Universidad Nacional del Altiplano
Escuela de Posgrado
Doctorado en Ciencias de la Computación

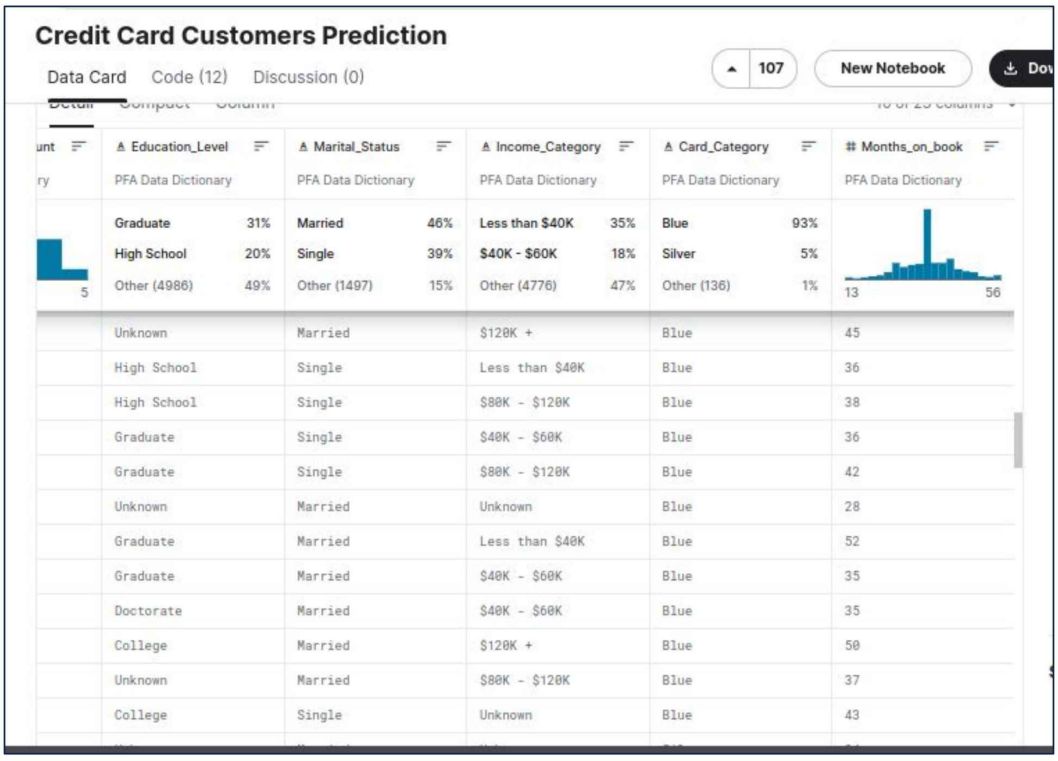
Minería de Datos

Prof. Dr. Ivar Vargas Belizario

ivargasbelizario@gmail.com

2024 - I

Structured data



Credit Card Customers Prediction

Data Card | Code (12) | Discussion (0) | 107 | New Notebook

Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book
PFA Data Dictionary	PFA Data Dictionary	PFA Data Dictionary	PFA Data Dictionary	PFA Data Dictionary
Graduate 31%	Married 46%	Less than $40K 35%	Blue 93%	13 – 56
High School 20%	Single 39%	$40K - $60K 18%	Silver 5%	
Other (4986) 49%	Other (1497) 15%	Other (4776) 47%	Other (136) 1%	
Unknown	Married	$120K +	Blue	45
High School	Single	Less than $40K	Blue	36
High School	Single	$80K - $120K	Blue	38
Graduate	Single	$40K - $60K	Blue	36
Graduate	Single	$80K - $120K	Blue	42
Unknown	Married	Unknown	Blue	28
Graduate	Married	Less than $40K	Blue	52
Graduate	Married	$40K - $60K	Blue	35
Doctorate	Married	$40K - $60K	Blue	35
College	Married	$120K +	Blue	50
Unknown	Married	$80K - $120K	Blue	37
College	Single	Unknown	Blue	43

Structured data

High-dimensional Data
Curse of Dimensionality

X^{nxm} y

	F1	F2	...	Fm	Label
n	0.3	0.5	0.8	0.4	0.5
	0.2	0.2	0.9	0.3	0.4
	...	...	...	...	...
	0.1	0.3	0.7	0.2	0.3

Label:
- Discrete or
- continuous

n: instances
m: features or attributes

Structured data

Características:

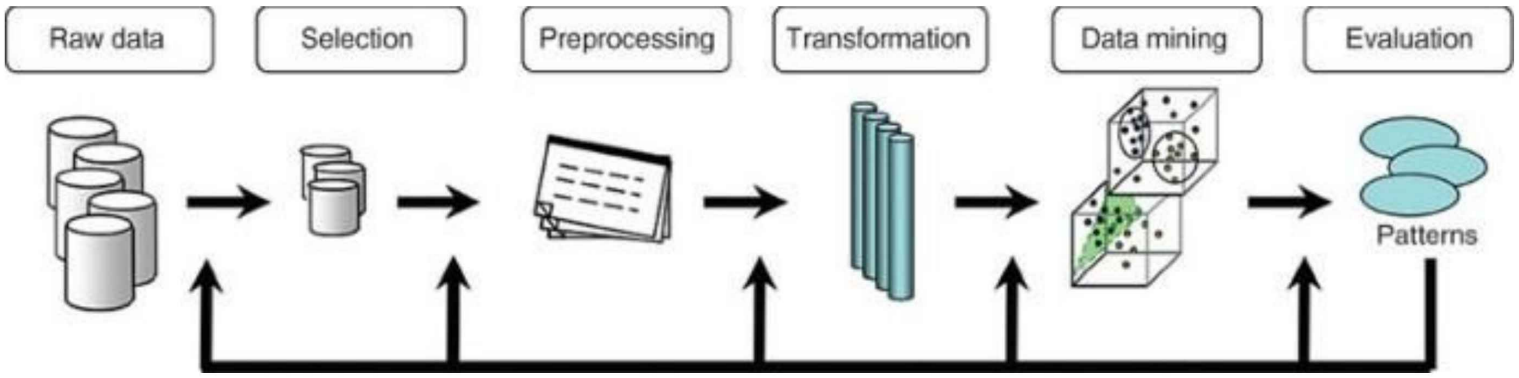
1. Grande tamaño de los datos:

1 Byte = 8 bits
1 Kilobyte (KB) = 1024 bytes
1 Megabyte (MB) = 1024 kilobytes
1 Gigabyte (GB) = 1024 megabytes
1 Terabyte (TB) = 1024 gigabytes
1 Petabyte (PB) = 1024 terabytes
1 Exabyte (EB) = 1024 petabytes
1 Zettabyte (ZB) = 1024 exabytes (2016-> tráfico en internet)
1 Yottabyte (YB) = 1024 zettabytes

2. Complexidad:

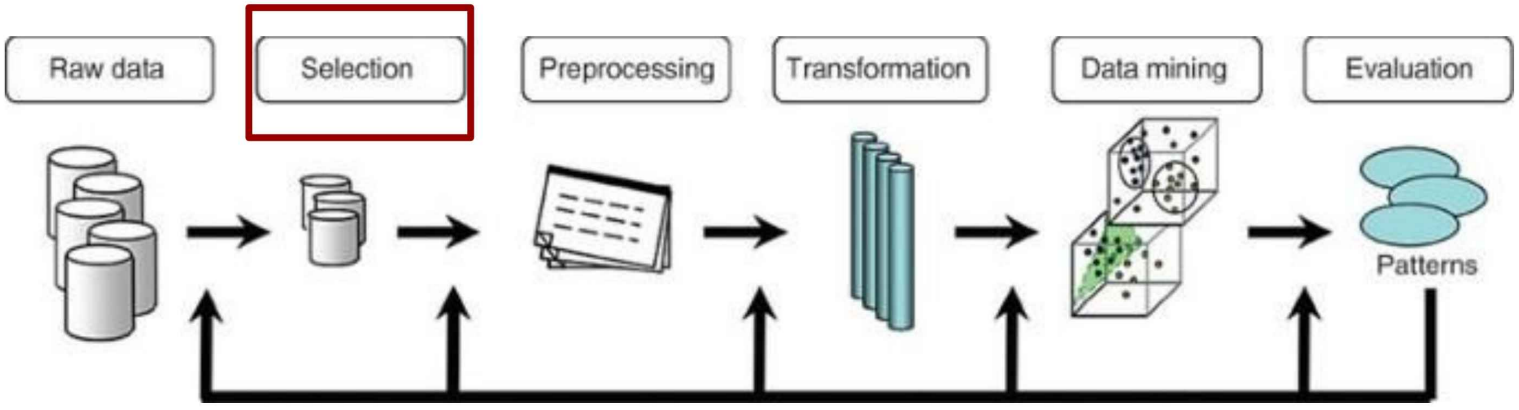
- **No estructurados: Imágenes, videos, audios, etc.**
- **Sufren variación en el tiempo**

