



UNIVERSIDAD
DE GRANADA



Sistemas Inteligentes para la Gestión de la Empresa

Tema 2: Depuración y calidad de datos

E.T.S. de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Juan Gómez Romero
jgomez@decsai.ugr.es

Departamento de Ciencias de la
Computación e Inteligencia Artificial
<http://decsai.ugr.es>

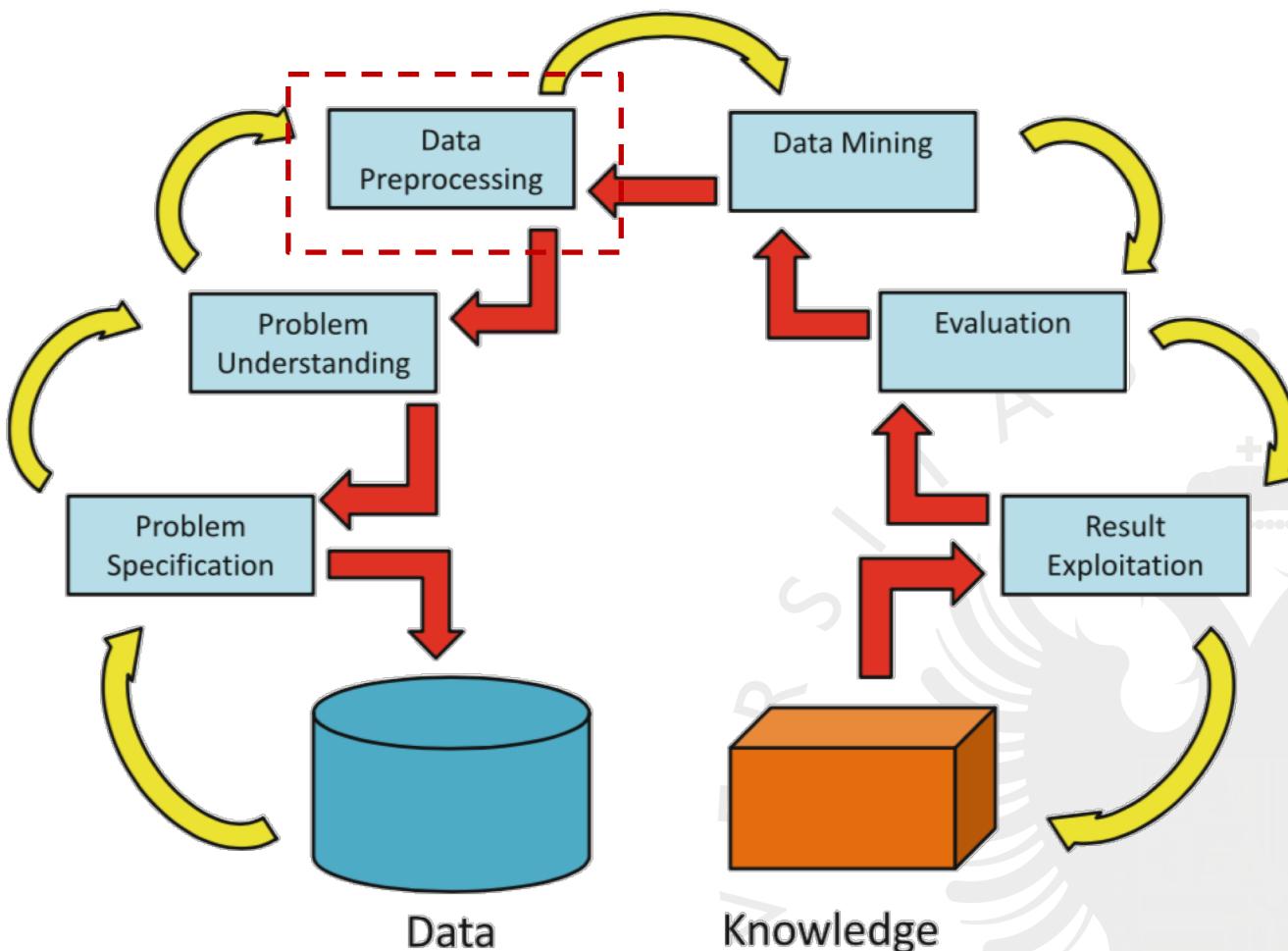
S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Springer.

G. Grolemund, H. Wickham (2017) *R for Data Science*. O'Reilly.

Tidyverse <https://www.tidyverse.org>

P. Casas (2018) *Data Science Live Book*
<https://livebook.datascienceheroes.com>

H. Wickham (2016) *Elegant Graphics for Data Analysis*. Springer.



**S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 1.
Introduction. Springer.**

Objetivo: Predecir si un pasajero sobrevive o no en función de una serie de variables relativas a la edad, género, etc.

Variables (9):

- **survival** = {0, 1}
- **pclass** = {1st, 2nd, 3rd}
- **sex**: sexo
- **age**: edad
- **sibsp**: número de parientes (hermano/a, hermanastro/a) / cónyuge (esposo/a) a bordo
- **parch**: número de padres (madre/padre) / hijos a bordo (hijo/a, hijastro/a)
- **ticket**: número de ticket
- **fare**: precio del ticket
- **cabin**: número del camarote
- **embarked**: puerto de embarque

Datasets:

- <https://www.kaggle.com/c/titanic/data>

Métrica:

- % de pasajeros correctamente clasificados (*accuracy*)

Formato:

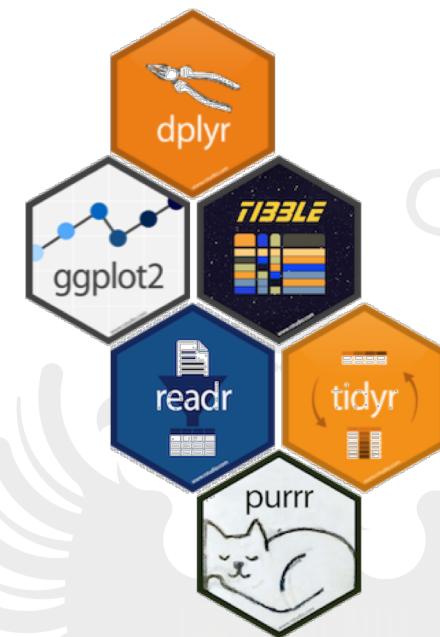
- Entrenamiento (*train.csv*)
- Validación (*test.csv*):
 - No incluye el valor objetivo
 - Enviar un fichero .csv con cabecera + 418 entradas
 - Cada entrada incluye dos columnas: PassengerId, Survived
 - Ejemplo: <https://www.kaggle.com/c/titanic/data>

Tutoriales:

- Kaggle R tutorial on Machine Learning (Datacamp)
<https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learning-the-sinking-of-the-titanic#gs.null>
- Getting started with R (Trevor Stephens) <http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/>
- Exploring the Titanic Dataset (Megan L. Risdal)
<https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic/notebook>

tidyverse

- <https://www.tidyverse.org>
- Documentación: *R for Data Science* website
<http://r4ds.had.co.nz>
 - dplyr: *pipes*
 - ggplot2: gráficos
 - tibble: *data.frame* mejorado
 - readr: lectura de ficheros
 - tidyr: funciones de transformación de datos
 - purrr: programación funcional



Preprocesamiento

Conjunto de tareas destinadas a la preparación de los datos previas al uso de algoritmos de extracción de conocimiento.

Dificultad

Proceso manual – consume el 60-80% del tiempo dedicado al análisis de datos (Adriaans & Zantinge, 1996)

