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, ….

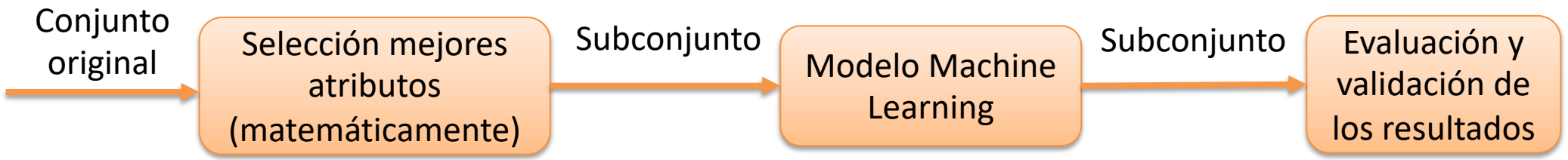
- **Métodos basados en modelo o métodos de envolvente (wrapper):** la bondad se evalúa respecto a la calidad de un modelo de machine learnig entrenado a partir de los datos reducidos (utilizando algún método de validación interna). Hay que controlar el sobreajuste (overfitting).

  Suelen ser más lentos y costosos que los anteriores.

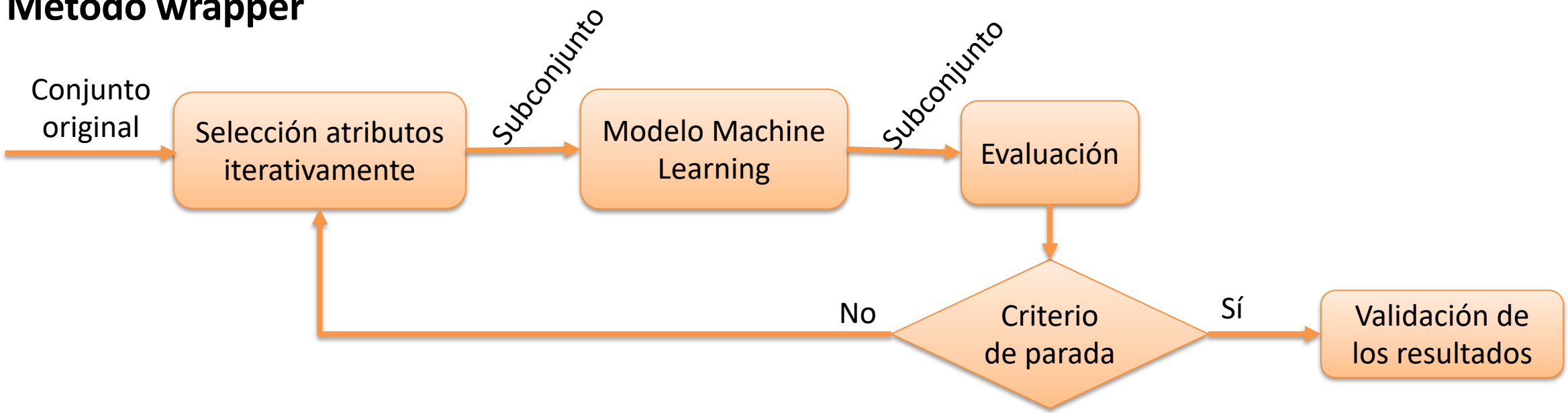
  Suelen dar muy buenos resultados gracias a su naturaleza exhaustiva.