Structured data (pipeline)



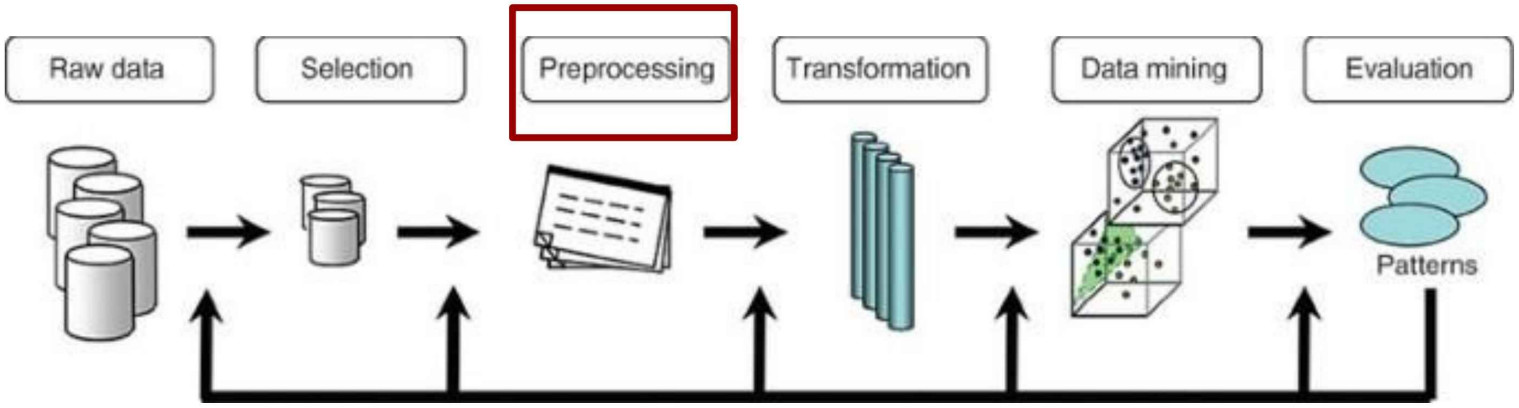
[4] https://doi.org/10.1007/978-1-4899-7993-3_1134-2

Structured data (pipeline)



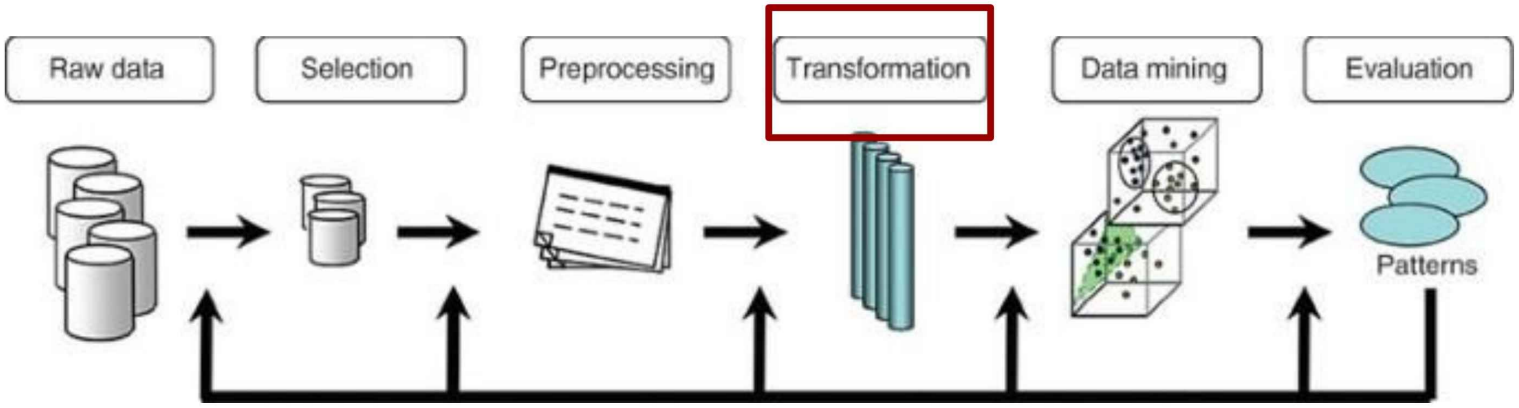
- Identifica el objetivo (problema)
- Selección con patrones relevantes

Structured data (pipeline)



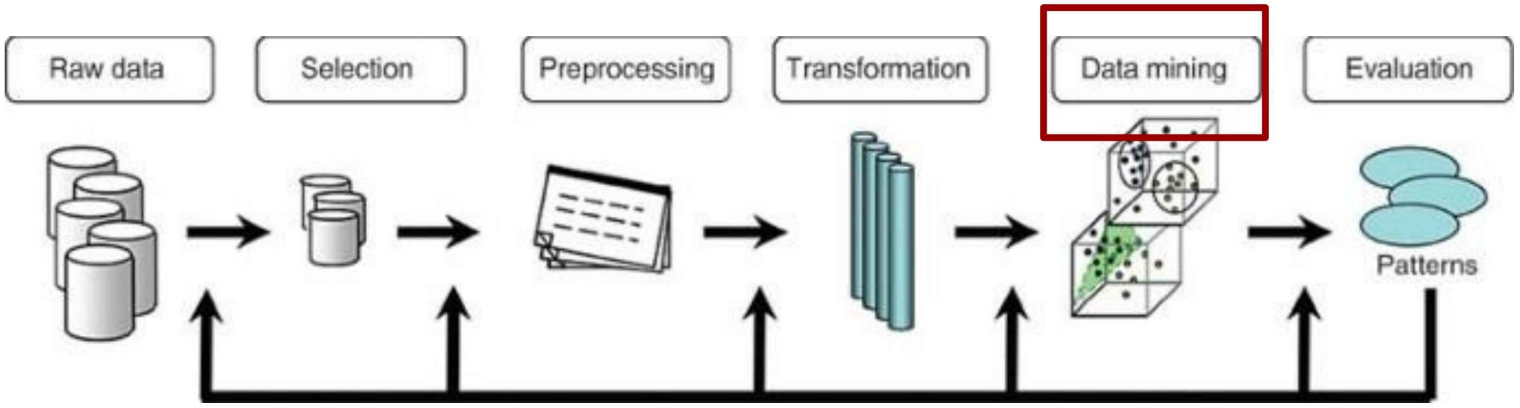
- Limpieza
- Aumentar la calidad de los datos:
  - Completar datos faltantes
  - Eliminar instancias duplicadas

Structured data (pipeline)



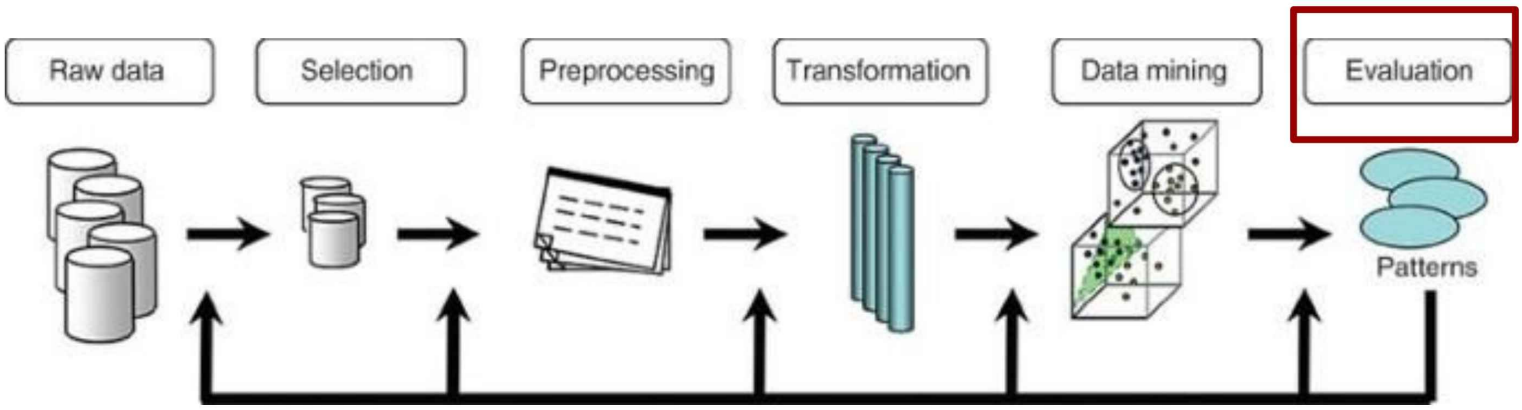
- Eliminar atributos correlacionados e irrelevantes.
- Se crea una nueva definición de atributos com mayor significancia.