Objetivos

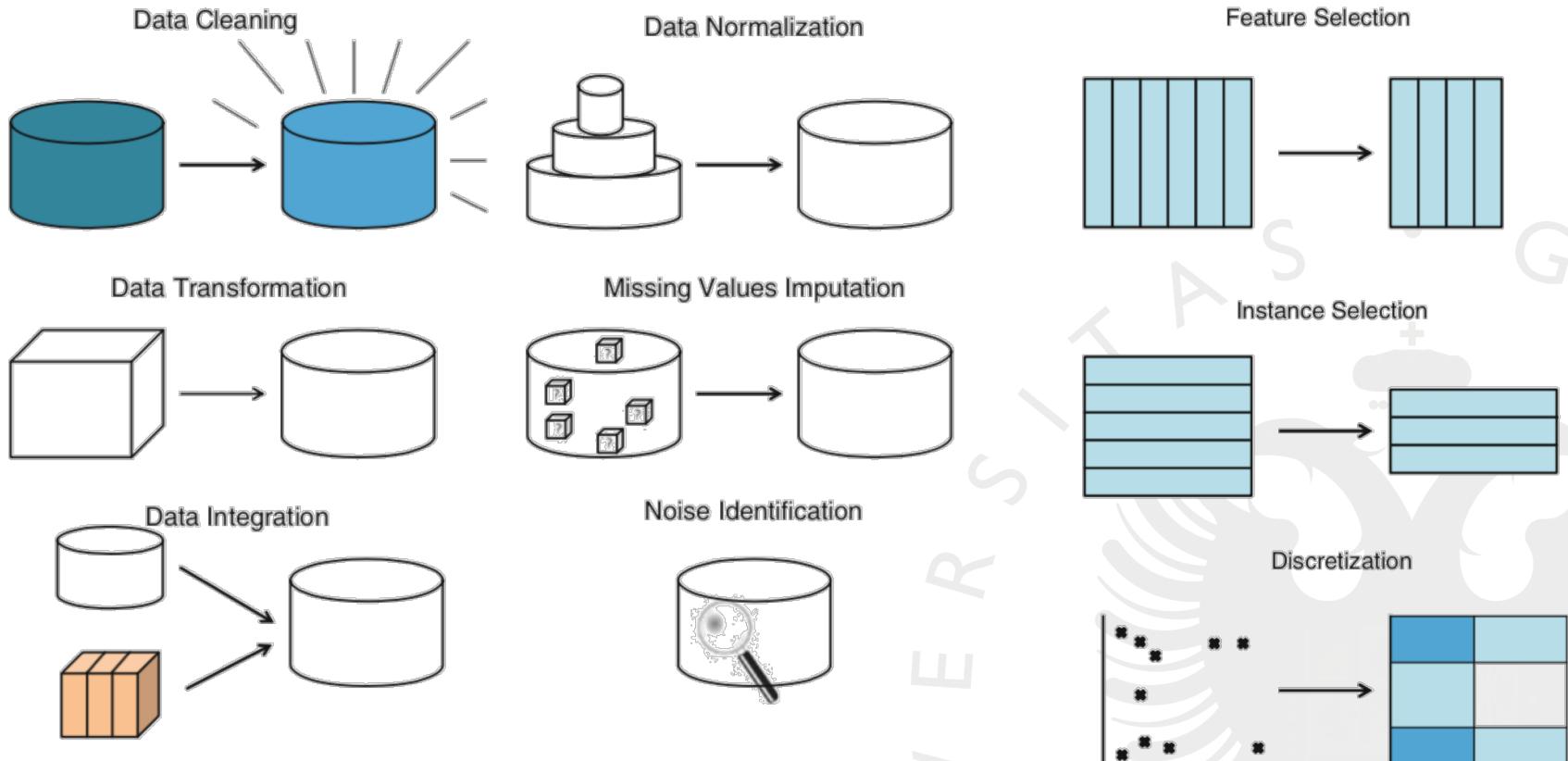
Mejora de la calidad

incompletos, ruido, inconsistentes

Reducción del tamaño

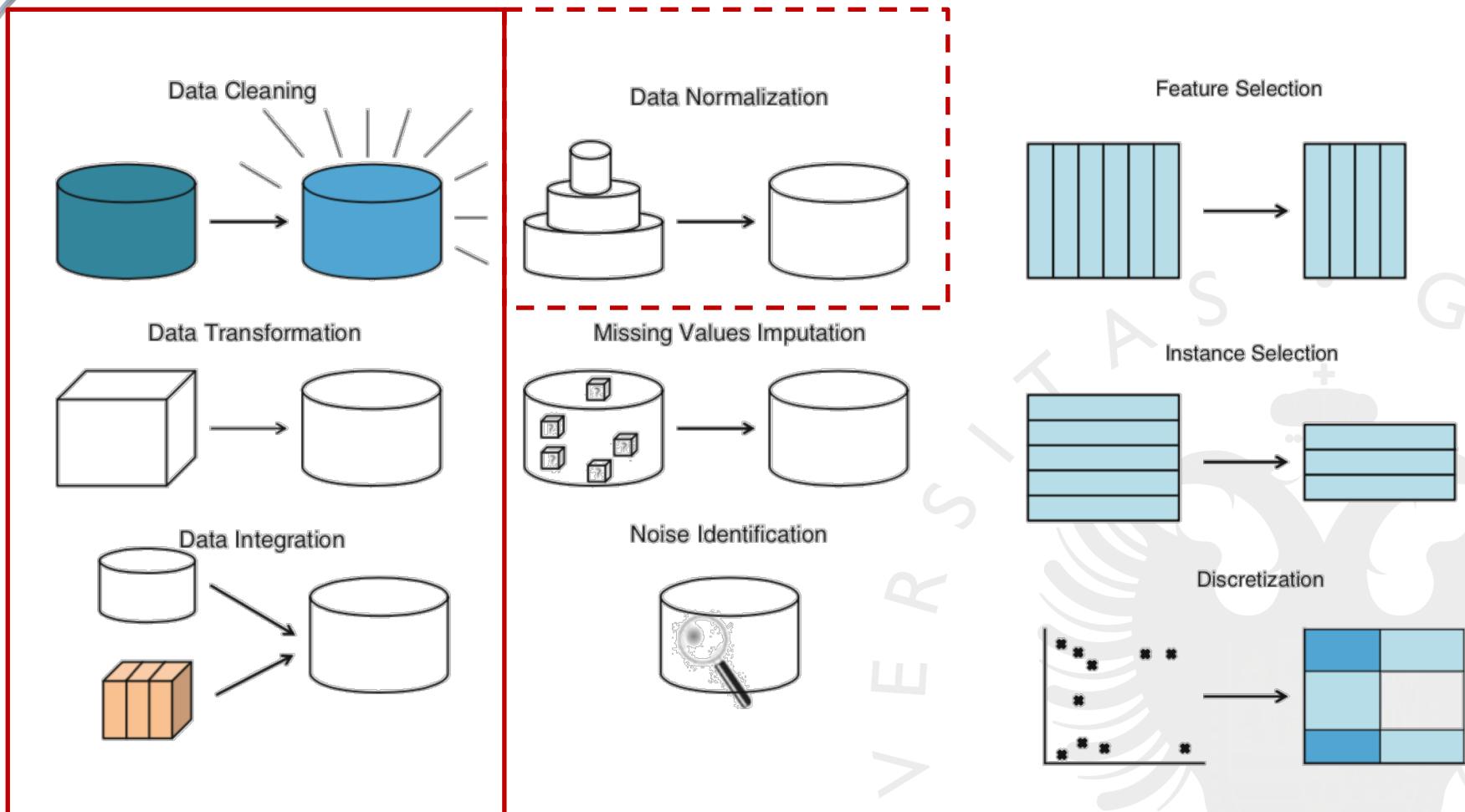
selección, eliminación, discretización

Pre-procesamiento de datos



S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 1. Introduction. Springer.

Integración, limpieza y transformación



S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 1. Introduction. Springer.

Integración

Combinación de datos de diferentes fuentes

Similar a ETL (*extraction, transformation & load*)

Integración de esquema

Unificación de la codificación

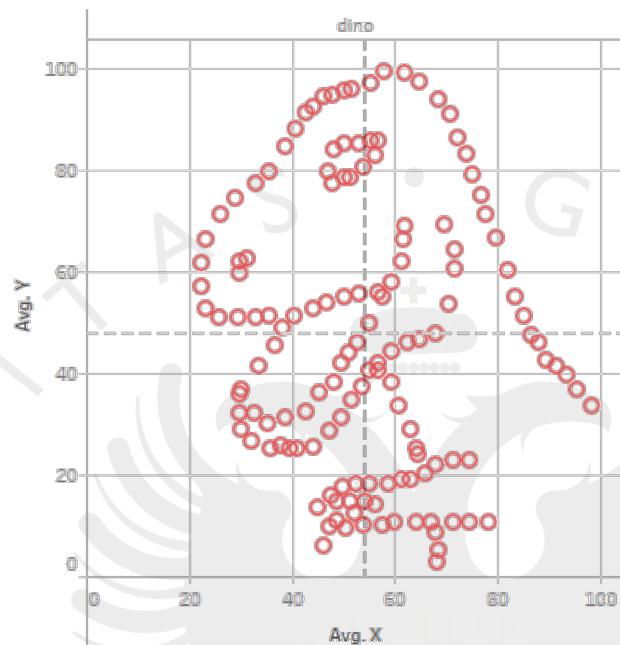
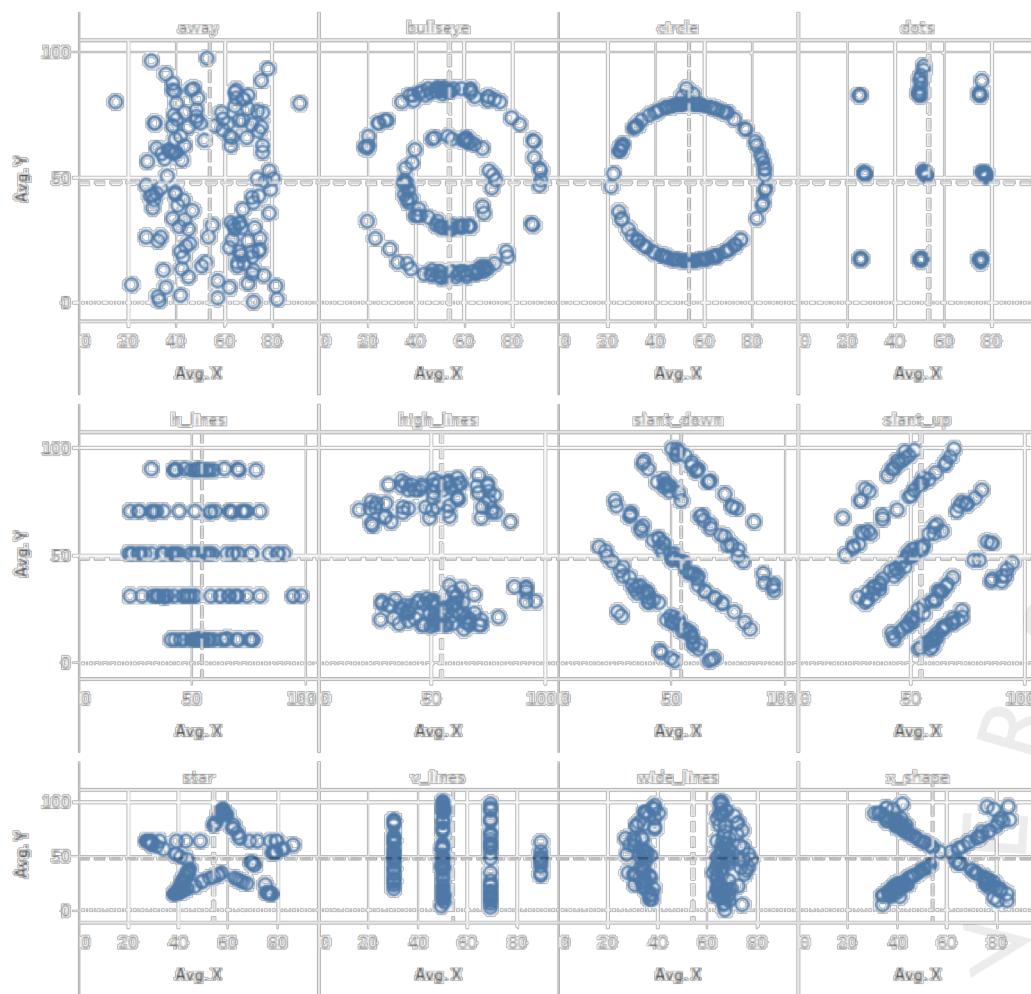
Detección de duplicados e inconsistencias

Redundancias

Análisis de correlaciones

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B}$$

Datasaurus Dozen



J. Matejka, G. Fitzmaurice (2017) A taxonomy of dirty data. ACM SIGCHI Conference on Human Factors in Computing Systems. (Online: <https://www.autodeskresearch.com/publications/samestats>)

Limpieza

Resolver inconsistencias

Rellenar valores perdidos (*)

Suavizar ruido (*)

Identificar *outliers*

...

W. Kim, B. Choi, E.-D. Hong, S.-K. Kim (2003) A taxonomy of dirty data. *Data Mining and Knowledge Discovery* 7, 81-99.

Transformación

Convertir, derivar, resumir...