II. Métodos

Método de filtro



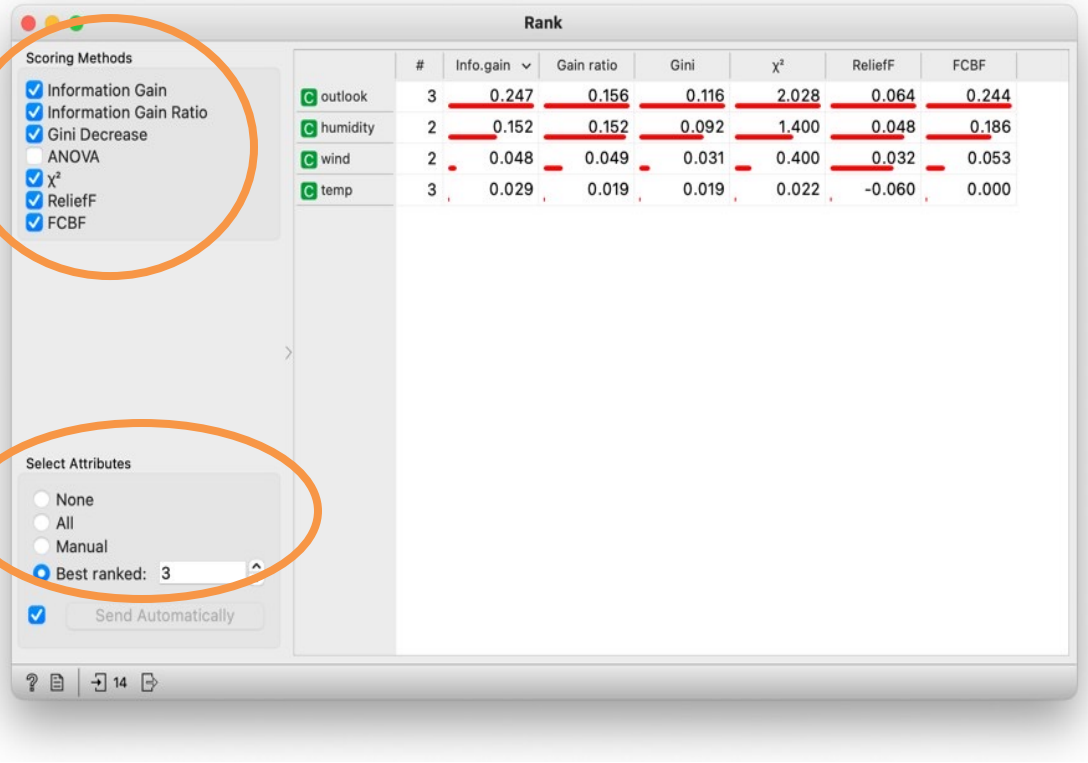
Método wrapper



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.

III. Selección de atributos en Orange

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

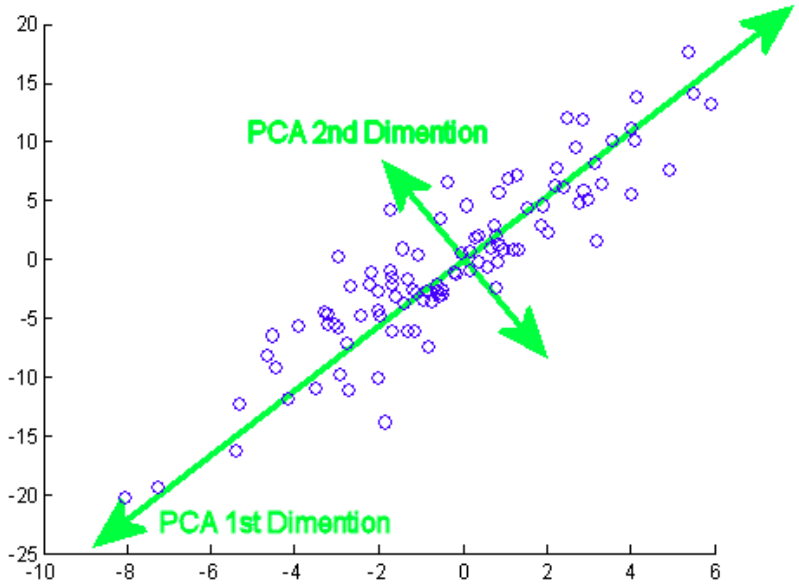
- Permite realizar un análisis de componentes principales y reducir la dimensionalidad de los datos.
- Genera **nuevos atributos** como **combinación lineal** de los atributos (todos o alguna selección).
- Los nuevos atributos se ordenan por importancia (varianza explicada).
- Objetivo: Cubrir al menos 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)**:**

