

# TEMA 5. SELECCIÓN DE ATRIBUTOS

---

# Contenidos

---

- I. Introducción
- II. Métodos
- III. Selección de atributos en Orange:
  - a. Rank
  - b. PCA
- IV. Actividad

# I. Introducción

---

- Seleccionar el subconjunto más reducido de atributos tal que no se afecte negativamente al aprendizaje.
- Las técnicas de selección de atributos consisten en reducir el número de atributos de un conjunto de datos con dos objetivos:
  - Reducir el **coste computacional** del algoritmo de aprendizaje
  - Obtener modelos **más precisos** eliminando atributos irrelevantes que pudieran sesgar la búsqueda hacia modelos menos adecuados.

# I. Introducción

---

Los atributos se pueden caracterizar como:

- **Relevantes.** Son los atributos que tienen influencia en la clase y su papel no puede ser asumido por el resto de atributos.
- **Irrelevantes.** Son aquellos que no tienen influencia. Se asemejan a atributos cuyos valores hubieran sido generados aleatoriamente para cada ejemplo.
- **Redundantes.** Existe redundancia si un atributo puede tomar el papel de otro. Estadísticamente hablando se puede decir que están correlacionados.

# I. Introducción

---

¿Por qué usar selección de atributos?

- Elimina datos ruidosos, redundantes o irrelevantes
- Mejora del rendimiento predictivo
- Mejor visualización y comprensión de los datos
- Reducción del tiempo de entrenamiento/predicción
- Reducción de las necesidades de almacenamiento

# II. Métodos

---

Métodos generales para seleccionar características:

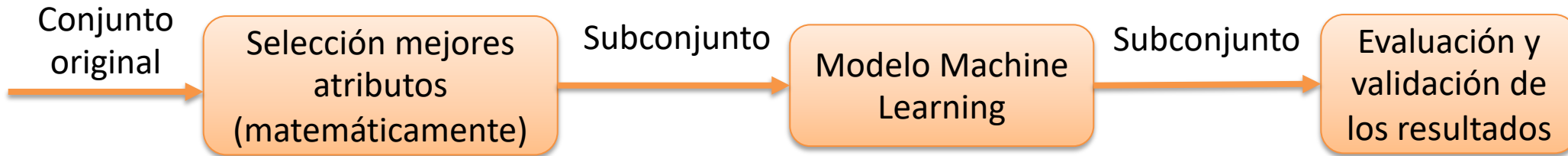
- **Métodos de filtro:** se evalúan los atributos (individual o conjuntamente) mediante criterios matemáticos. No usan modelos de machine learning. ANOVA, Chi-cuadrado, Pearson, ....
- **Métodos basados en modelo o métodos de envoltante (wrapper):** la bondad se evalúa respecto a la calidad de un modelo de machine learning entrenado a partir de los datos reducidos (utilizando algún método de validación interna). Hay que controlar el sobreajuste (overfitting).

Suelen ser más lentos y costosos que los anteriores.

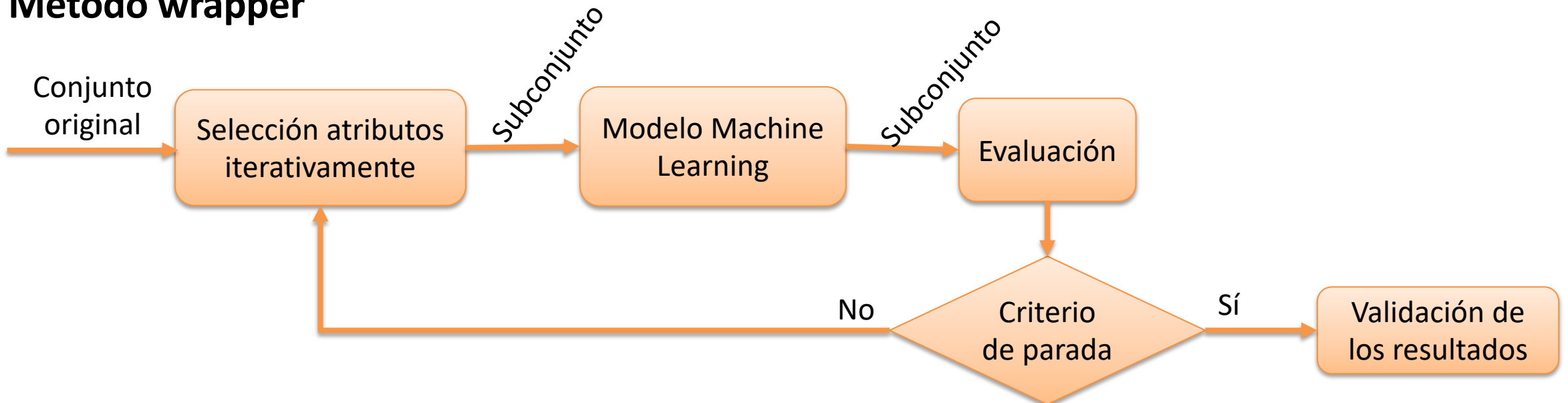
Suelen dar muy buenos resultados gracias a su naturaleza exhaustiva.