Agregación

Generalización

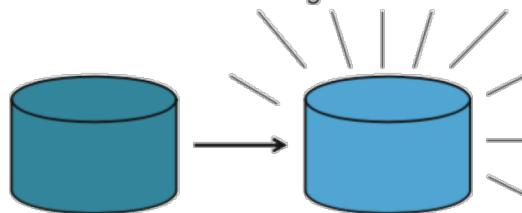
Normalización

Otras transformaciones

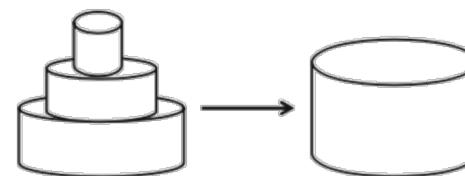
T. Y. Lin (2002) Attribute Transformation for Data Mining I: Theoretical Explorations. *International Journal of Intelligent Systems* 17, 213-222.

S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 3 Data Preparation Basic Model. Springer.

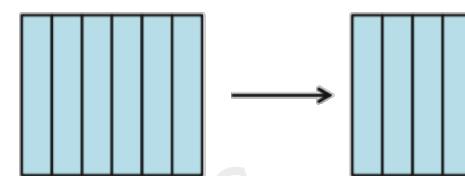
Data Cleaning



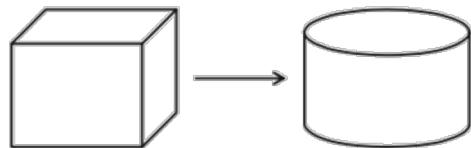
Data Normalization



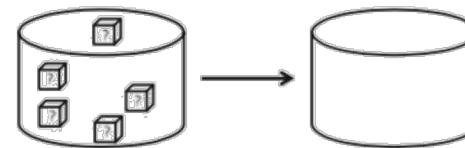
Feature Selection



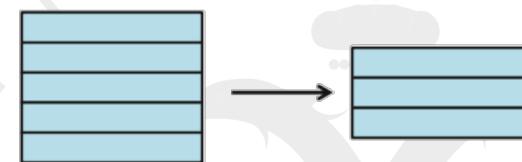
Data Transformation



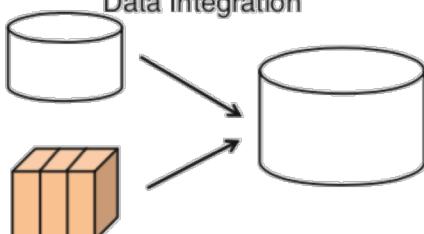
Missing Values Imputation



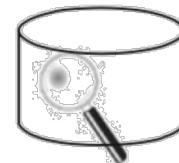
Instance Selection



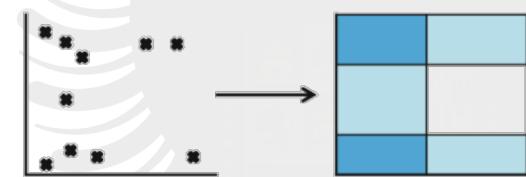
Data Integration



Noise Identification



Discretization



Datos imperfectos

S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 1. Introduction. Springer.

Valores perdidos

No disponibles

Eliminar

Asignar manualmente

Asignar valor global

Rellenar con media/desviación

Rellenar con valor más probable

S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 4 Dealing with Missing Values. Springer.

S. Van Buuren (2018) multiple-imputation.com (<http://stefvanbuuren.nl/mi/software.html>)

Valores perdidos

MICE

<https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/>

Multivariate Imputation by Chained Equations

- Software para imputar valores perdidos o incompletos basado en FCS (*fully conditional specification*)
 - Sustituir valores perdidos por “valores más probables”, estimados mediante inferencia a partir del resto del dataset

Facilita:

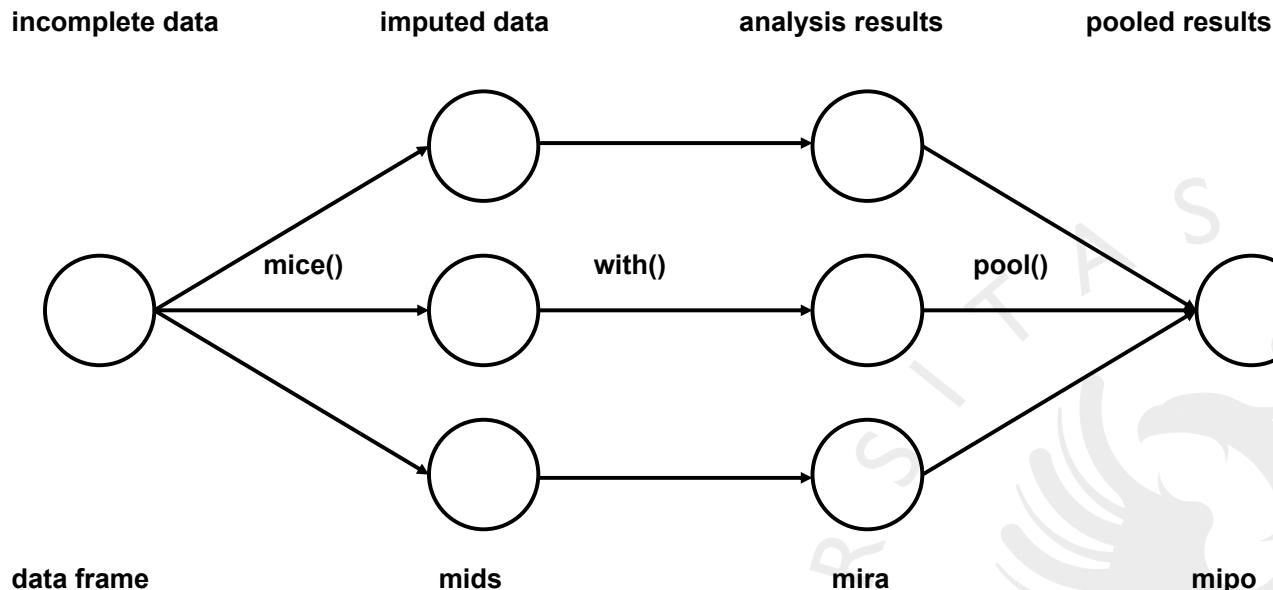
- Crear modelos predictivos para imputar valores perdidos
- Utilizar datasets con diferentes tipos de imputaciones para el análisis (*pooling*)

Idealmente, el modelo de predicción debe:

- Tener en cuenta el proceso que generó los valores perdidos
- Mantener las relaciones entre los datos
- Manejar incertidumbre de esas relaciones

Alternativas

- Amelia, caret (kNN)



S. van Buuren, K. Groothuis-Oudshoorn (2011) `mice`: Multivariate Imputation by Chained Equations in *R. Journal of Statistical Software*, 45(3), 1 - 67.

Valores perdidos

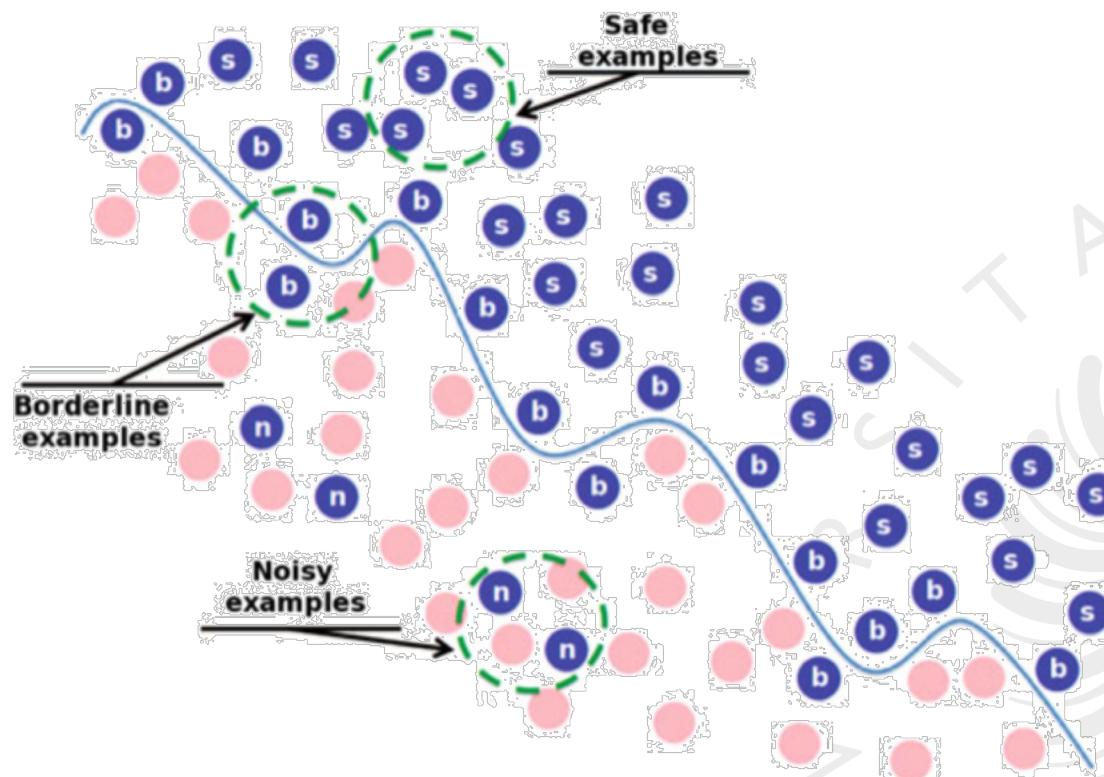
MICE

<https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/>

Selección del modelo de imputación

1. MAR (*Missing at Random*) vs MNAR (*Missing Not at Random*)
2. Forma del modelo de imputación (estructura, distribución del error)
3. Conjunto de variables que se usarán como predictores
4. Imputar variables derivadas de variables incompletas
5. Orden de imputación de las variables
6. Imputaciones iniciales y número de iteraciones
7. Número de *datasets* a la salida

Valores con ruido



S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 5 Dealing with Noisy Data. Springer.

