricas: k-Means
  - Clustering de variables categóricas: k-Modes

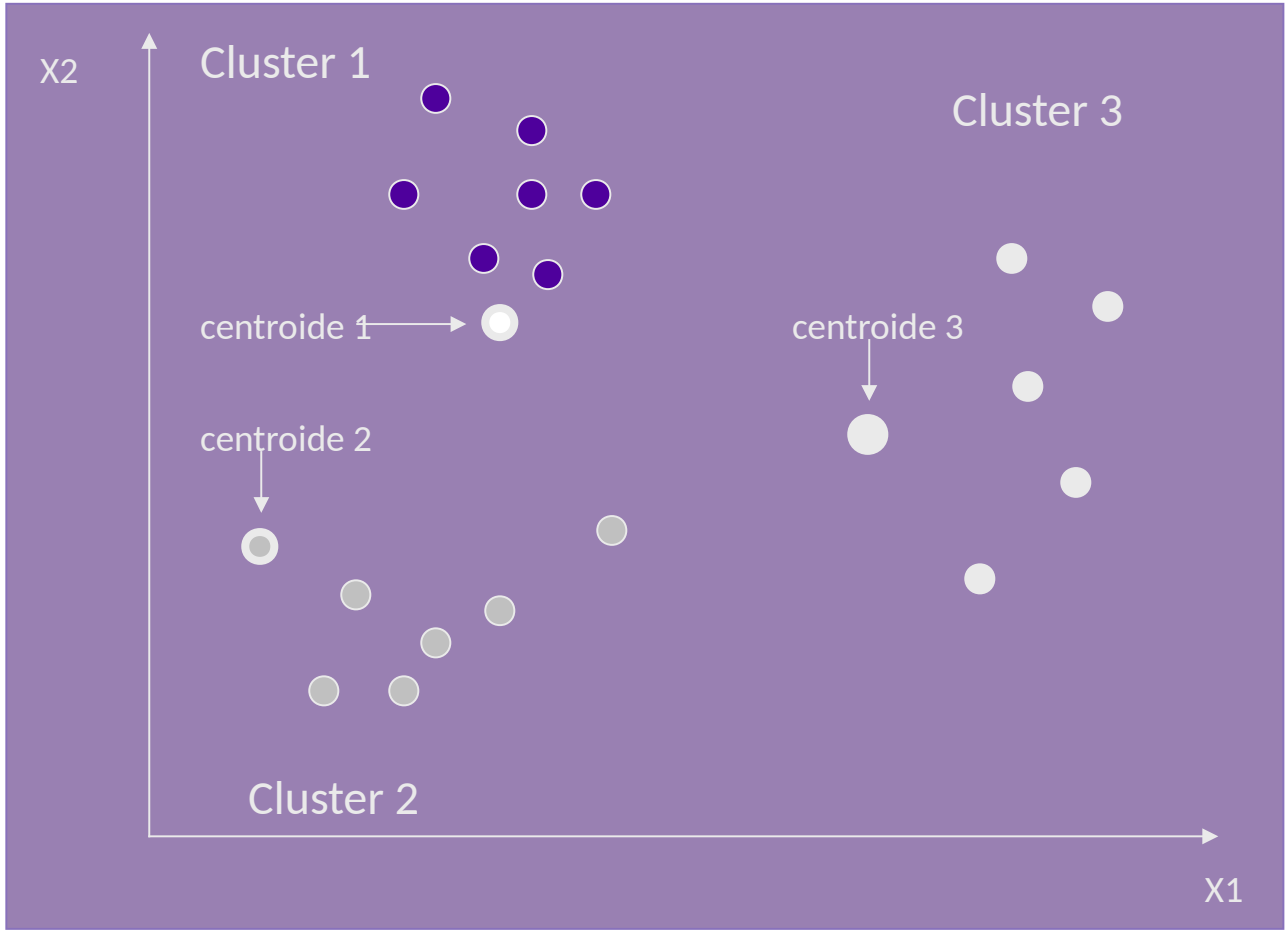
Clustering kMeans



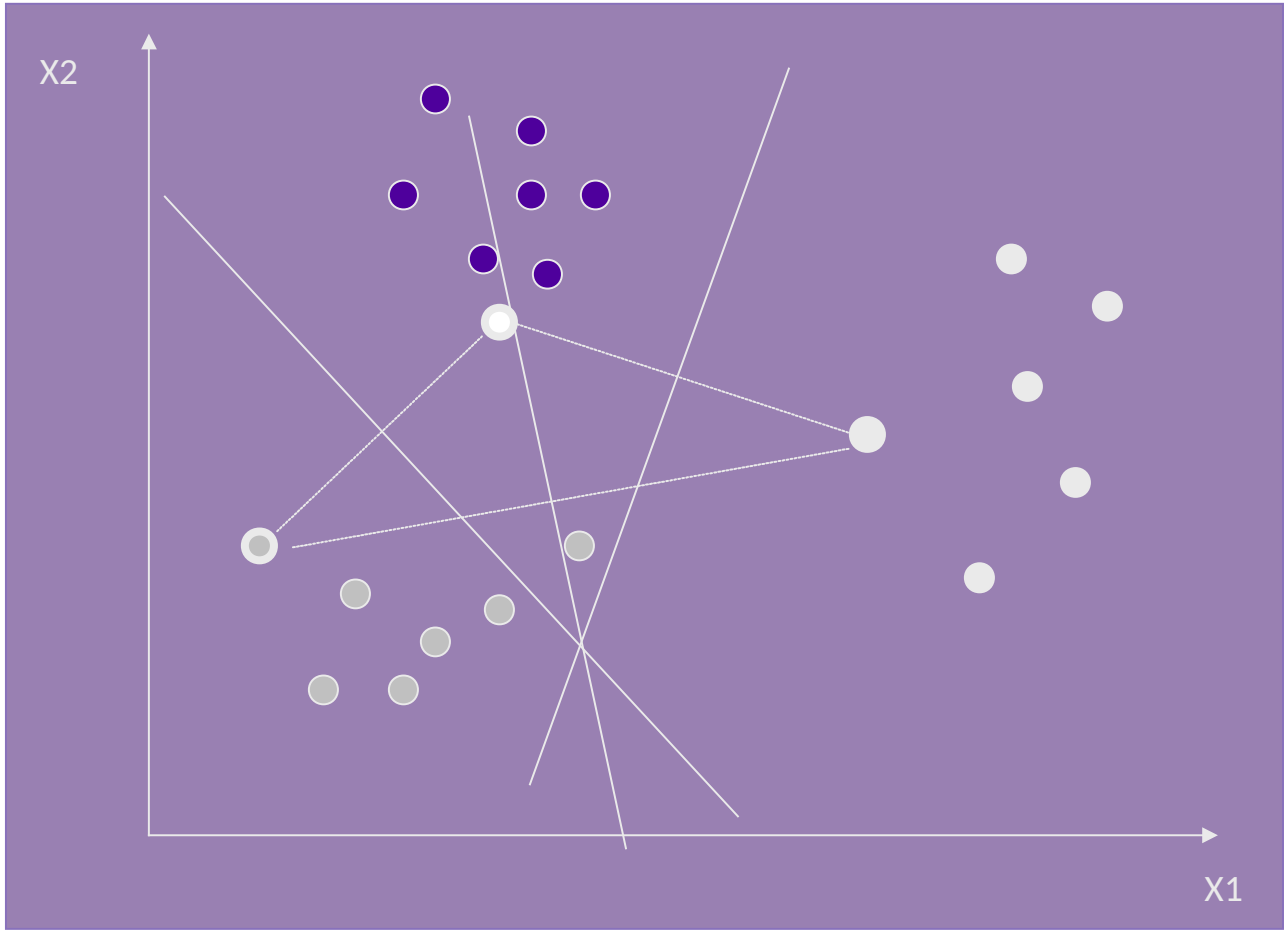
Clustering kMeans



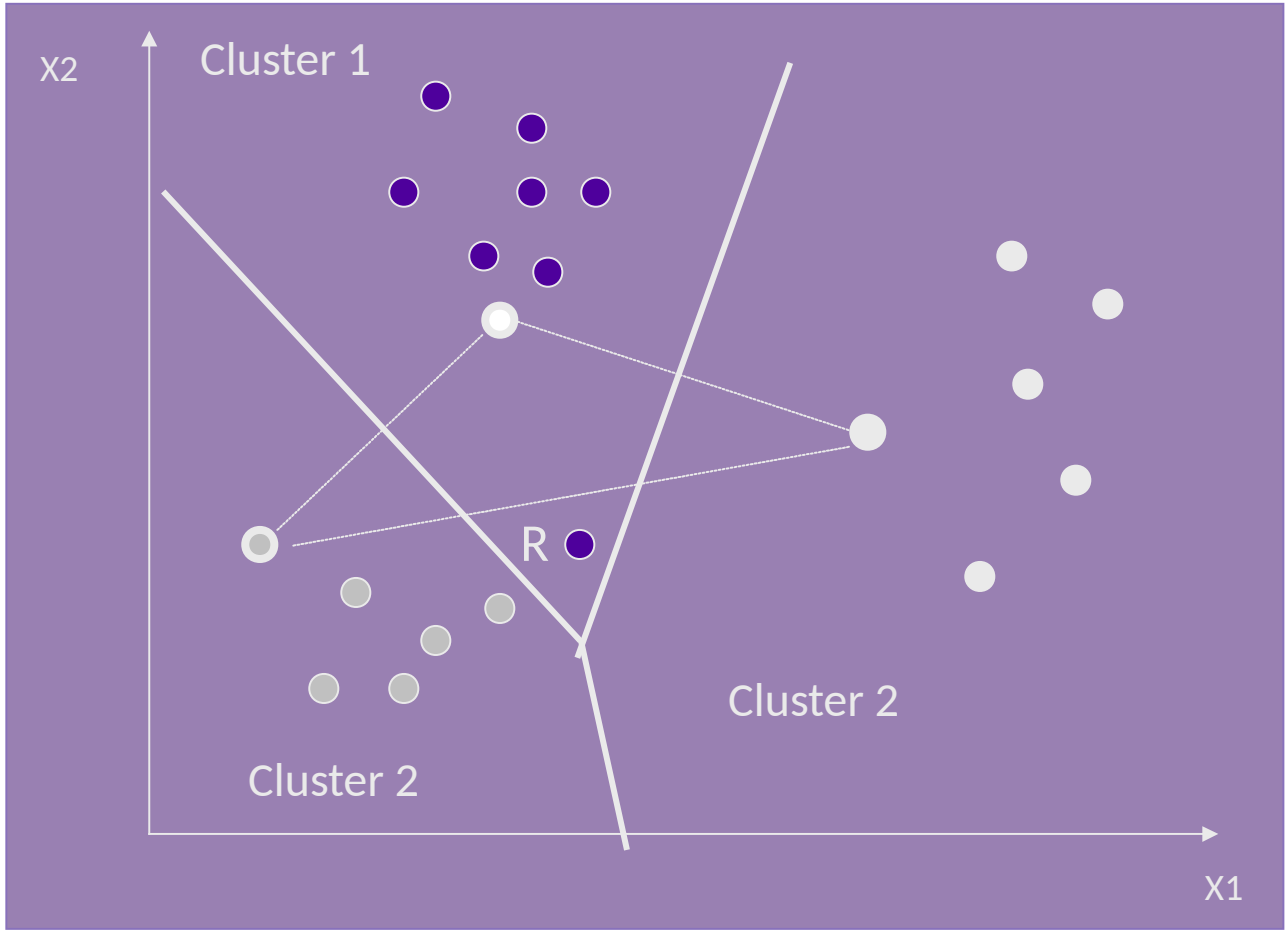