# II. Métodos

## Método de filtro



## Método wrapper



# III. Selección de atributos en Orange

Nodo **Rank** permite evaluar **individualmente** los atributos de un conjunto de datos supervisados con respecto a la clase.



Rank

- No se realiza búsqueda de subconjuntos (método de filtro).
- Se basa en métodos de evaluación univariantes.
- Proporciona una tabla de atributos con las puntuaciones de cada atributo según la/s métrica/s escogida/s.
- Soporta selección manual de atributos.

	#	Info.gain	Gain ratio	Gini	$\chi^2$	ReliefF	FCBF
outlook	3	0.247	0.156	0.116	2.028	0.064	0.244
humidity	2	0.152	0.152	0.092	1.400	0.048	0.186
wind	2	0.048	0.049	0.031	0.400	0.032	0.053
temp	3	0.029	0.019	0.019	0.022	-0.060	0.000



# III. Selección de atributos en Orange

---

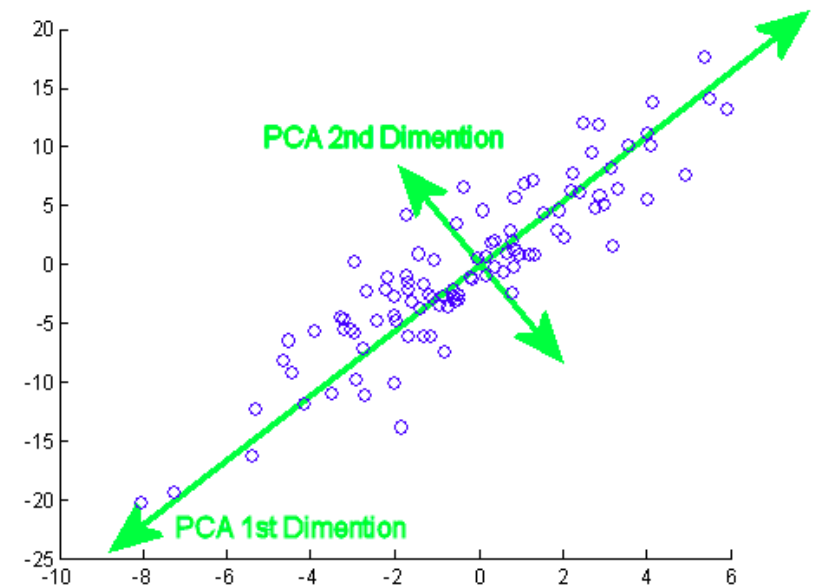
Nodo **PCA** (Principal Component Analysis):

- Permite realizar un análisis de componentes principales y reducir la dimensionalidad de los datos.
- Genera nuevos atributos como **combinación lineal** de los atributos (todos o alguna selección).
- Los nuevos atributos se ordenan por importancia (varianza explicada).
- Objetivo: Cubrir al menos una cantidad de la varianza total de los datos usando el menor número de componentes PCA posibles.
- No es necesario disponer de una clase (técnica no supervisada). No se puede asegurar que los atributos generados discriminen mejor las clases.