Structured data (pipeline)



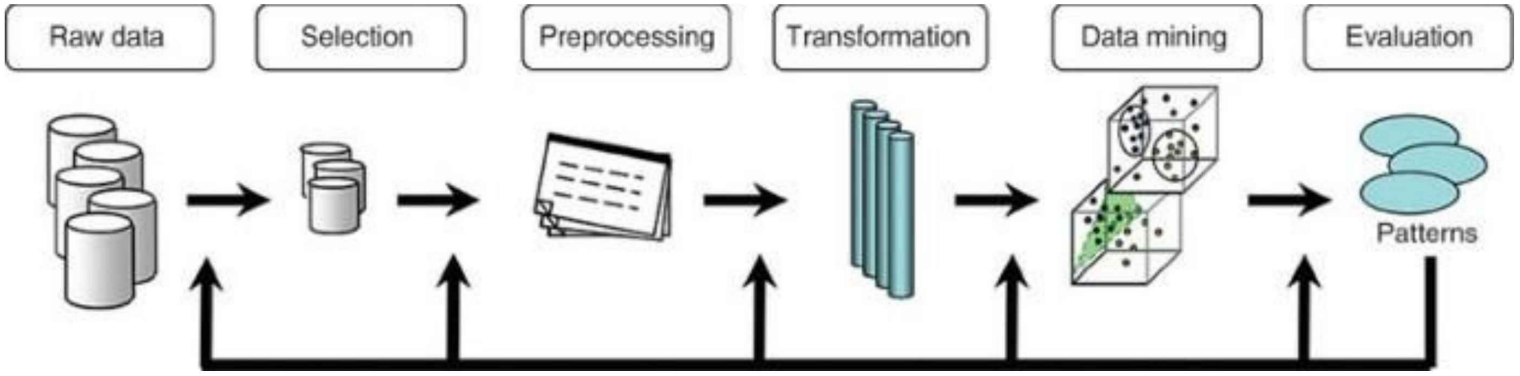
- Selecciona el tipo de algoritmo de minería de datos identificado en la etapa de **Selection**.
- Aprender patrones según:
  - Clasificación,
  - Agrupamiento
  - Regresión,
  - Asociación

Structured data (pipeline)



- Evalúa lo aprendido

Structured data



Goal and sampling | Cleaning | Dimensionality reduction | Learning patterns | Metrics: Precision, recall, accuracy, error, f1-score, etc.

Dimensionality reduction

X^{nxm} y

	F1	F2	…	Fm	Label
n	0.3	0.5	0.8	0.4	0.5
	0.2	0.2	0.9	0.3	0.4
	…	…	…	…	…
	0.1	0.3	0.7	0.2	0.3

X^{nxp} y

	F1	F2	…	Fp	Label
n	0.3	0.5	0.8	0.4	0.5
	0.2	0.2	0.9	0.3	0.4
	…	…	…	…	…
	0.1	0.3	0.7	0.2	0.3

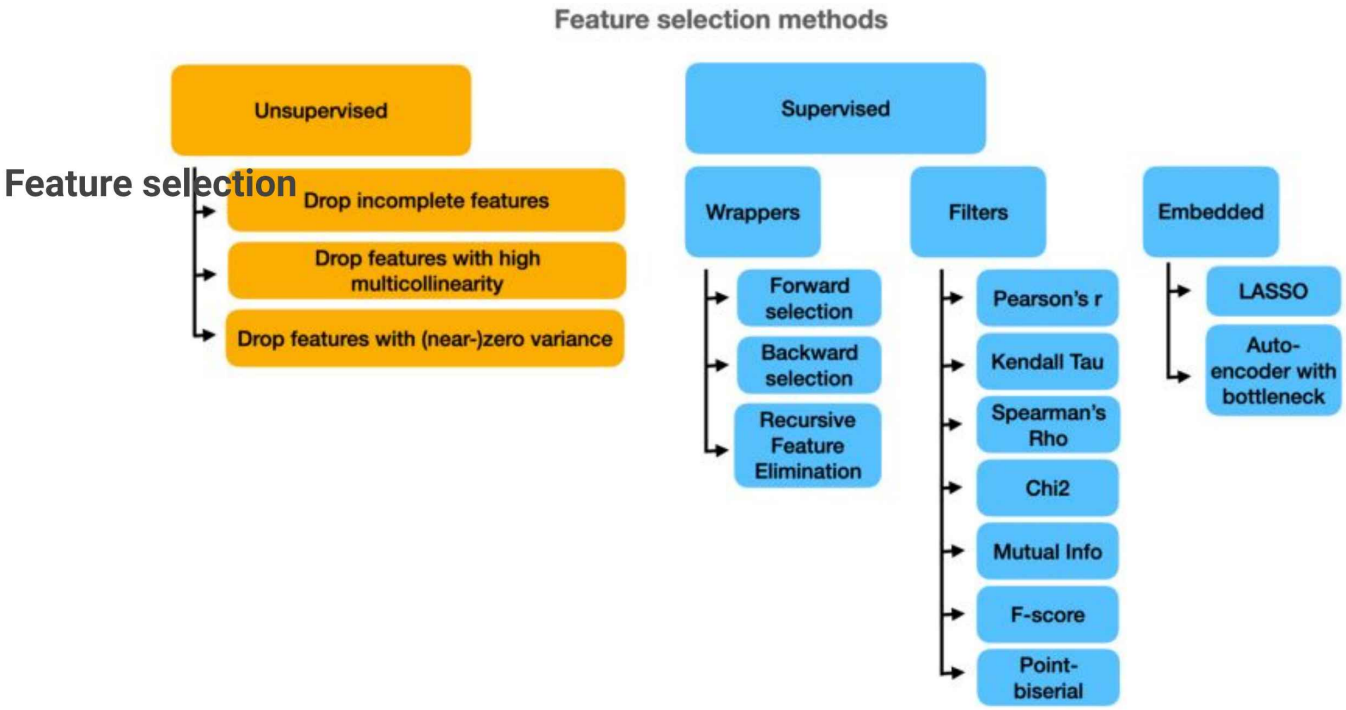
donde, p<m

Dimensionality reduction

Tipos

1. **Feature selection**: Los datos no varian, se seleccionan los mejores atributos manteniendo su integridad.
    i. Feature importance
    ii. Based in correlation matrix
2. **Feature extraction**: Los datos sí varían por fusión o transformación de características.
    i. PCA
    ii. T-SNE
    iii. UMAP
    iv. LSP

Dimensionality reduction



Dimensionality reduction

Feature extracti

- PCA
- T-SNE
- UMAP
- **LSP**

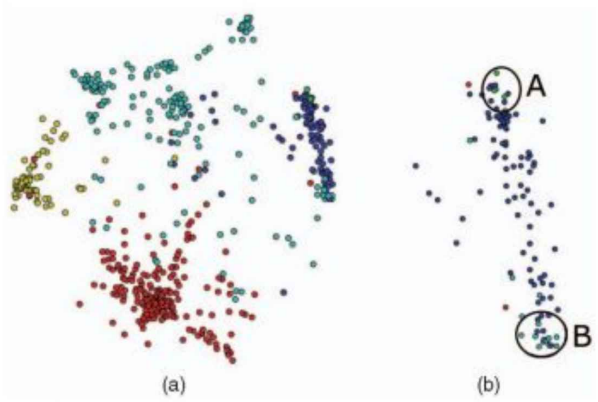


Fig. 2. Projection of a document collection composed of scientific papers in four different areas (colors indicate the areas). (a) Whole map. (b) Zoomed part.