Valores con ruido

NoiseFiltersR

<https://cran.r-project.org/web/packages/NoiseFiltersR/index.html>

Implementación de algoritmos de preprocesamiento para tratamiento de ruido de clase en problemas de clasificación

Eliminan las observaciones identificadas como ruidosas o modifican la clase asignada

Métodos basados en distancia (vecindario) o clasificación (frecuentemente mediante *ensembles* y *cross validation*)

Sintaxis

Dataset

Fórmula describiendo la etiqueta con ruido y las clases que se utilizarán para calcular la probabilidad de ruido

Salida

Datos sin ruido

Vector de índices eliminados

Métodos

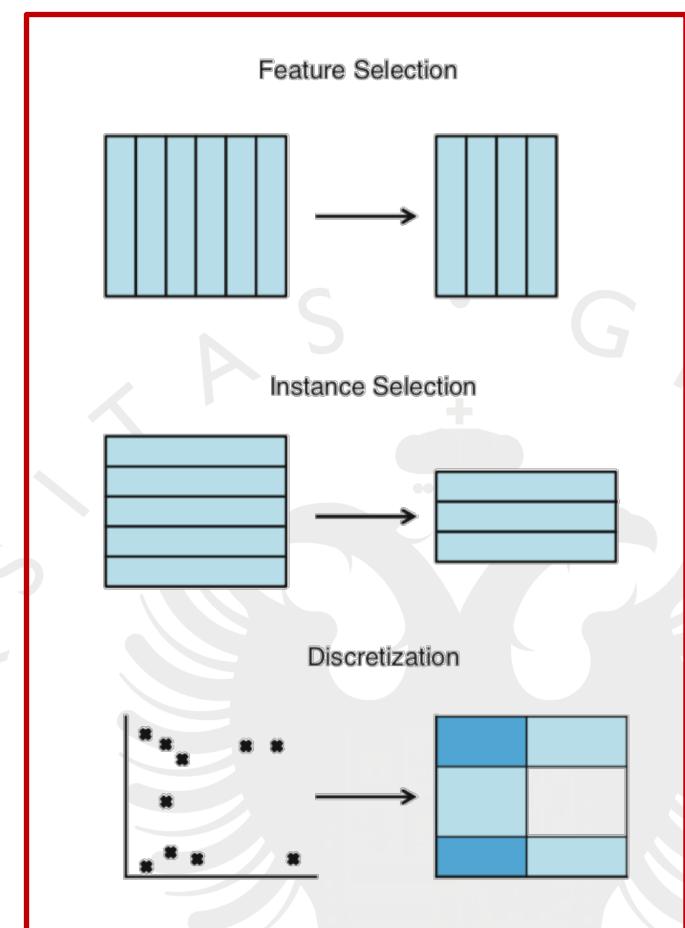
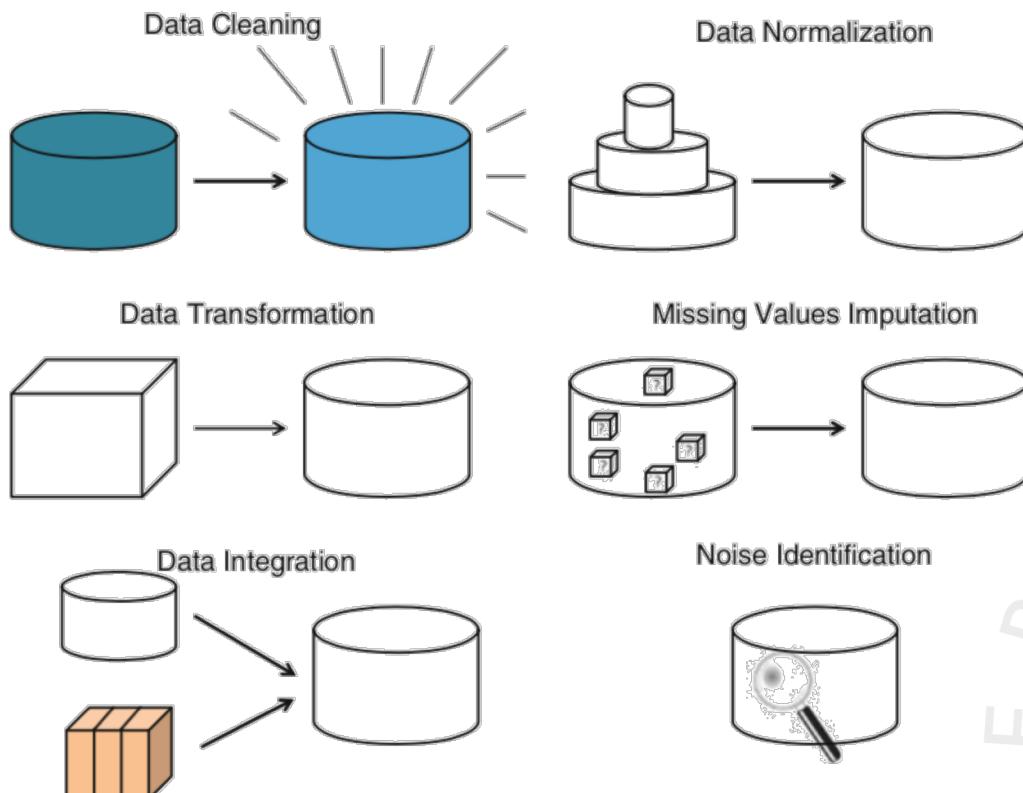
NoiseFiltersR

- AENN: All-k Edited Nearest Neighbors
- BBNR: Blame Based Noise Reduction
- C45ensembles: Classical Filters based on C4.5
- CNN: Condensed Nearest Neighbors
- CVCF: Cross-Validated Committees Filter
- DROP: Decremental Reduction Optimization Procedures
- dynamicCF: Dynamic Classification Filter
- edgeBoostFilter: Edge Boosting Filter
- EF: Ensemble Filter
- ENG: Editing with Neighbor Graphs
- ENN: Edited Nearest Neighbors
- EWF: Edge Weight Filter
- GE: Generalized edition
- HARF: High Agreement Random Forest
- hybridRepairFilter: Hybrid Repair-Remove

Filter

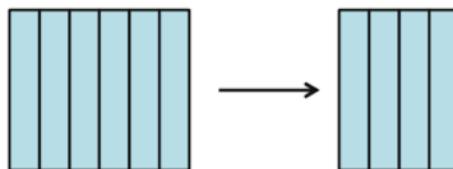
- INFCC: Iterative Noise Filter based on the Fusion of Classifiers
 - IPF: Iterative Partitioning Filter
 - ModeFilter: Mode Filter
 - ORBoostFilter: Outlier Removal Boosting Filter
 - PF: Partitioning Filter
 - PRISM: Preprocessing Instances that Should be Misclassified
 - RNN: Reduced Nearest Neighbors
 - saturationFilter: Saturation Filters
 - TomekLinks: TomekLinks
- summary: Summary method for class filter

Reducción de datos



S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 6. Data Reduction. Springer.

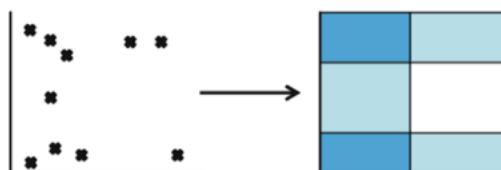
Feature Selection



Instance Selection



Discretization



Reducir
dimensionalidad >

Selección de características

Eliminar muestras
redundantes y conflictivas >

Selección de ejemplos

Simplificar dominio de una
variable >

Discretización

S. García, J. Luengo, F. Herrera (2015) Data Preprocessing in Data Mining. Ch. 7. Feature selection.
Springer.

Selección de características

Subconjunto de variables del problema que optimiza la posibilidad de crear un modelo de predicción correcto

Menos datos > Algoritmos más rápidos

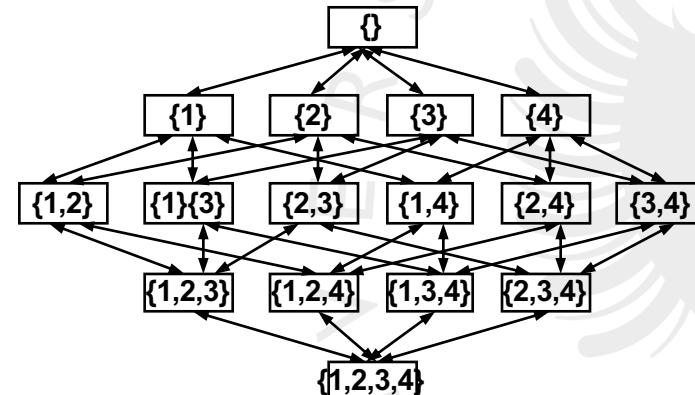
Mayor precisión > Más generalización

The curse of dimensionality

Resultados más simples > Más interpretables

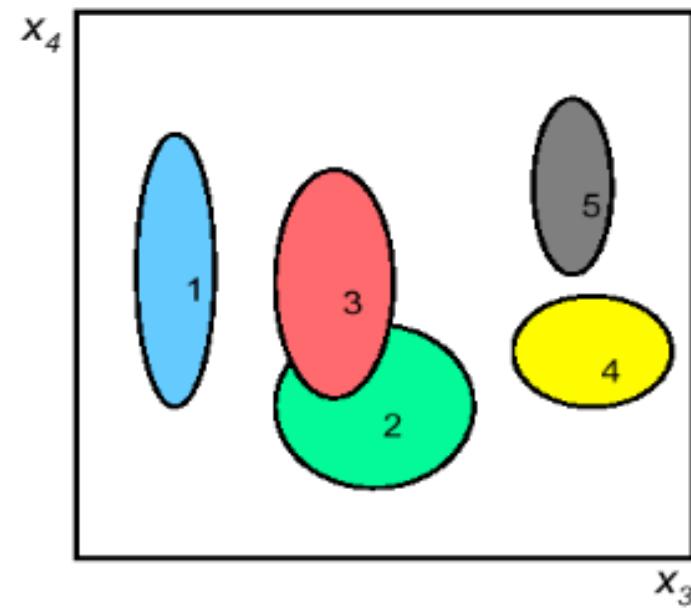
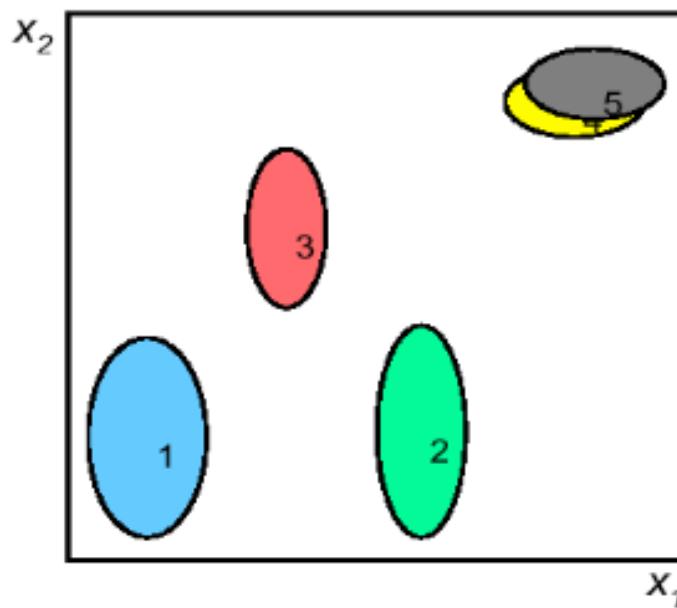
Puede verse como un problema de búsqueda