# III. Selección de atributos en Orange

Nodo **PCA** (Principal Component Analysis):

- Objetivos:
  - Identificar patrones ocultos en los datos.
  - Reducir la dimensionalidad de los datos quitando el ruido y la redundancia.
  - Identificar las variables correlacionadas.



# III. Selección de atributos en Orange

Nodo **PCA** (Principal Component Analysis):

- Cada componente PCA es una combinación lineal de los atributos.
- Línea roja: varianza cubierta por componente
- Línea verde: varianza acumulada cubierta por componentes

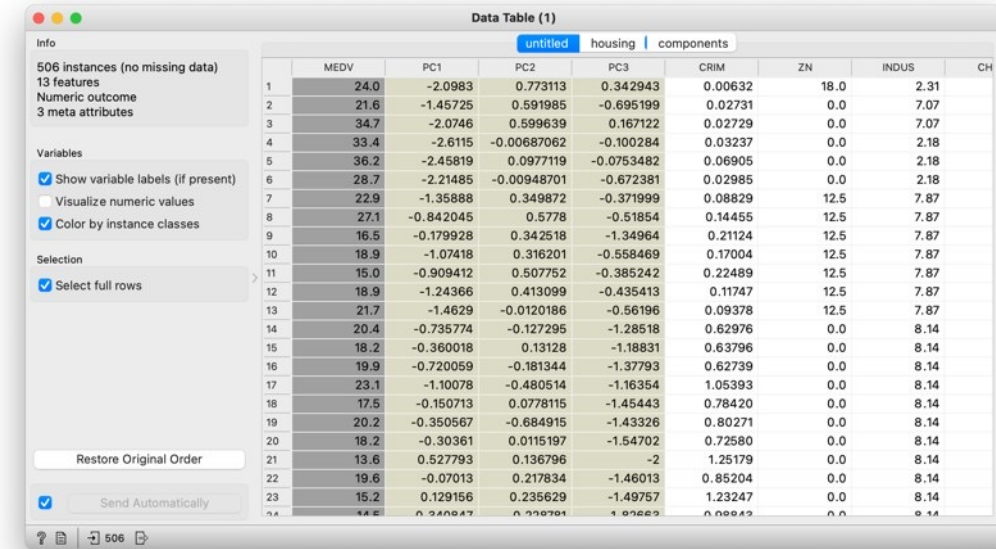


Gráfico de cobertura de varianza

# III. Selección de atributos en Orange

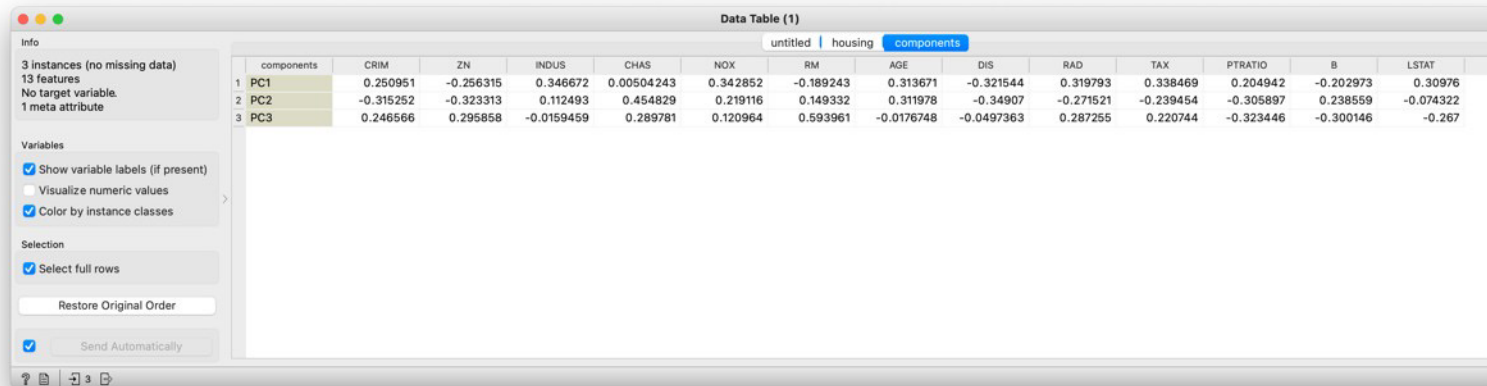
Nodo **PCA** (Principal Component Analysis) devuelve:

- Datos transformados: Pesos para las instancias en un nuevo sistema de coordenadas.
- Componentes principales: Descriptores del sistema, son los pesos de las componentes principales.