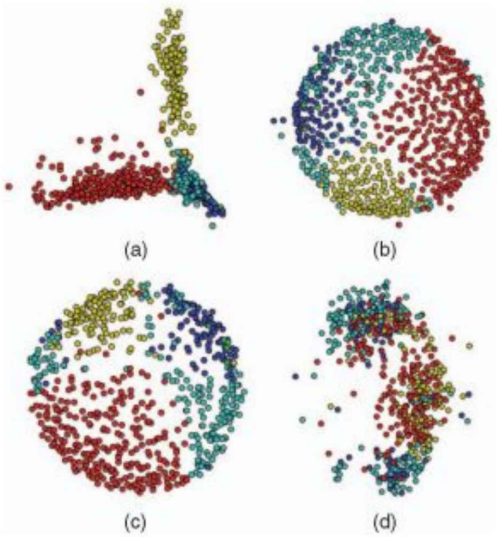
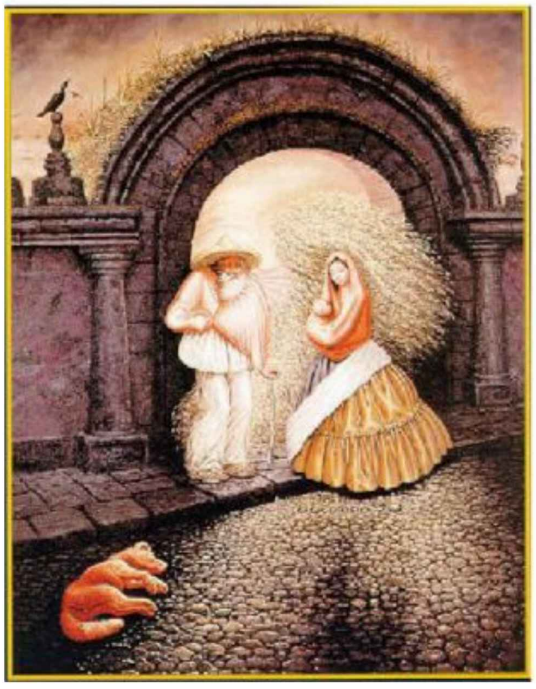


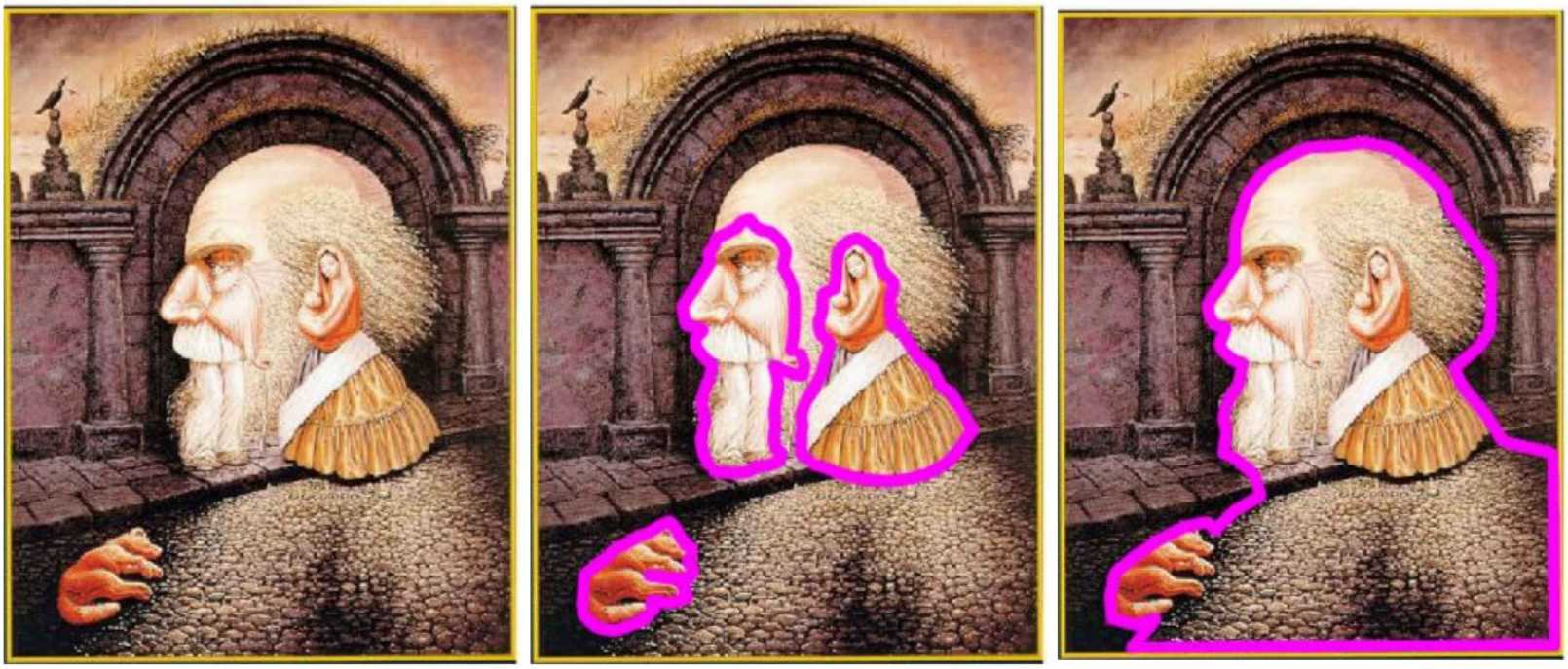
Fig. 11. Examples of projections generated using different techniques for the same data set used in the LSP projection presented in Fig. 2a. (a) PCA. (b) Sammon's mapping. (c) Original FDP model. (d) Approximated FDP model.