- 1) Selección
- 2) Evaluación
 - Filtro
 - Wrapper



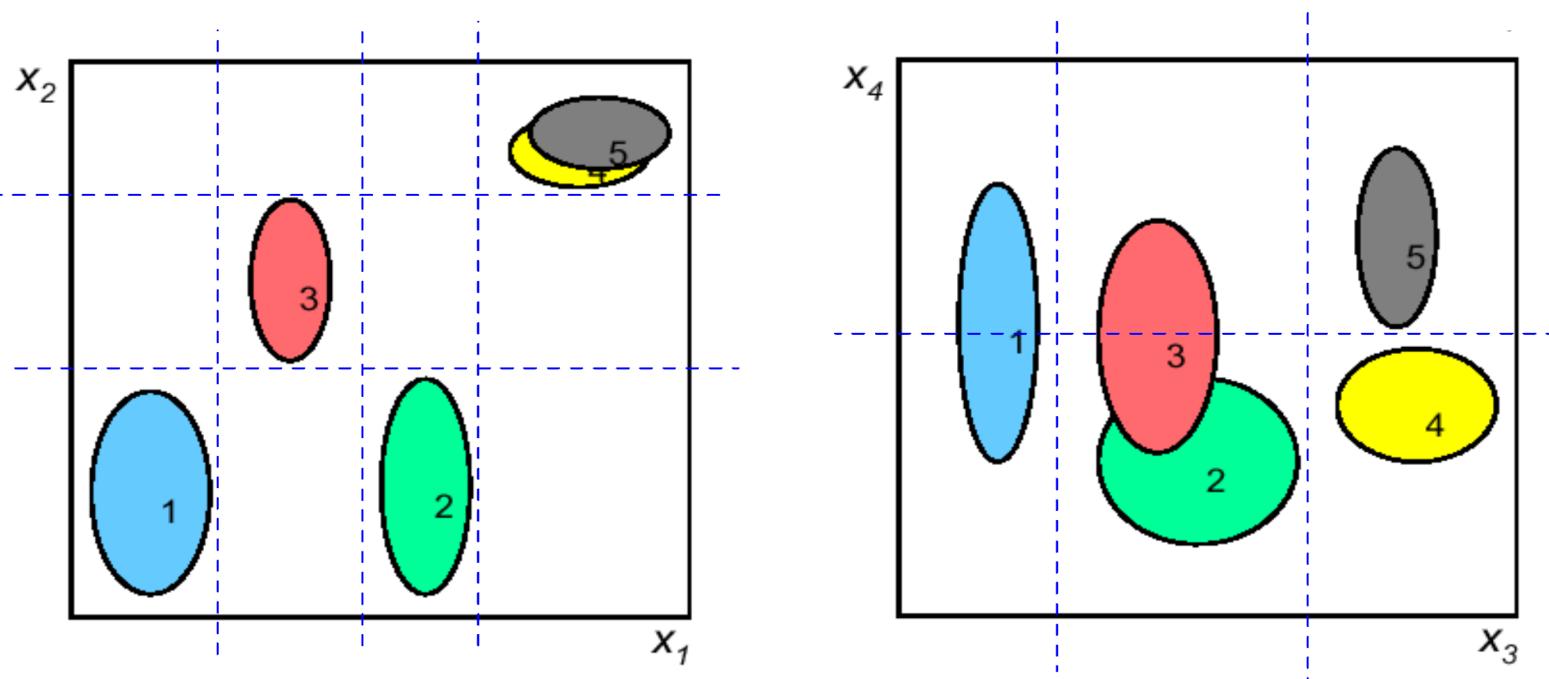
Selección de características

¿Qué subconjunto de variables seleccionamos?



Selección de características

¿Qué subconjunto de variables seleccionamos?



x_1 es mejor que **x_2**

x_3 es mejor que **x_4**

x_1 es mejor que **x_3**

... pero **(x_1, x_4)** separan todas las clases

Selección de características

Clasificación

Por forma de evaluación en la búsqueda

- **Filtro** (*filter*) La función objetivo evalúa los subconjuntos basándose en la información que contienen:
 - Distancias
 - Correlaciones
 - Teoría de la información
 - Consistencia y dependencia
- **Envolvente** (*wrapper*) La función objetivo consiste en aplicar la técnica de aprendizaje que se utilizará finalmente sobre la proyección de los datos al conjunto de variables candidato. El valor devuelto suele ser el porcentaje de acierto del clasificador construido

Por disponibilidad del objetivo de clasificación

Por amplitud de la búsqueda

Por la salida del algoritmo

Selección de características

Selección hacia adelante

Comienza con un conjunto vacío, al que va añadiendo secuencialmente el atributo que maximiza $U(S \cup x_i)$

Entrada

X: conjunto de atributos
U: criterio de evaluación

Proceso

```
S = Ø
do
     $x+ = \arg \max_{x \in X-S} U(S \cup x)$ 
    S = S ∪ {x+}
```

```
while not stop
```

Resultado

Conjunto de atributos

Selección de características

Selección hacia atrás

Comienza con el conjunto de todos las variables, del que se va eliminando secuencialmente el atributo que miniza $U(S - xi)$

Entrada

X: conjunto de atributos
U: criterio de evaluación

Proceso

```
S = X
do
     $x^- = \arg\max_{x \in S} U(S - x)$ 
    S = S - {x^-}
while not stop
```

Resultado

Conjunto de atributos

Selección de características

Enfoques mixtos

Selección l-más r-menos

Repite / calculos de $x+$ y r cálculos de $x-$

Selección bidireccional

Ejecución paralela de adelante y atrás

Selección flotante

Selección l-más r-menos que no fija a priori / y r

Construcción de árboles de decisión

Identifican los predictores en orden de prioridad

> Todos estos algoritmos son lineales

Añaden o eliminan atributos de uno en uno, pueden caer en óptimos locales

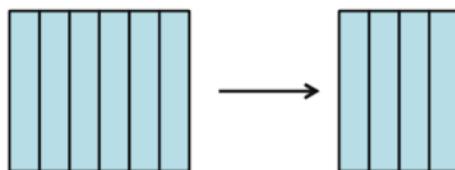
+ Algoritmos exponenciales

El número de subconjuntos evaluados aumenta exponencialmente

+ Algoritmos estocásticos

Añaden aleatoriedad para escapar de óptimos locales

Feature Selection



Reducir
dimensionalidad >

Selección de
características

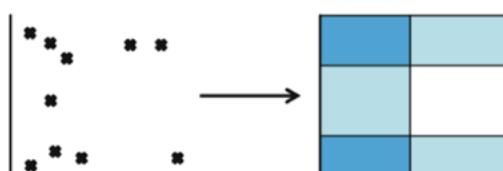
Instance Selection



Eliminar muestras
redundantes y conflictivas >

Selección de ejemplos

Discretization



Simplificar dominio de una
variable >

Discretización

S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 8. Instance Selection. Springer.

Selección de ejemplos

Menos datos > Algoritmos más rápidos

Mayor precisión > Más generalización

Resultados más simples > Más interpretables

Soluciones

Muestreo

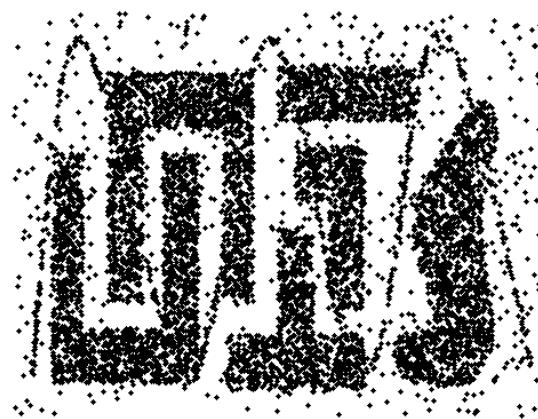
Selección de prototipos

Aprendizaje activo

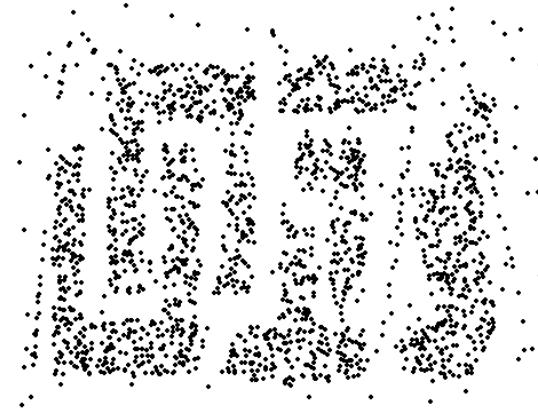
S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch. 8. Instance Selection.
Springer.