	MEDV	PC1	PC2	PC3	CRIM	ZN	INDUS	CH
1	24.0	-2.0983	0.773113	0.342943	0.00632	18.0	2.31	
2	21.6	-1.45725	0.591985	-0.695199	0.02731	0.0	7.07	
3	34.7	-2.0746	0.599639	0.167122	0.02729	0.0	7.07	
4	33.4	-2.6115	-0.00687062	-0.100284	0.03237	0.0	2.18	
5	36.2	-2.45819	0.0977119	-0.0753482	0.06905	0.0	2.18	
6	28.7	-2.21485	-0.00948701	-0.672381	0.02985	0.0	2.18	
7	22.9	-1.35888	0.349872	-0.371999	0.08829	12.5	7.87	
8	27.1	-0.842045	0.5778	-0.51854	0.14455	12.5	7.87	
9	16.5	-0.179928	0.342518	-1.34964	0.21124	12.5	7.87	
10	18.9	-1.07418	0.316201	-0.558469	0.17004	12.5	7.87	
11	15.0	-0.909412	0.507752	-0.385242	0.22489	12.5	7.87	
12	18.9	-1.24366	0.413099	-0.435413	0.11747	12.5	7.87	
13	21.7	-1.4629	-0.0120186	-0.56196	0.09378	12.5	7.87	
14	20.4	-0.735774	-0.127295	-1.28518	0.62976	0.0	8.14	
15	18.2	-0.360018	0.13128	-1.18831	0.63796	0.0	8.14	
16	19.9	-0.720059	-0.181344	-1.37793	0.62739	0.0	8.14	
17	23.1	-1.10078	-0.480514	-1.16354	1.05393	0.0	8.14	
18	17.5	-0.150713	0.0778115	-1.45443	0.78420	0.0	8.14	
19	20.2	-0.350567	-0.684915	-1.43326	0.80271	0.0	8.14	
20	18.2	-0.30361	0.0115197	-1.54702	0.72580	0.0	8.14	
21	13.6	0.527793	0.136796	-2	1.25179	0.0	8.14	
22	19.6	-0.07013	0.217834	-1.46013	0.85204	0.0	8.14	
23	15.2	0.129156	0.235629	-1.49757	1.23247	0.0	8.14	
24	14.5	0.240847	0.228791	-1.92669	0.08842	0.0	8.14	

Fuente: [https://www.cienciadedatos.net/documentos/35\\_principal\\_component\\_analysis](https://www.cienciadedatos.net/documentos/35_principal_component_analysis)



	components	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
1	PC1	0.250951	-0.256315	0.346672	0.00504243	0.342852	-0.189243	0.313671	-0.321544	0.319793	0.338469	0.204942	-0.202973	0.30976
2	PC2	-0.315252	-0.323313	0.112493	0.454829	0.219116	0.149332	0.311978	-0.34907	-0.271521	-0.239454	-0.305897	0.238559	-0.074322
3	PC3	0.246566	0.295858	-0.0159459	0.289781	0.120964	0.593961	-0.0176748	-0.0497363	0.287255	0.220744	-0.323446	-0.300146	-0.267

Datos reducidos (tomando sólo PC1, PC2 y PC3 como atributos)

Combinación lineal de cada componente PCA

# IV. Actividad: Selección de atributos

---

1. Cargue el dataset **MONK's 1**
2. Utilice el nodo **Distributions** para observar los atributos
3. Utilice el nodo **Rank** y elija los dos mejores atributos según:
  - a) Information Gain
  - b) Gini Decrease
4. ¿Cuáles son los dos mejores atributos? ¿Coinciden con lo que se observó con el nodo Distributions?
5. Realice un análisis de componentes principales usando el nodo **PCA**