PCA

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

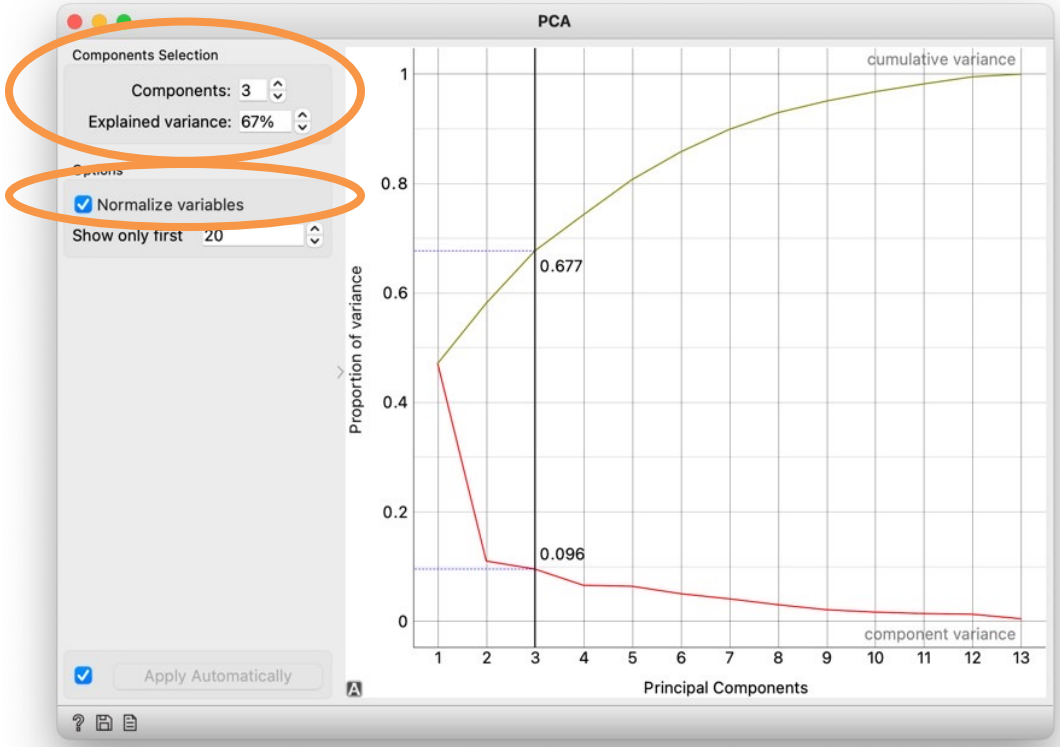


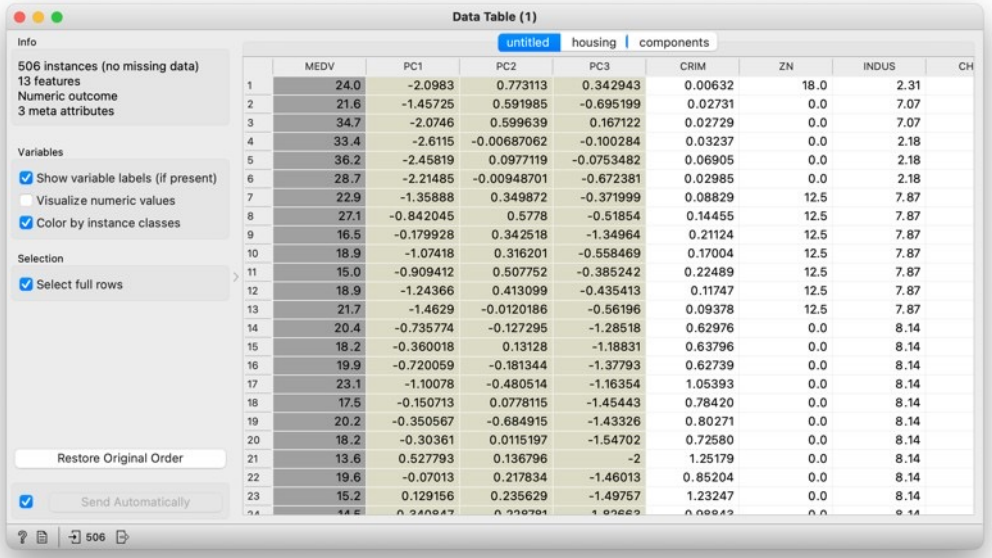
Gráfico de cobertura de varianza

III. Selección de atributos en Orange

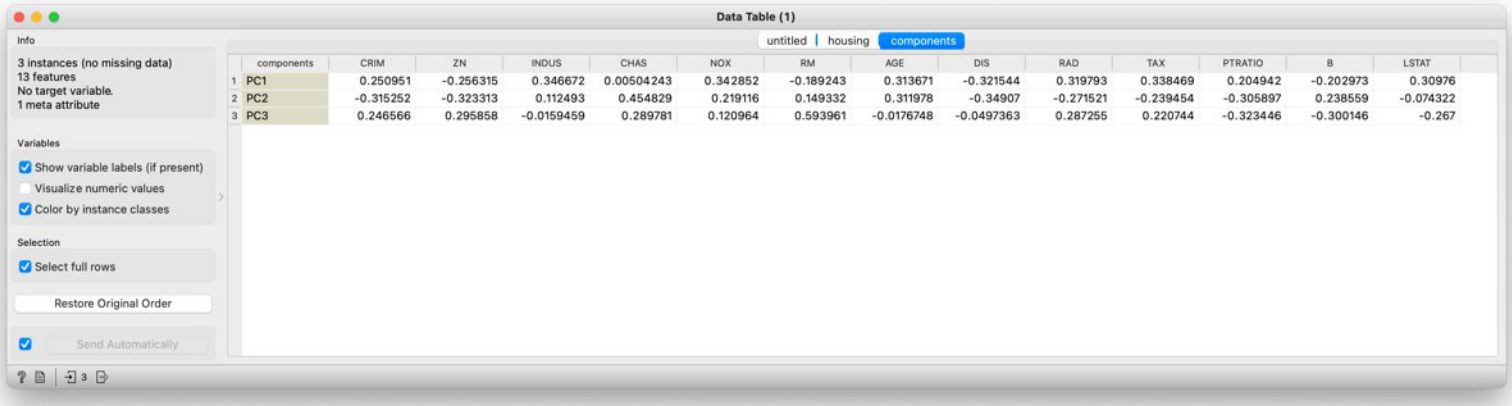
Nodo **PCA** (Principal Component Analysis) devuelve:

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

Fuente: https://www.cienciadedatos.net/documentos/35_principal_component_analysis



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**