T. Reinartz (2002) A unifying view on instance selection. *Data Mining and Knowledge Discovery* 6, 191-210.

Selección de ejemplos



8000 puntos

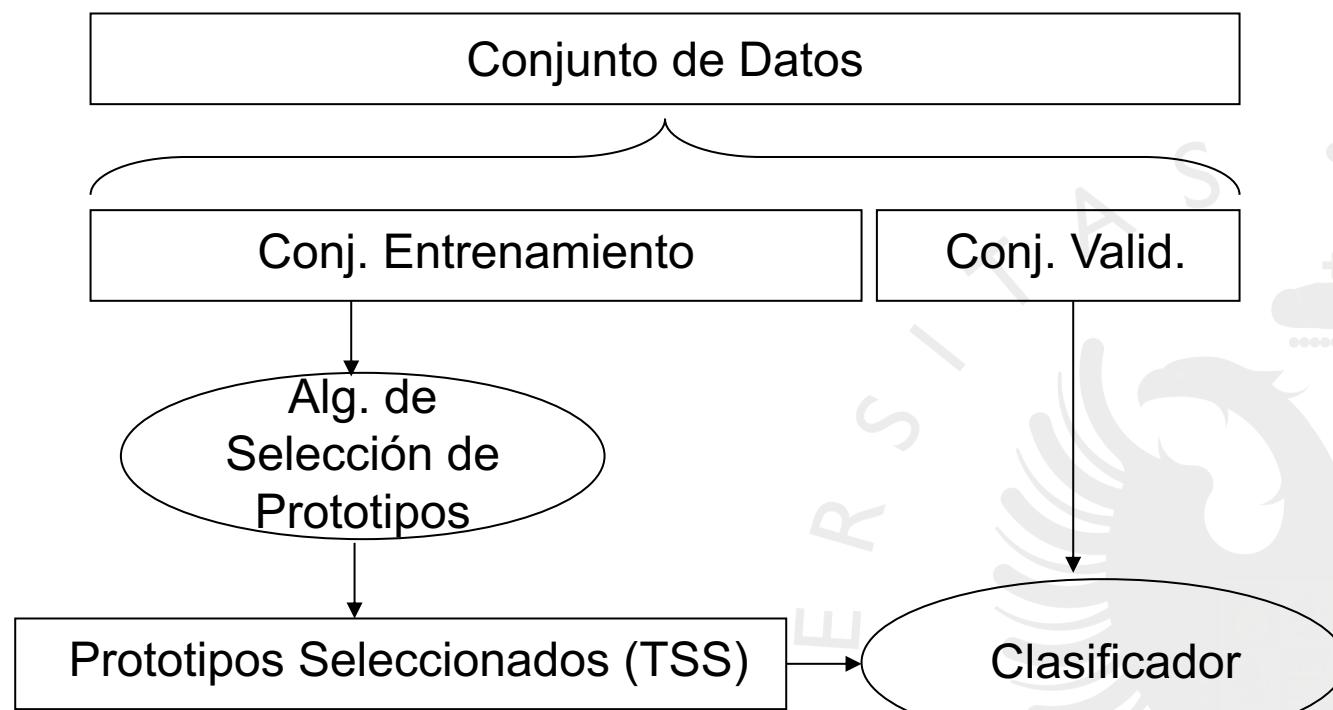


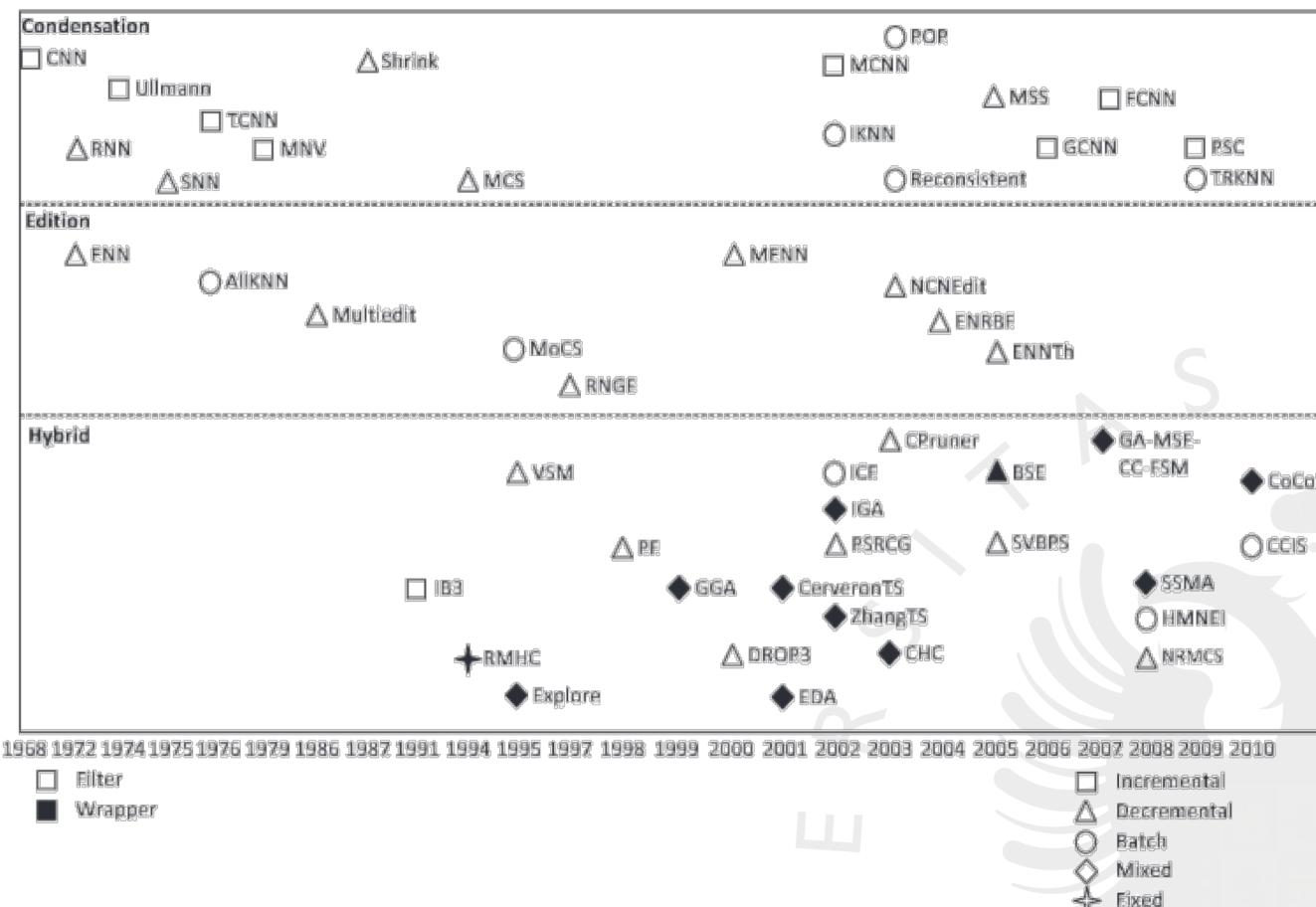
2000 puntos



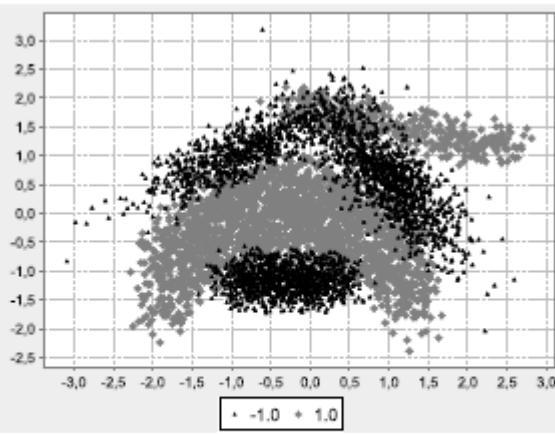
500 puntos

Selección de ejemplos

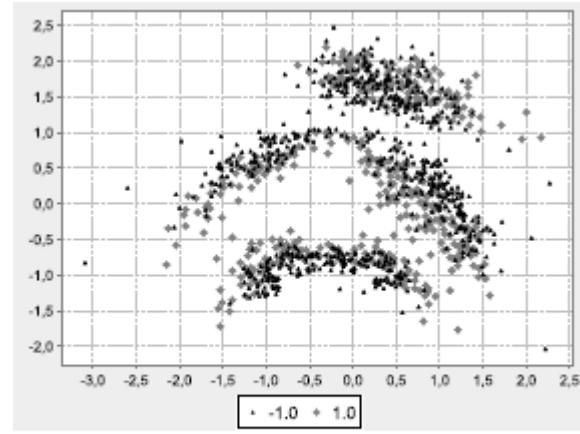




S. García, J. Luengo, F. Herrera (2015) Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 417-435.

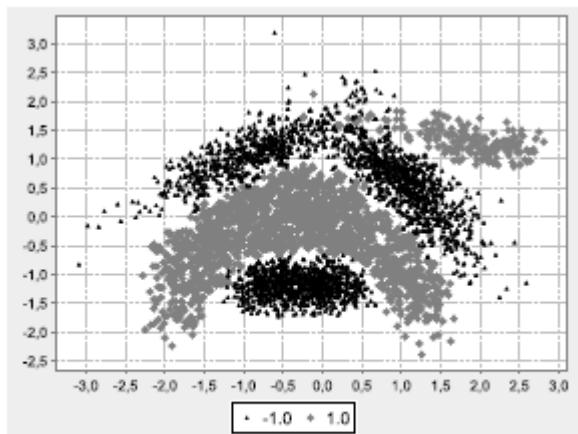


(a) Banana
(0.8751, 0.7476)

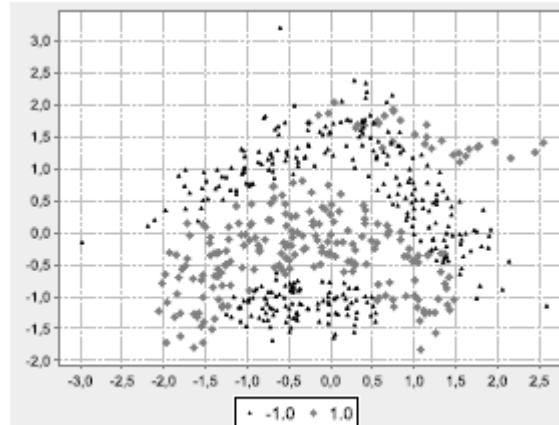


Original

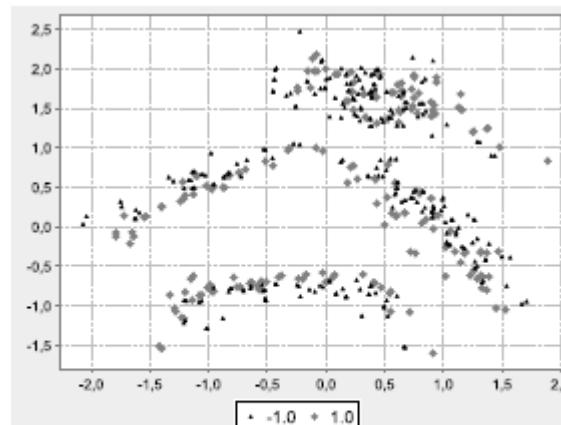
(b) CNN (0.7729, 0.8664, 0.7304)



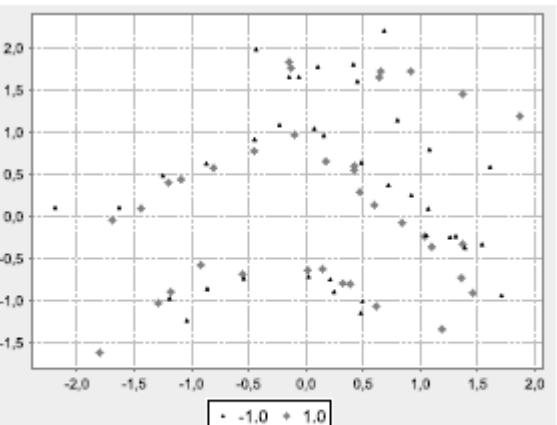
(h) AllKNN (0.1758, 0.8934, 0.7831)