Clustering kMeans



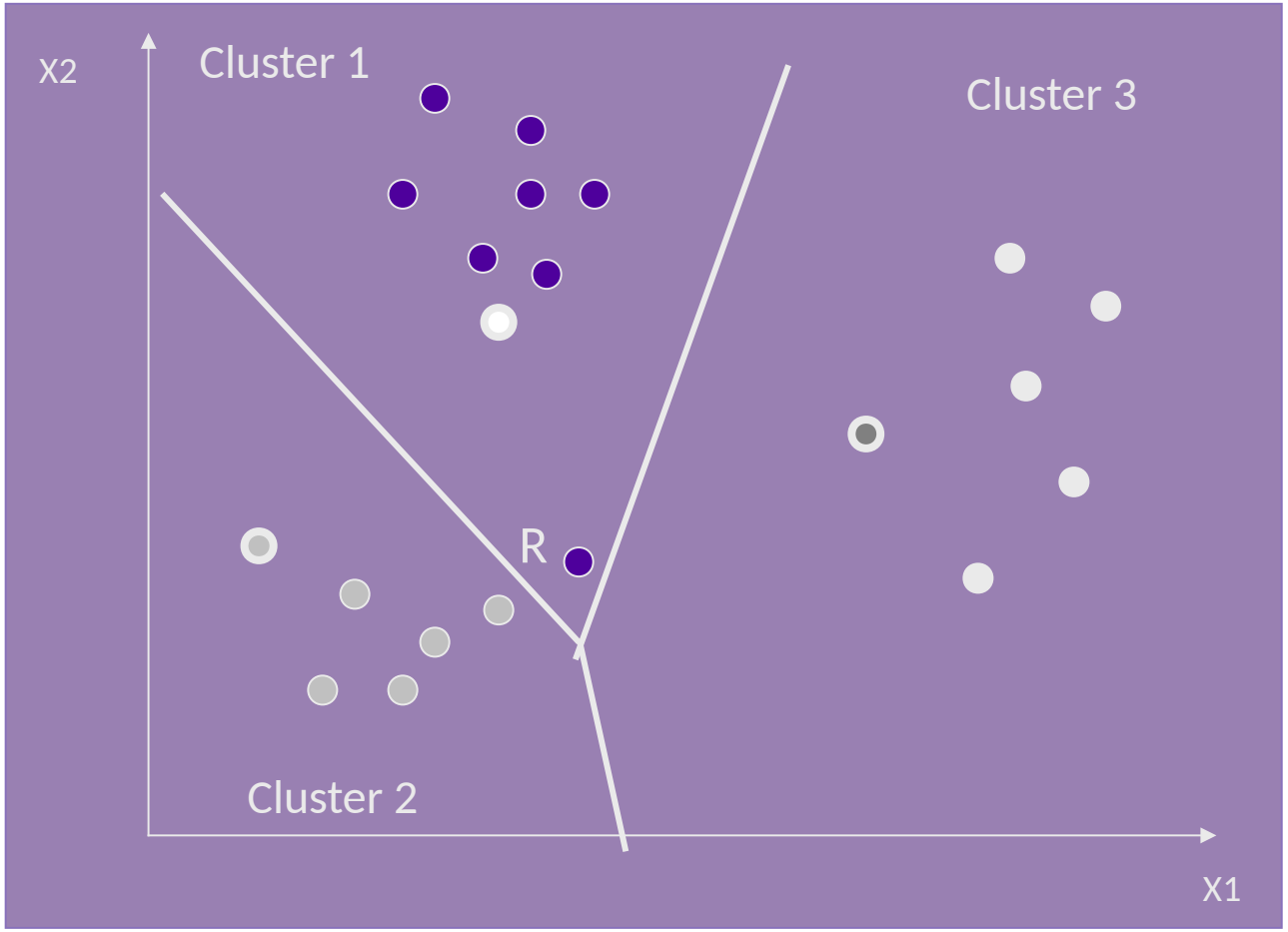
Clustering kMeans



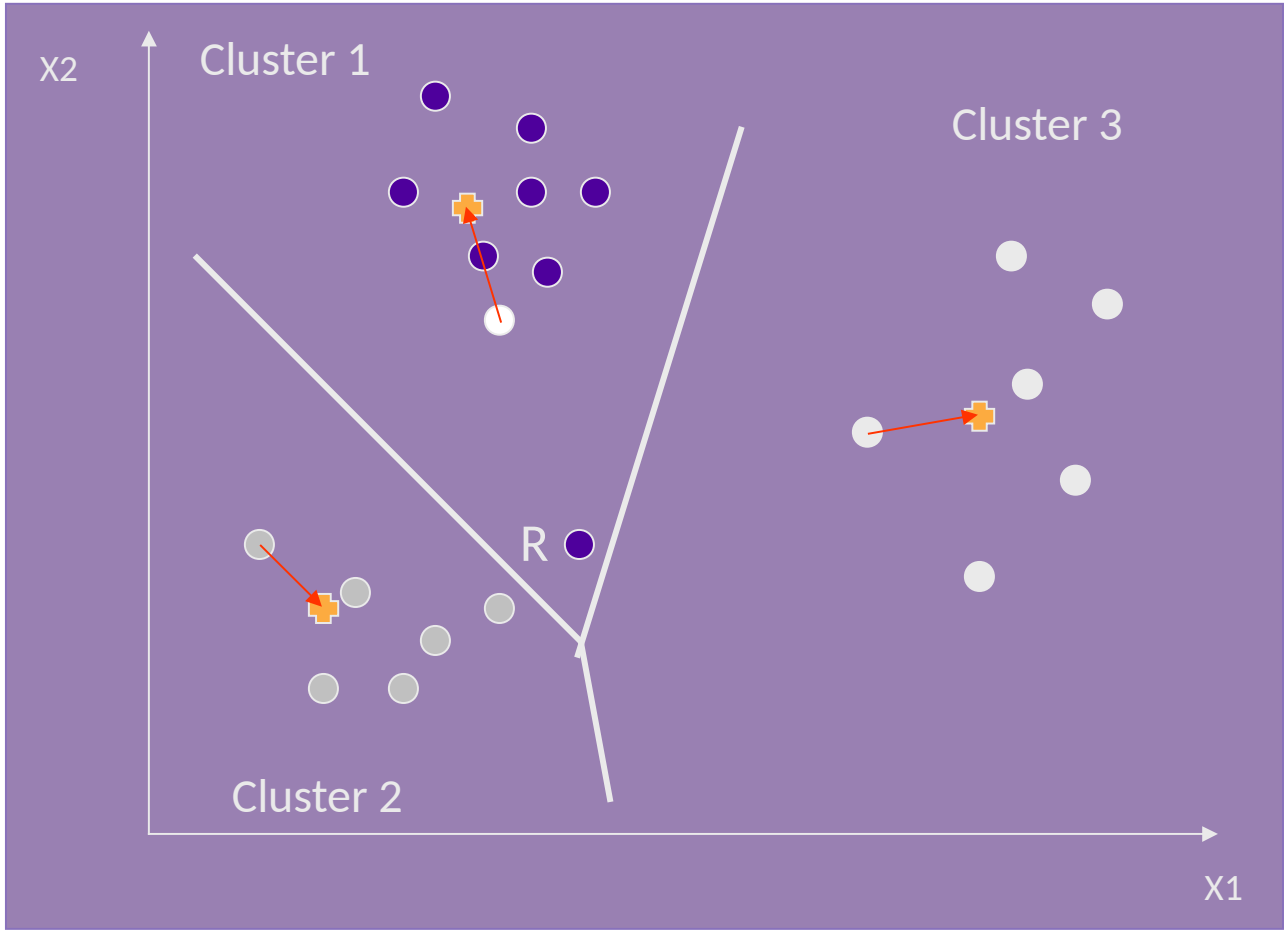
Clustering kMeans



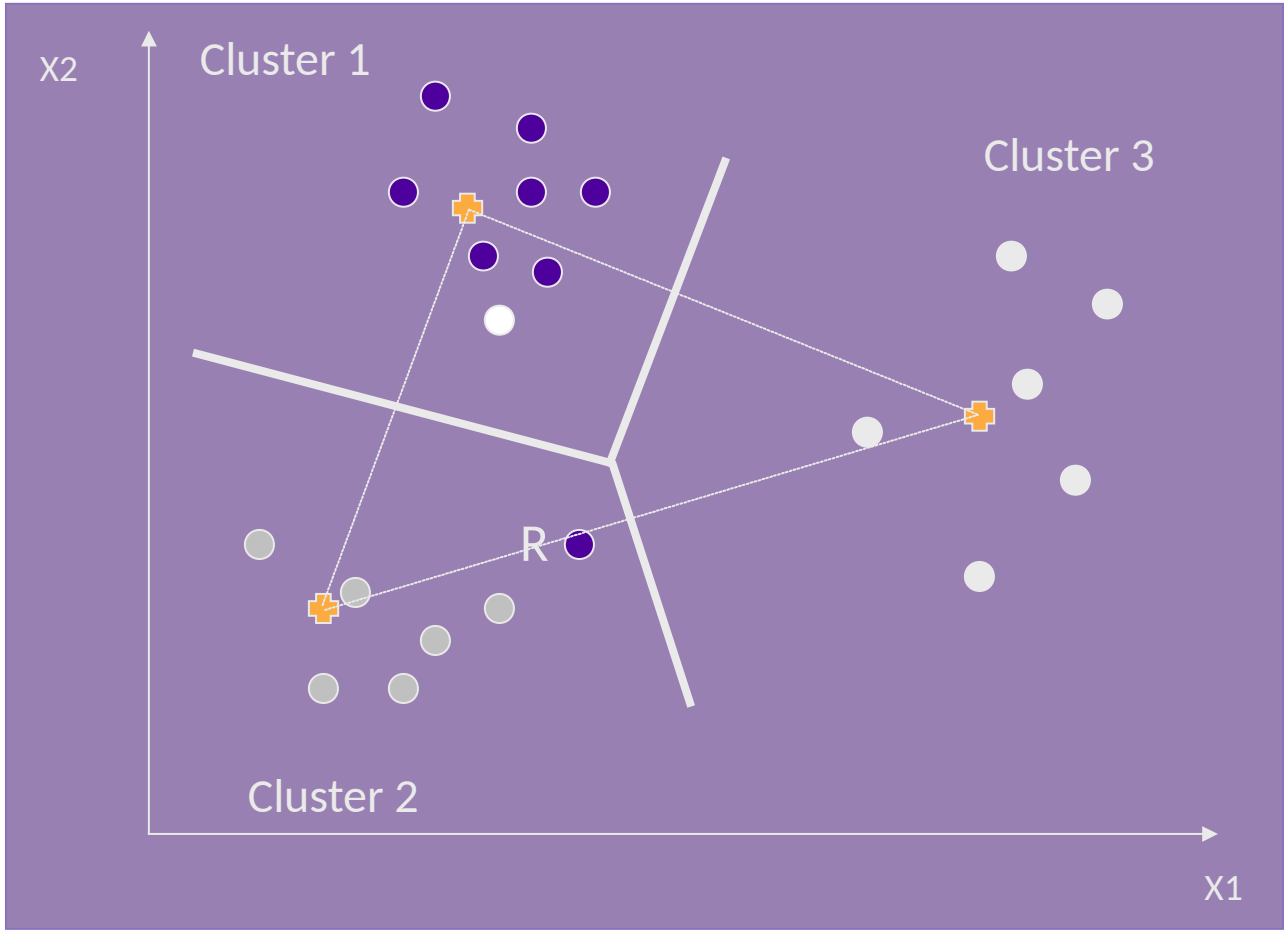
Clustering kMeans



Clustering kMeans



Clustering kMeans



k-Means

- Hiperparámetros:
  - Cantidad de clusters
  - Cantidad de iteraciones (ajuste de centroides)

Referencias

Frameworks de Data Mining:

- **KNIME**: https://www.knime.com/, IDE o Java jar
- **Apache Spark ML Lib**: https://spark.apache.org/mllib, Java jar + deploy sobre Apache Spark
- **TensorFlow**: https://www.tensorflow.org/?hl=es, Python Lib
- **R Packages**: CARET, NNET, KERNLAB
- R Lib + R Studio
- **Encog**: https://www.heatonresearch.com/encog/, Javascript lib, Java jar o C++ lib
- **DeepLearning4j**: https://deeplearning4j.org/, Java jar

Referencias (2)

Teoría de Data Mining y Estadísticas para Data Scientist:

- 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
- Bruce, P., Bruce A. Practical Statistics for Data Scientists: 50 Essential Concepts, 1st Ed, O'Really Media, 2017