[5] https://doi.org/10.1109/TVCG.2007.70443

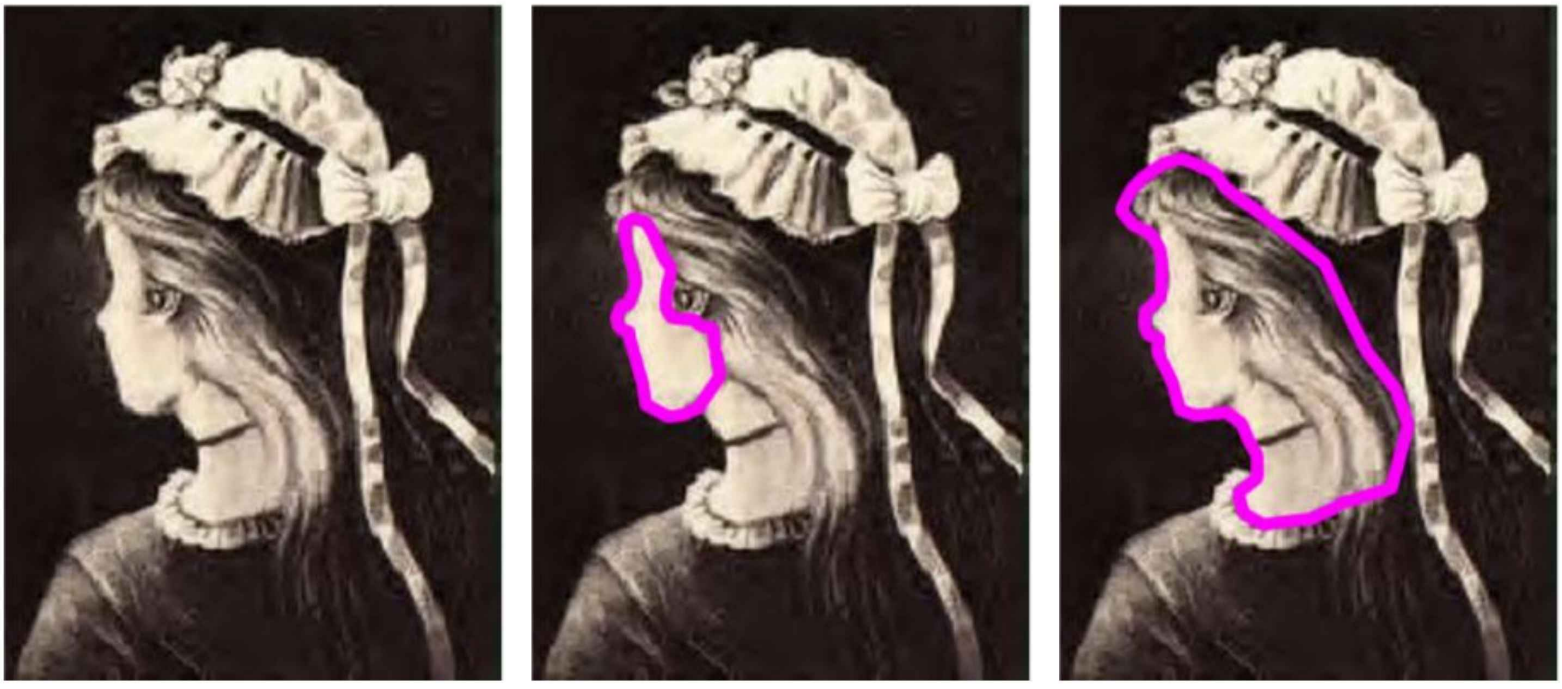
Clustering



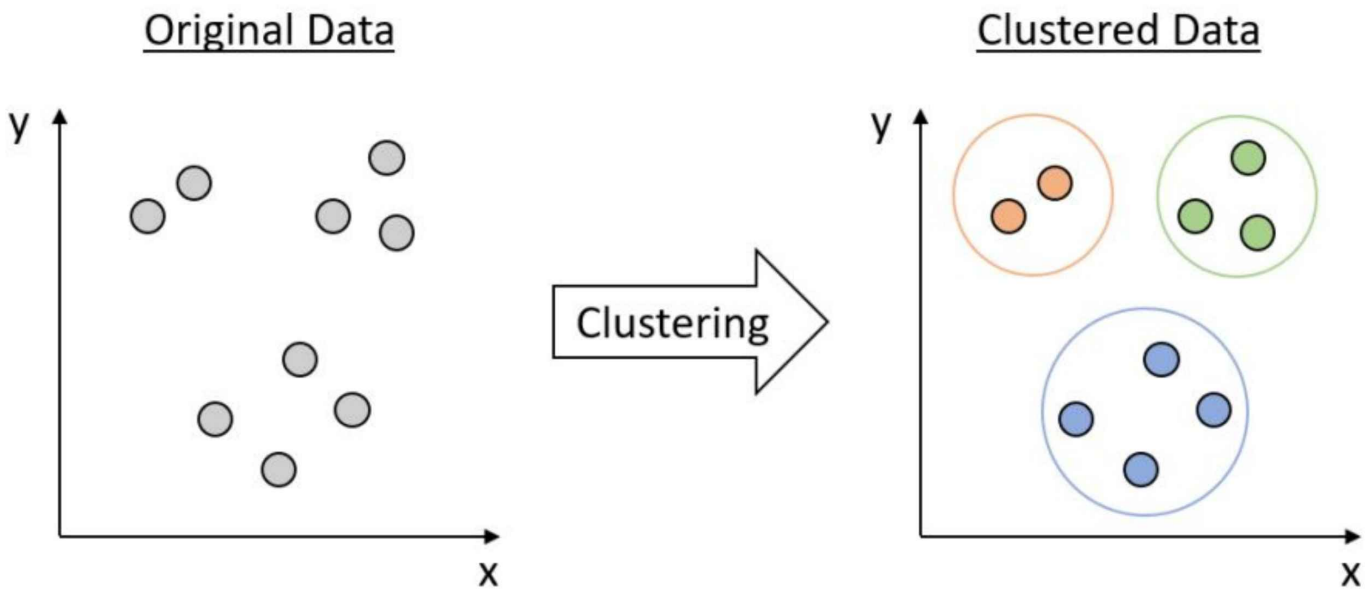
Clustering



Clustering



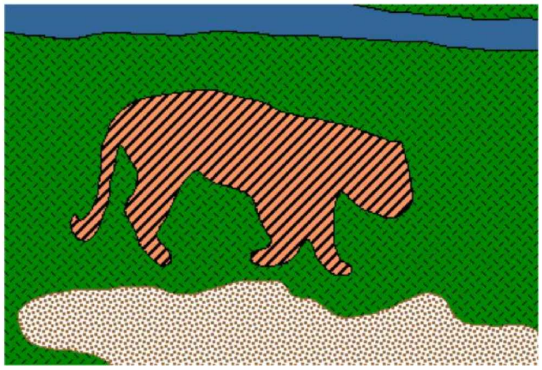
Clustering

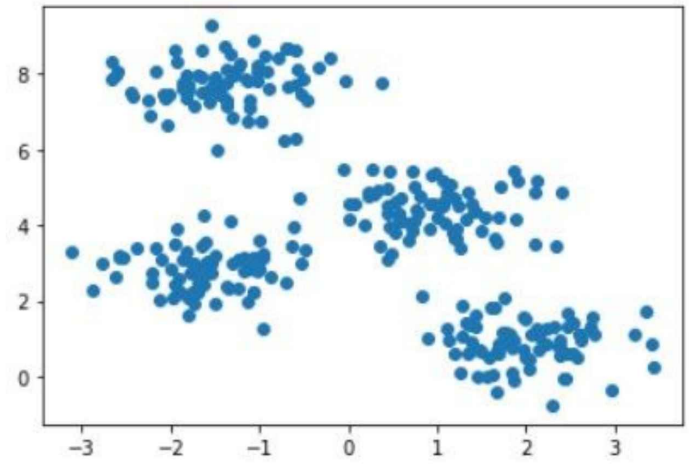


Clustering

- K-means clustering
- Hierarchical clustering
- Spectral clustering
- Mean shift algorithm
- etc.

Clustering - K-means clustering



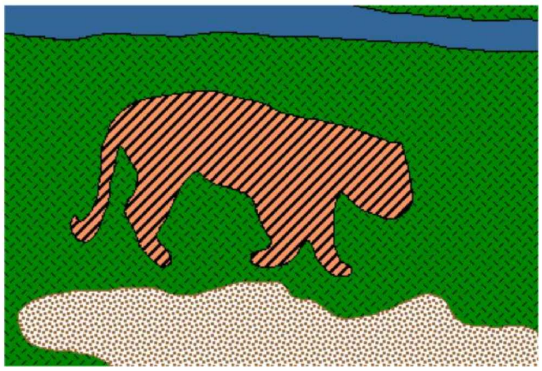


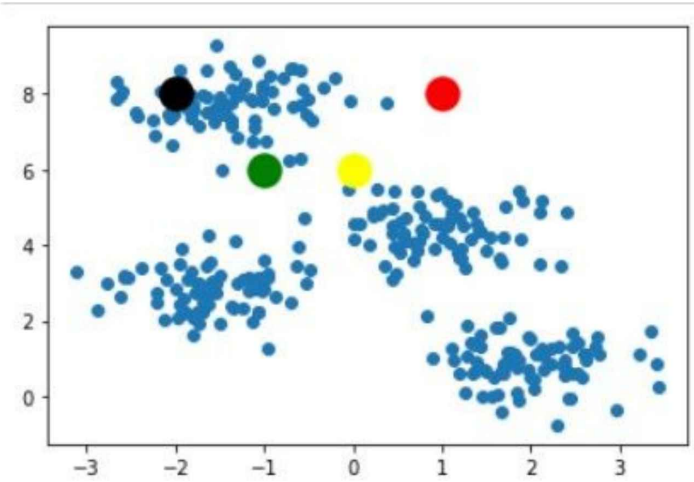
number of clusters — number of cases — case i — centroid for cluster j

objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

Distance function

Clustering - K-means clustering



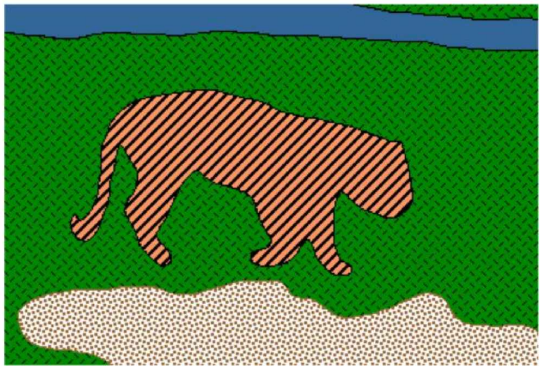


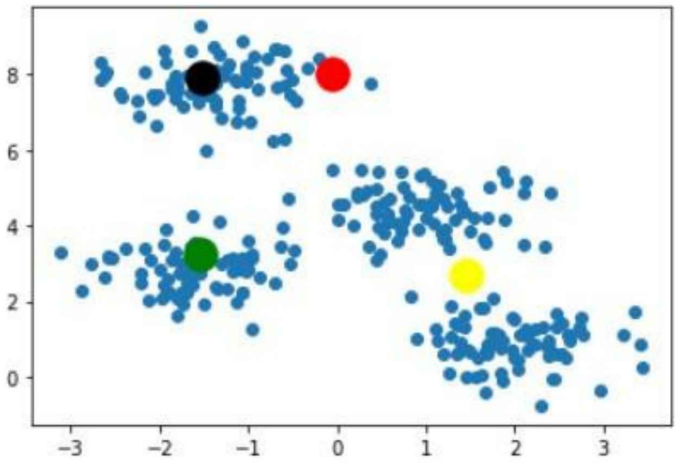
number of clusters — number of cases — case i — centroid for cluster j

objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

Distance function

Clustering - K-means clustering



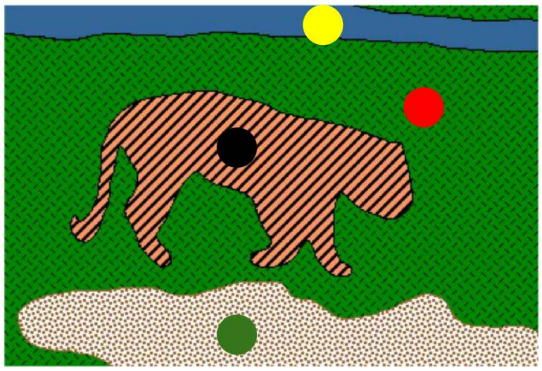


number of clusters — number of cases — centroid for cluster j — case i

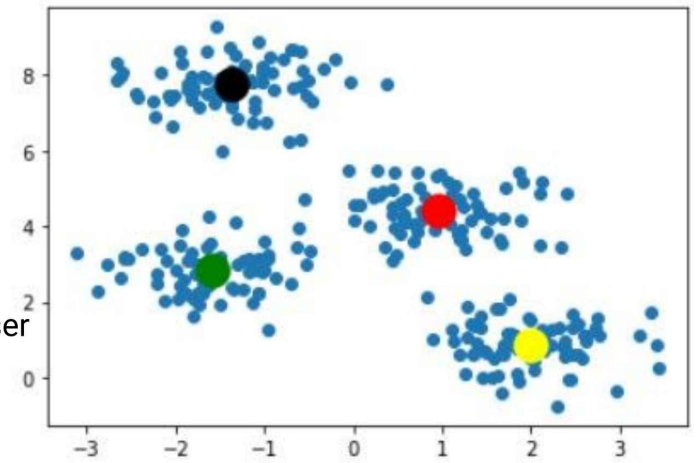
objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

Distance function

Clustering - K-means clustering



k = is indicated by the user

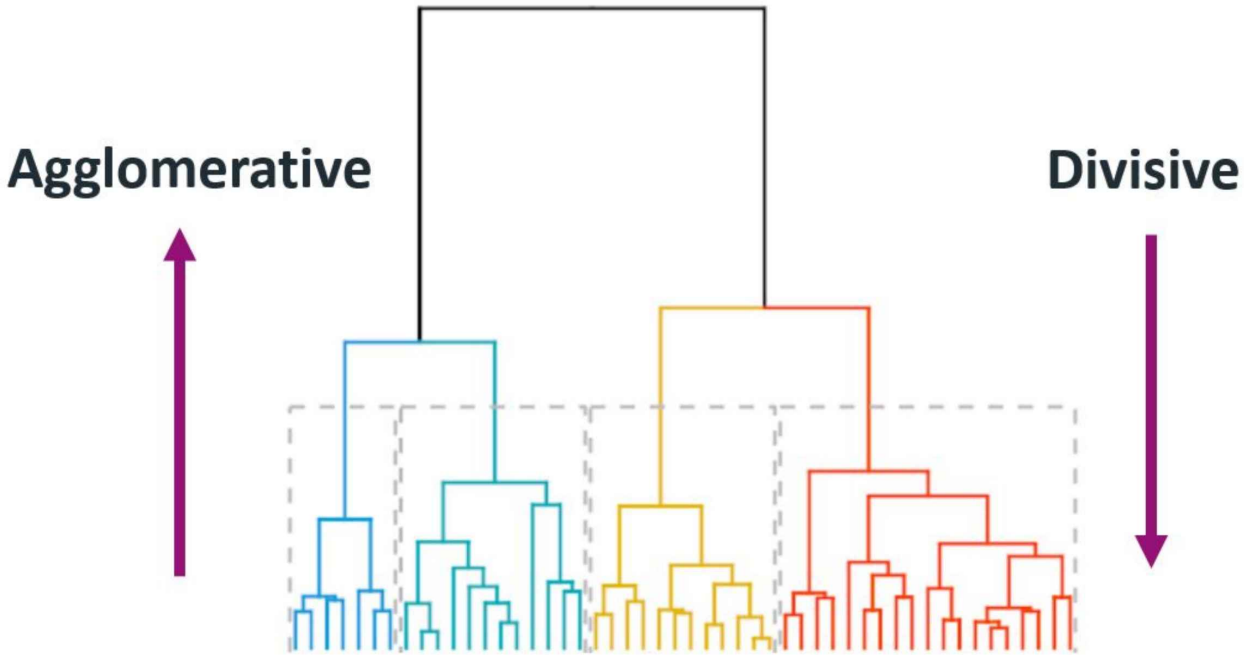


number of clusters — number of cases — centroid for cluster j — case i

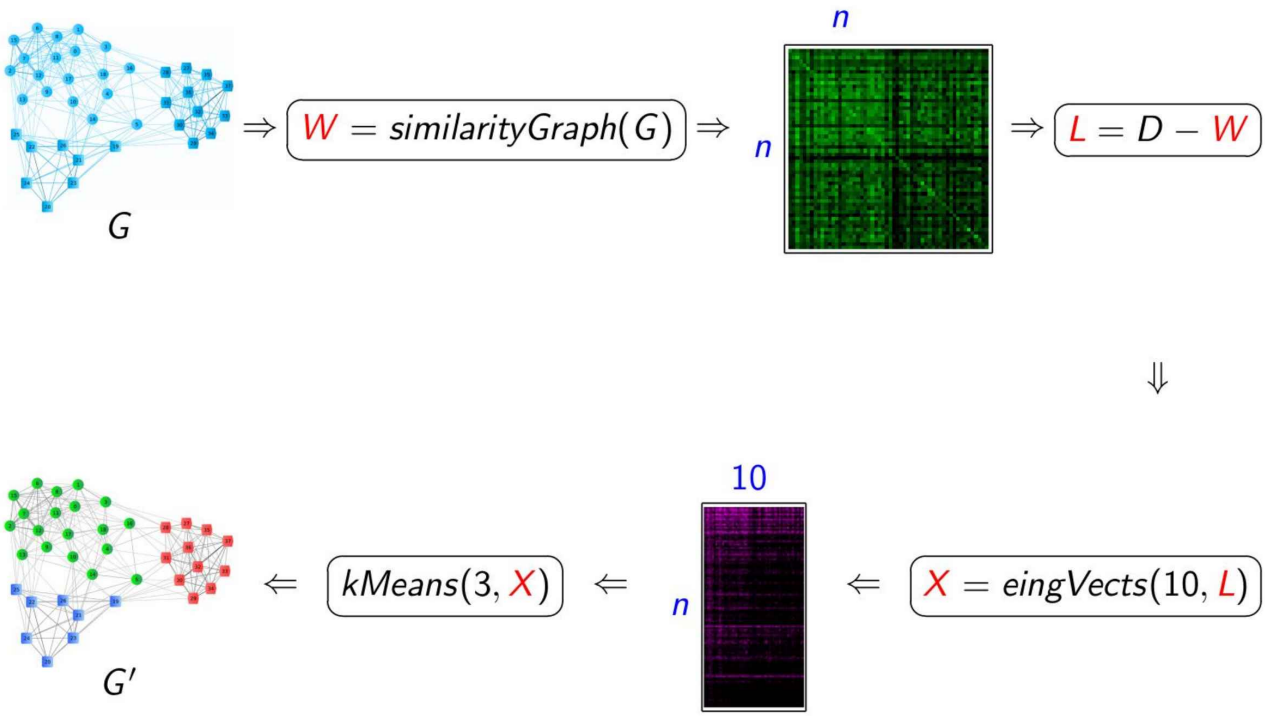
objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

Distance function

Clustering - Hierarchical clustering



Clustering - Spectral clustering



[6] https://doi.org/10.1109/34.868688

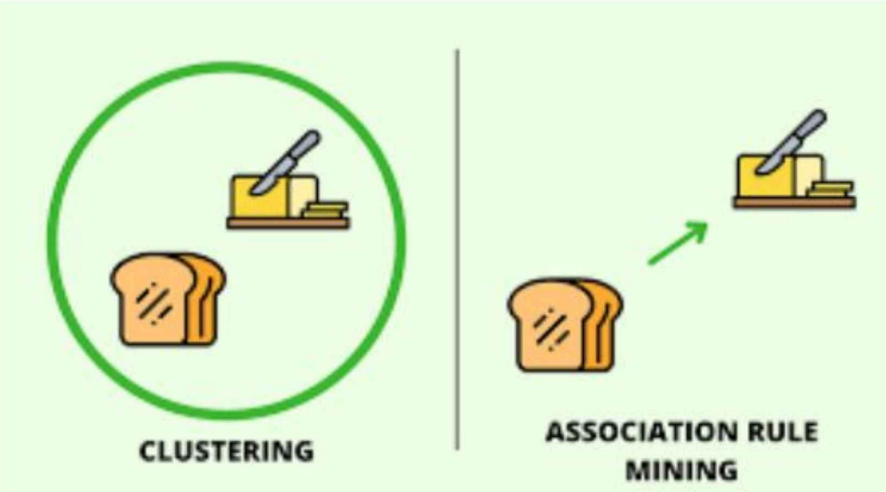
Association

{Bread, Egg} => {Milk}

Antecedent Consequent

Itemset = {Bread, Egg, Milk}

Association



Classification

Classification

X				y
F1	**F2**	**...**	**Fn**	**Label**
0.3	0.5	0.8	0.4	0
0.2	0.2	0.9	0.3	1
...	...	...	...	...
0.1	0.3	0.7	0.2	0