(k) RMHC (0.9000, 0.8972, 0.7915)
(e)
(0.9151, 0.8696, 0.7356)



DROP3



(l) SSMA (0.9879, 0.8964, 0.7900)

Selección de ejemplos

Conjuntos de datos no balanceados

Problemas con presencia de clases desigual

Diagnóstico médico

E-commerce

Ciberseguridad

Es sencillo obtener un clasificador con un alto porcentaje de clasificación correcta, pero no son útiles

Aproximaciones

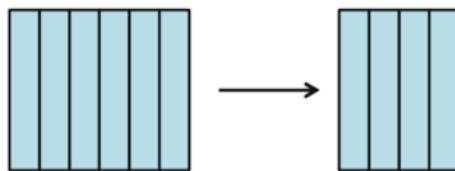
Reducción de datos de las clases mayoritarias

Sobremuestreo de la clases minoritarias

Generación de instancias artificiales

Hibridación entre selección de instancias y características

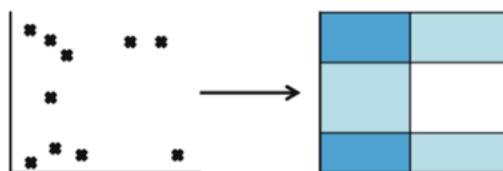
Feature Selection



Instance Selection



Discretization



Reducir
dimensionalidad >

Selección de
características

Eliminar muestras
redundantes y conflictivas >

Selección de ejemplos

Simplificar dominio de una
variable >

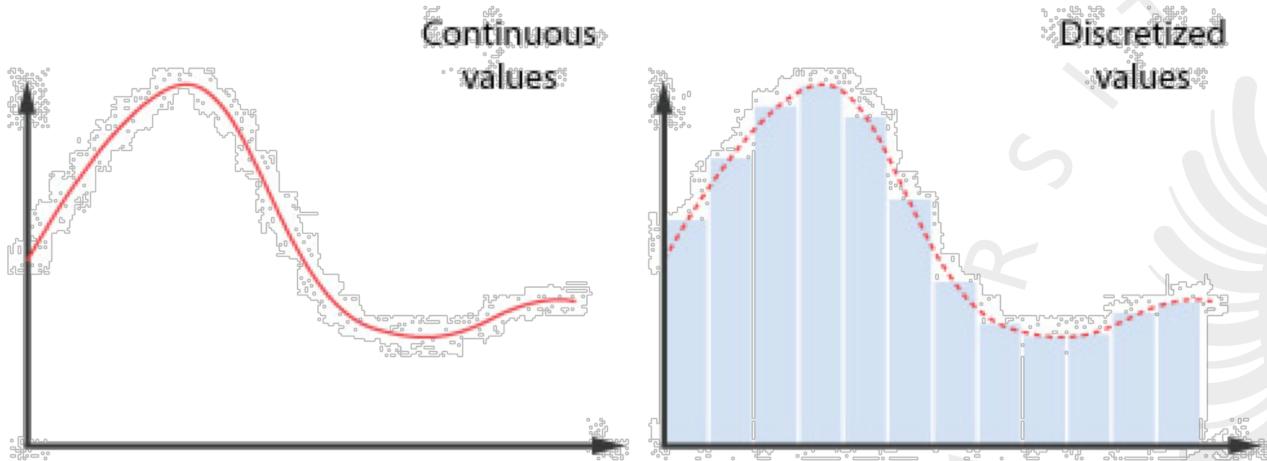
Discretización

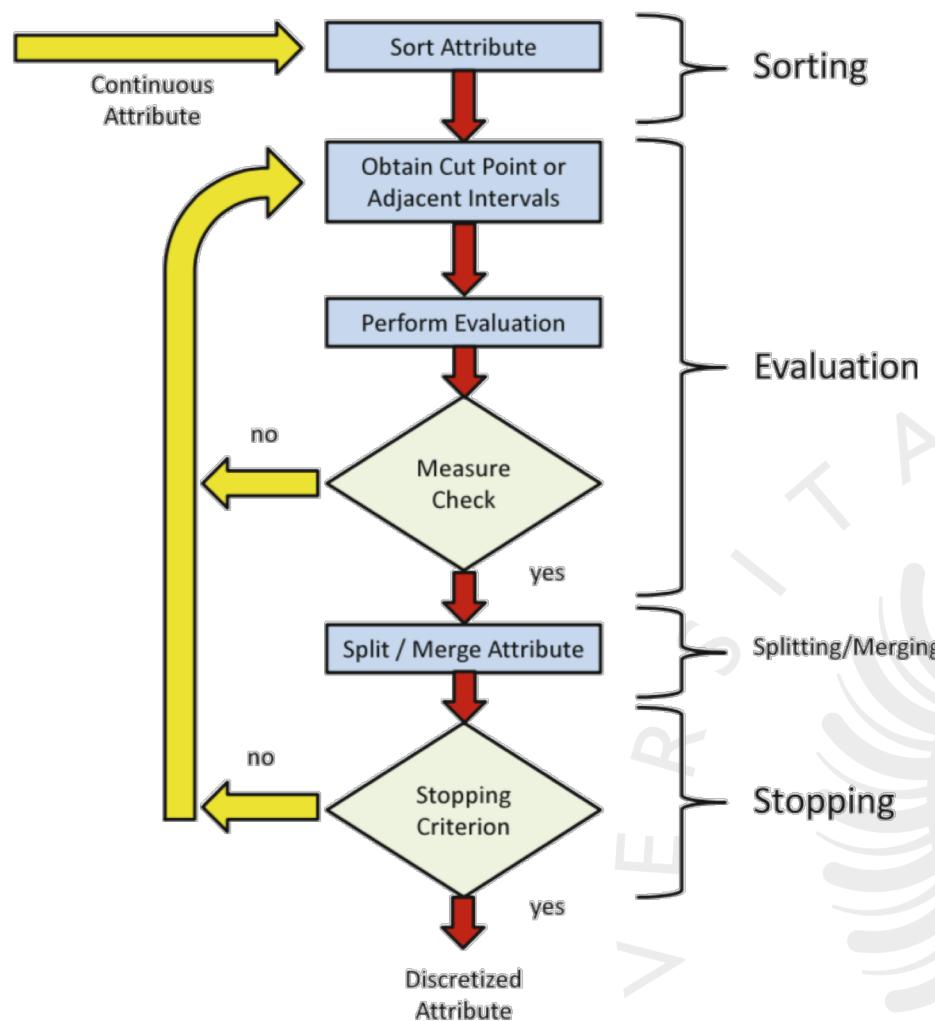
S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch.9. Discretization. Springer.

Discretización

Los valores discretos son más fáciles de manejar en aprendizaje automático

Transformar valores ordenados (numéricos) en valores nominales (categorías o intervalos)





S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Ch.9. Discretization. Springer.

Manual

Dirigido por el experto y el analista de datos

Algoritmos no supervisados [sin información de clase]

Intervalos de igual amplitud

Intervalos de igual frecuencia

Clustering

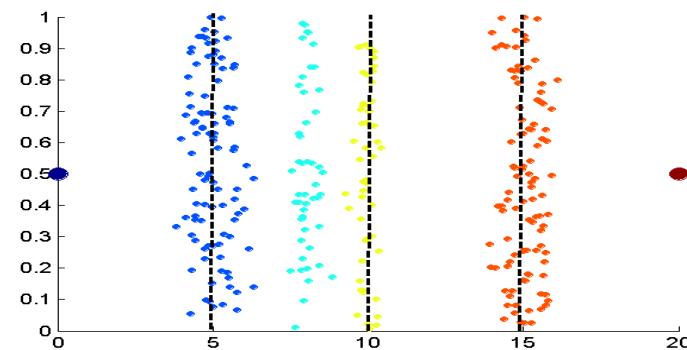
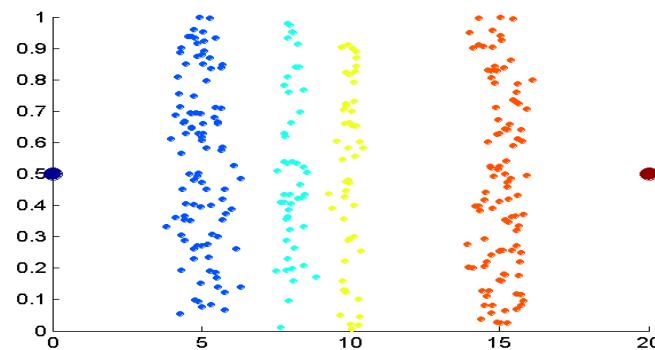
Algoritmos supervisados [con información de clase]

Entropía

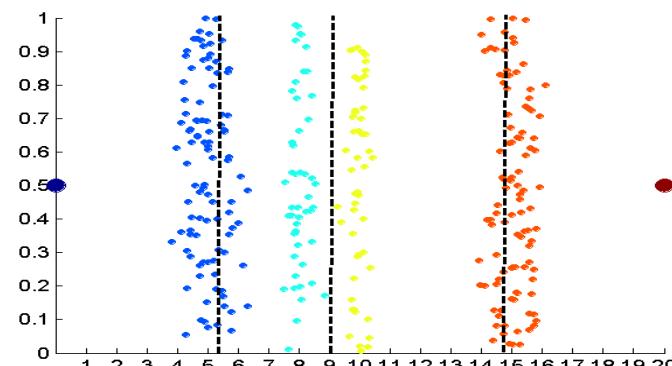
Chi-square

...