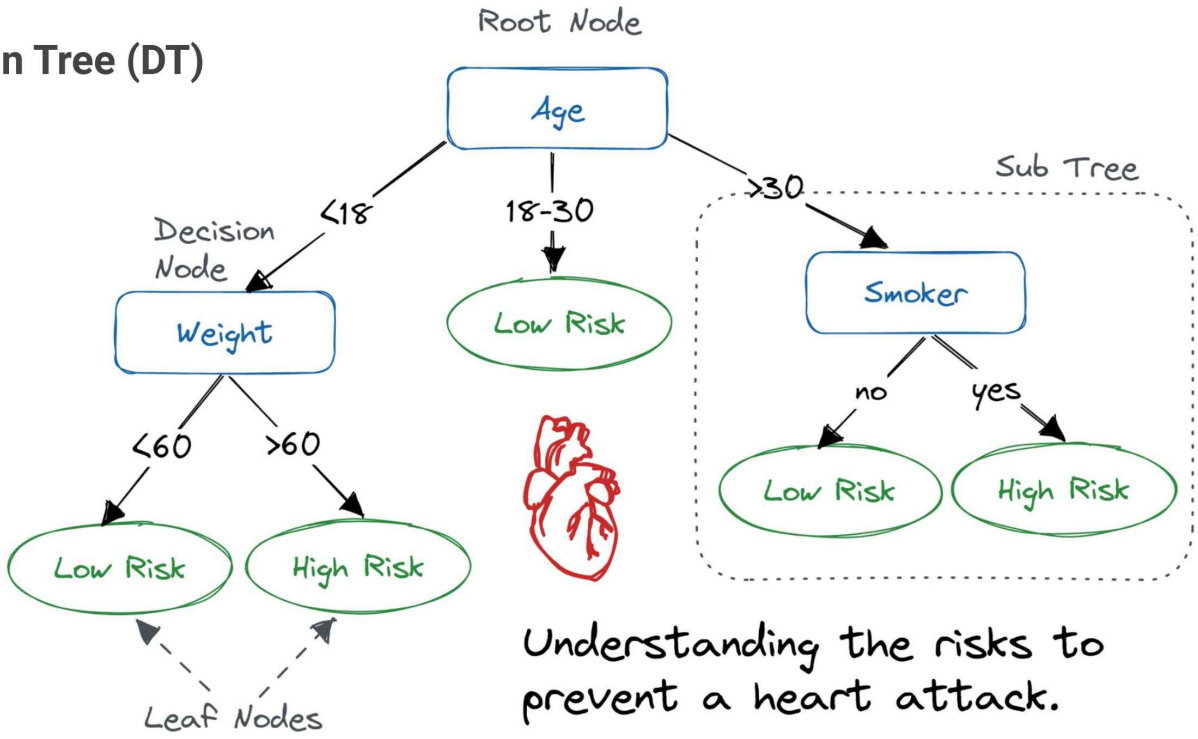
- y: son **datos discretos**, es decir número enteros positivos que representan objetos **finitos**.
  - 0: Cat; 1: Dog
  - 0: Cancer; 1: No cancer
  - 0: Virus; 1: Bacteria; 2: hongos

Classification

- Decision Tree (DT)
- Random Forest (RF)
- Multi-layer Perceptron (MLP)
- etc.

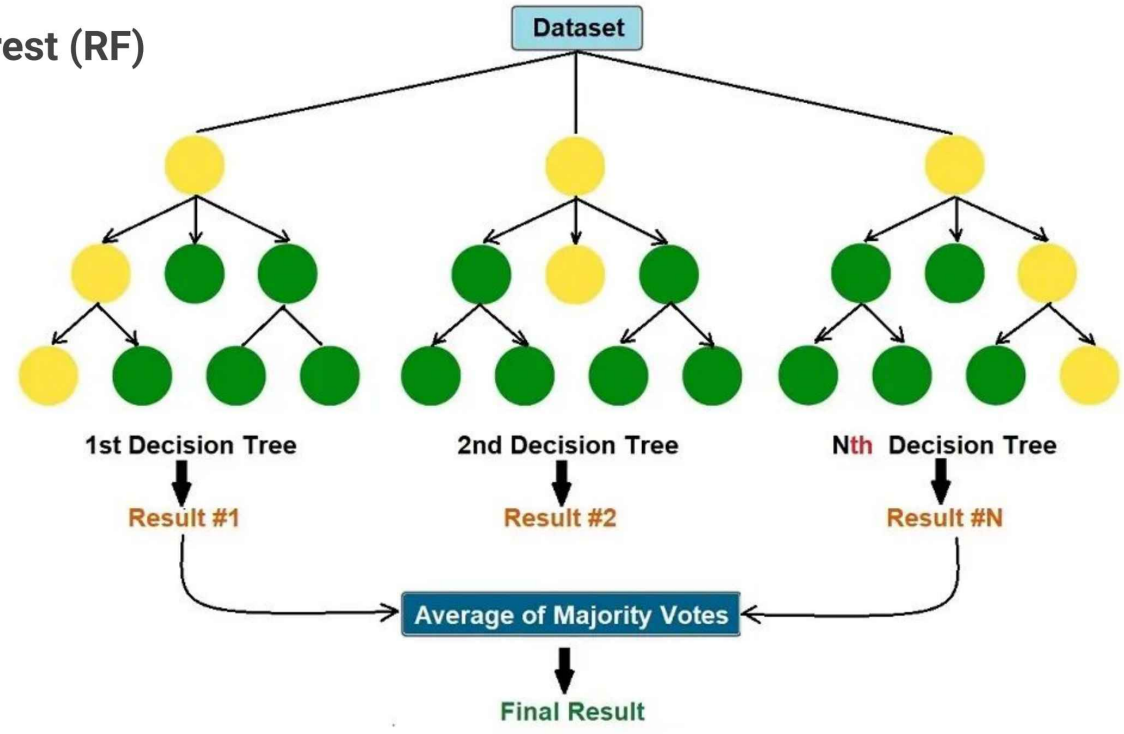
Classification

Decision Tree (DT)



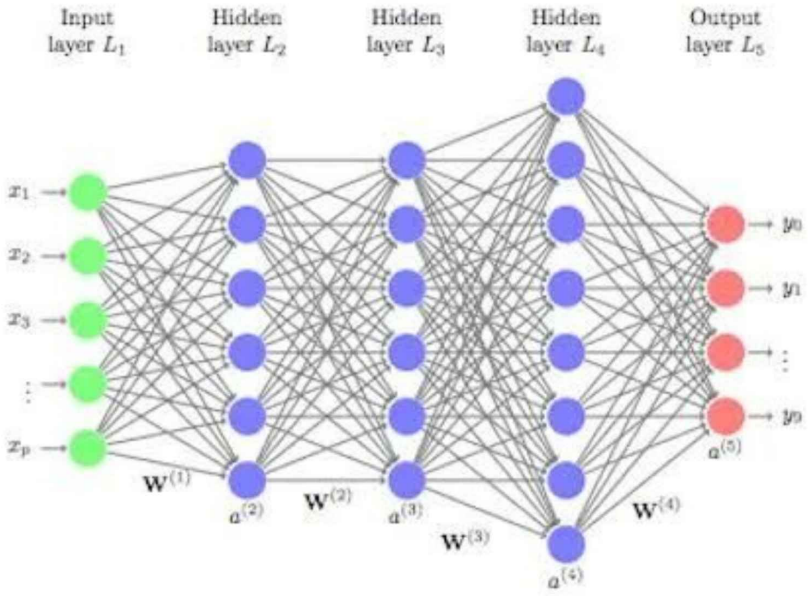
Classification

Random Forest (RF)



Classification

Multi-layer Perceptron (MLP)