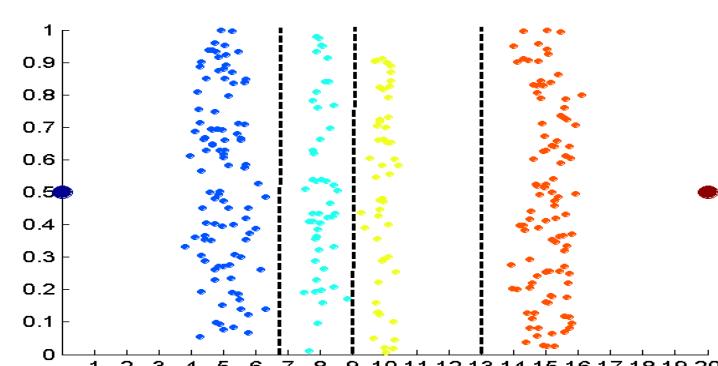
Algoritmos no supervisados



Igual anchura de intervalo

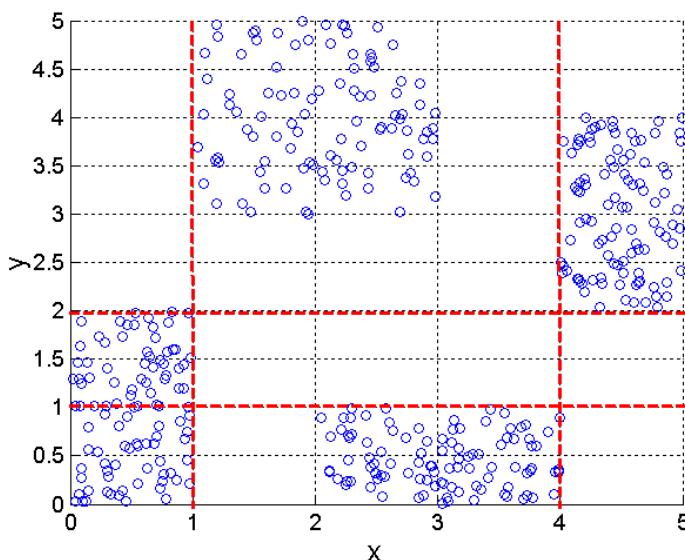


Igual frecuencia

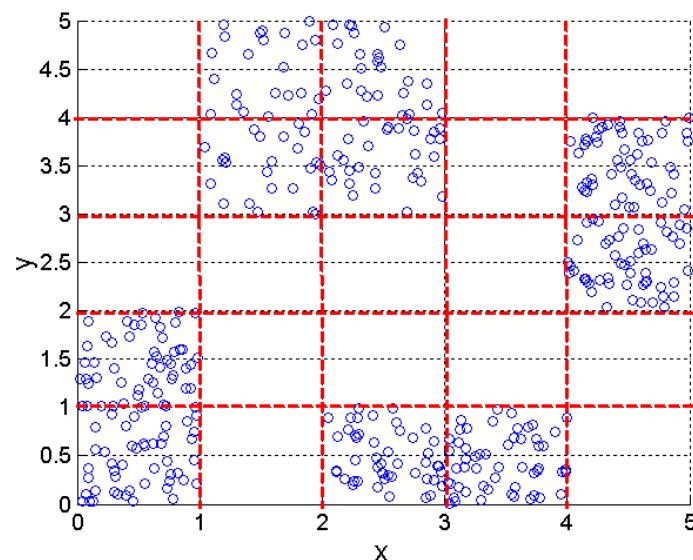


K-medias

Algoritmos supervisados

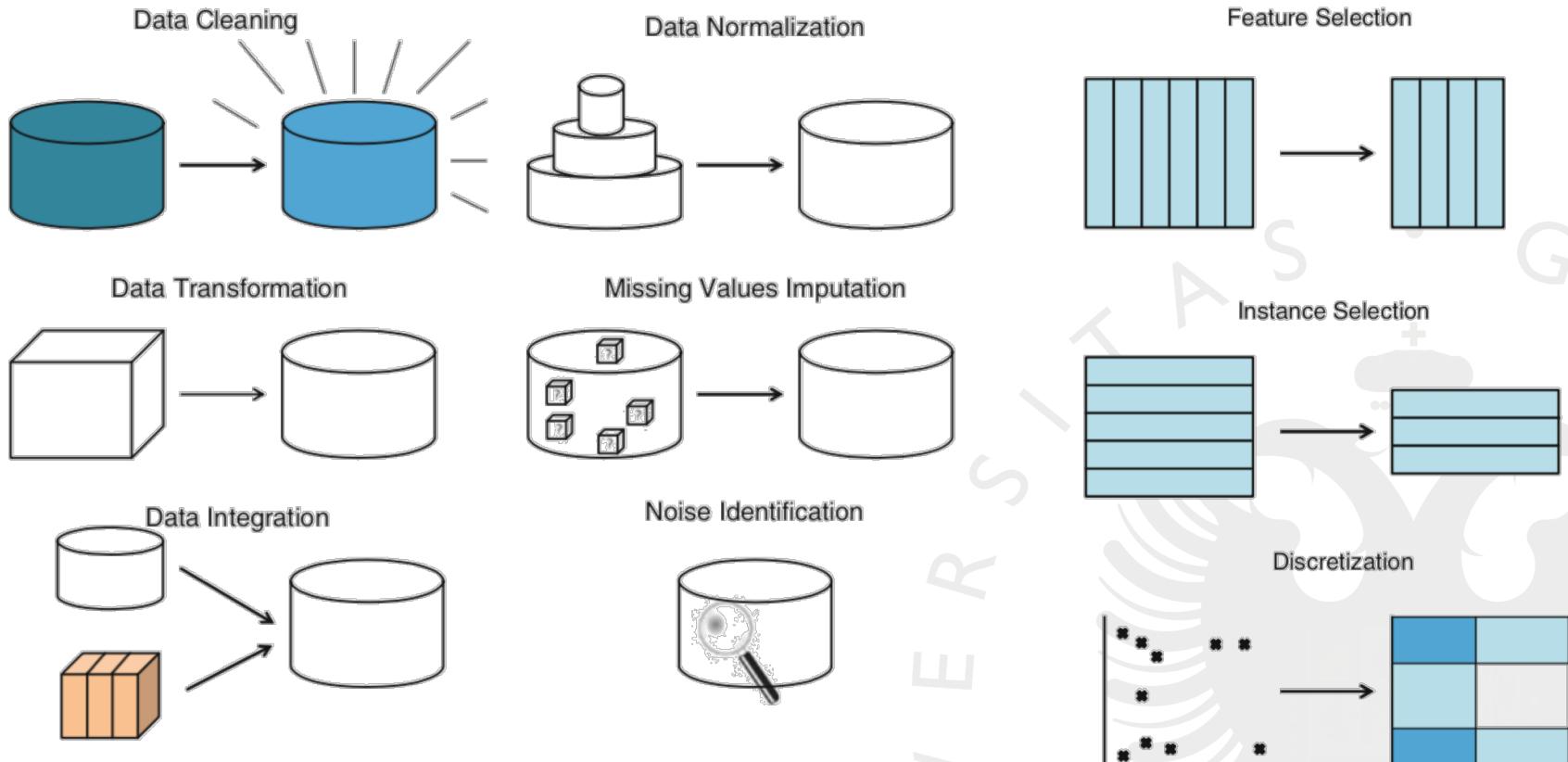


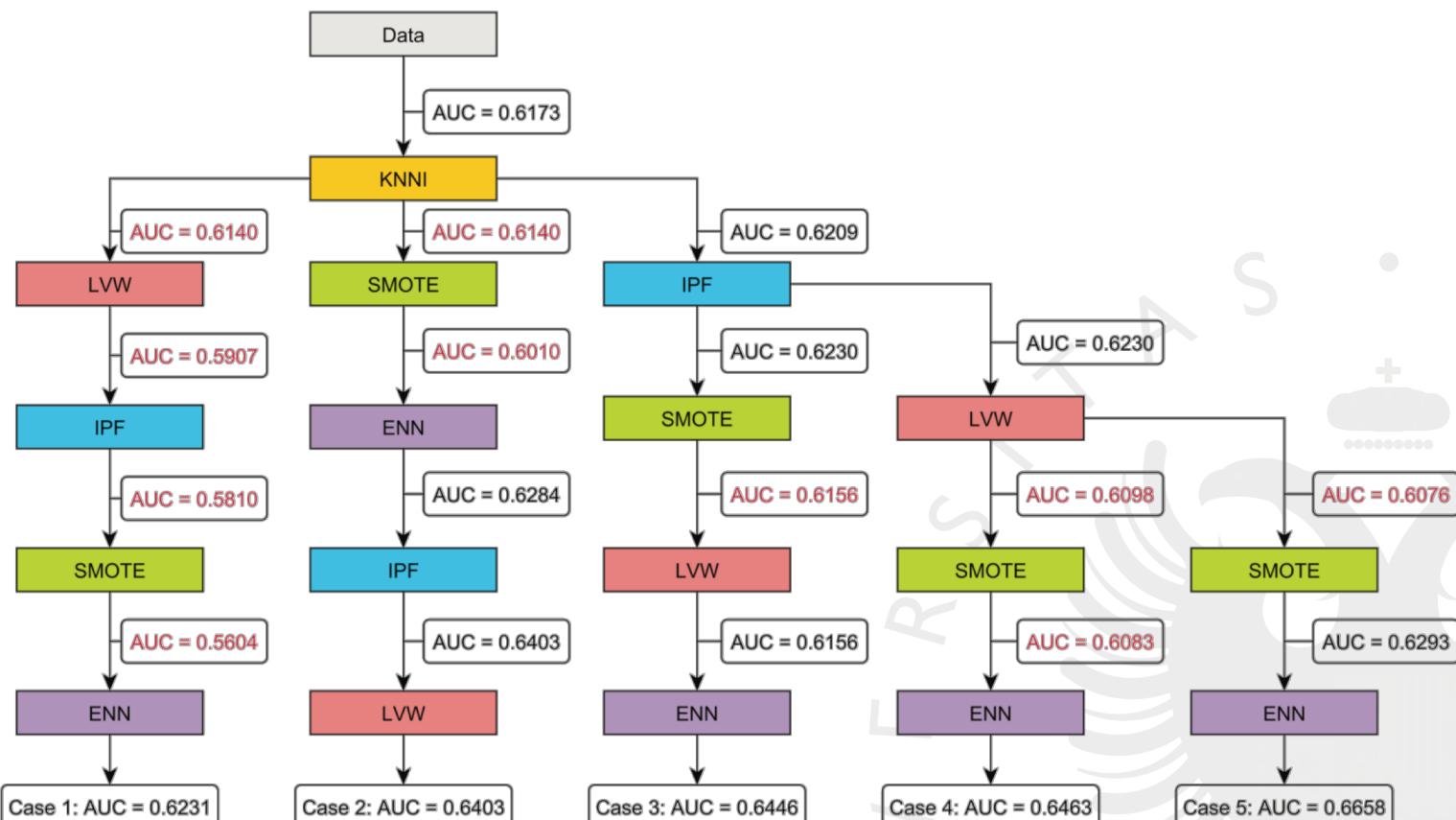
3 intervalos



5 intervalos

Pre-procesamiento de datos





S. García, J. Luengo, F. Herrera (2016) Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems* 98, 1–29.