Regression

Regression

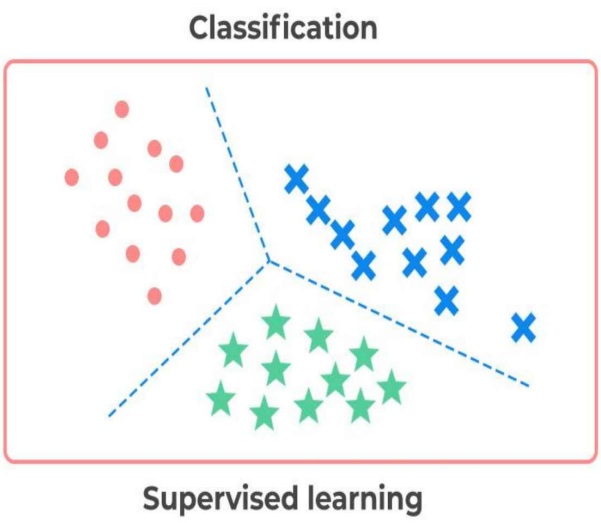
X				y
F1	**F2**	**...**	**Fn**	**Label**
0.3	0.5	0.8	0.4	0.5
0.2	0.2	0.9	0.3	0.4
...	...	...	...	...
0.1	0.3	0.7	0.2	0.3

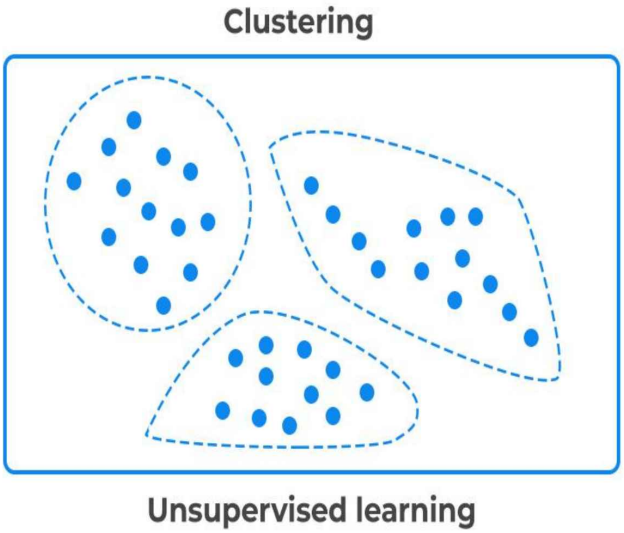
- Y: son datos **continuos** dentro de un **rango infinito** de valores. Generalmente son representados por valores reales.
  - Temperatura
  - Consumo de energia
  - Latitud, Longitud

Regression

- Decision Tree Regression (DTR)
- Random Forest Regression(RFR)
- Multi-layer Perceptron Regression (MLPR)
- etc.

Classification x Clustering

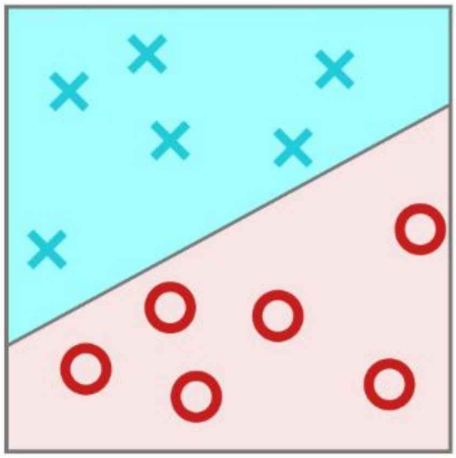




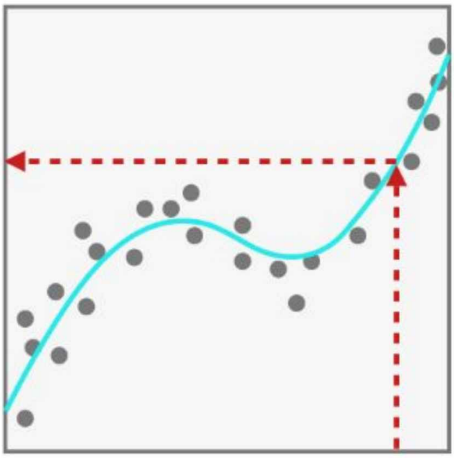
Se determina la clase (discreto)

Se determina el agrupamiento (discreto)

Classification x Regression



Classification

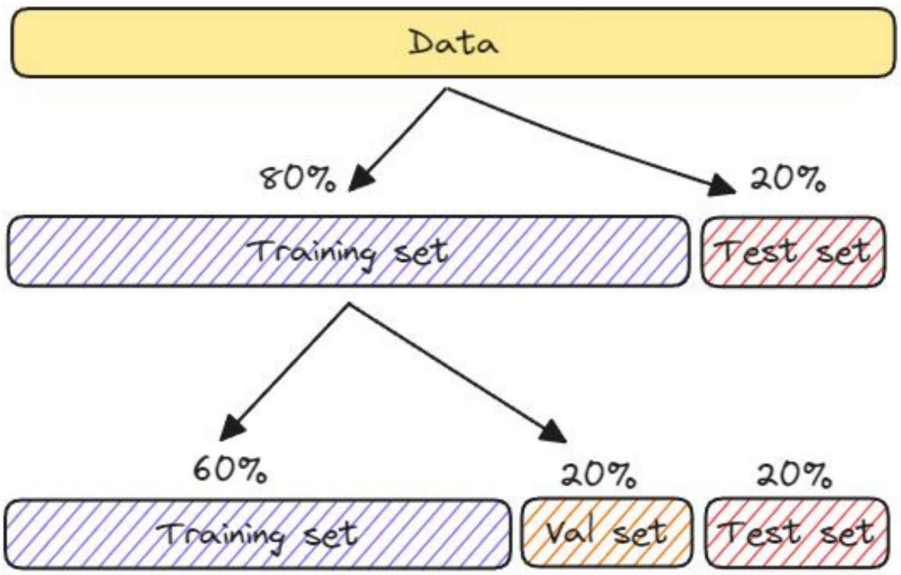


Regression

Se determina la clase (discreto)

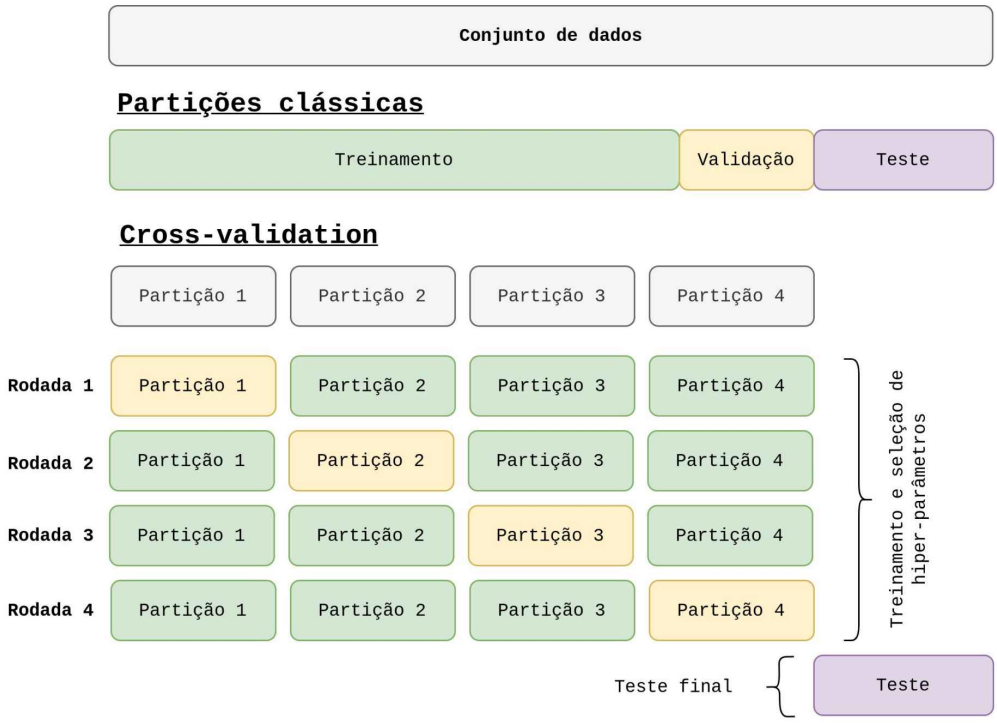
Se determina valores continuous

Classification and Regression



{X => y}

Classification and Regression



{X => y}

https://computacaointeligente.com.br/conceitos/avaliando-performance-cross-validation/

Examples

Digit Recognizer (profesor)
https://www.kaggle.com/competitions/digit-recognizer/

Shelter Animal Outcomes (Trabajo Encargado 1 - TE1)
https://www.kaggle.com/competitions/shelter-animal-outcomes/

Universidad Nacional del Altiplano
Escuela de Posgrado
Doctorado en Ciencias de la Computación

Minería de Datos

Gracias

Prof. Dr. Ivar Vargas Belizario

ivargasbelizario@gmail.com

2024 - I