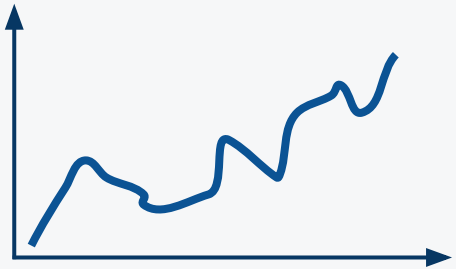
Detección de anomalías

PORTAFOLIOS | EQUIPO 1
Caso de estudio 2

Detección de anomalías

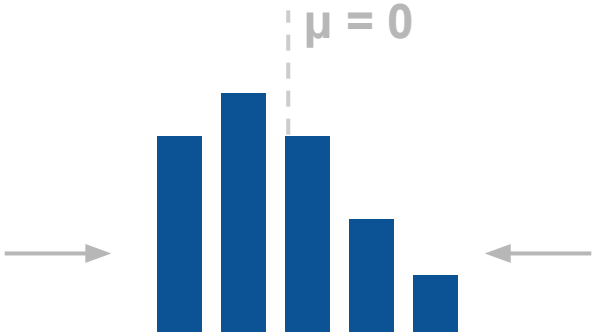
Forma **supervisada** VS Forma **no supervisada**



Detección de anomalías

2 Problemas

Distribución **sesgada**



Datos **normalizados**

Distancias

Distancias

¿Cómo detectar un outlier?

La detección se basa en la suposición de que los datos atípicos se encuentran separados de la mayoría de los datos del conjunto de datos.

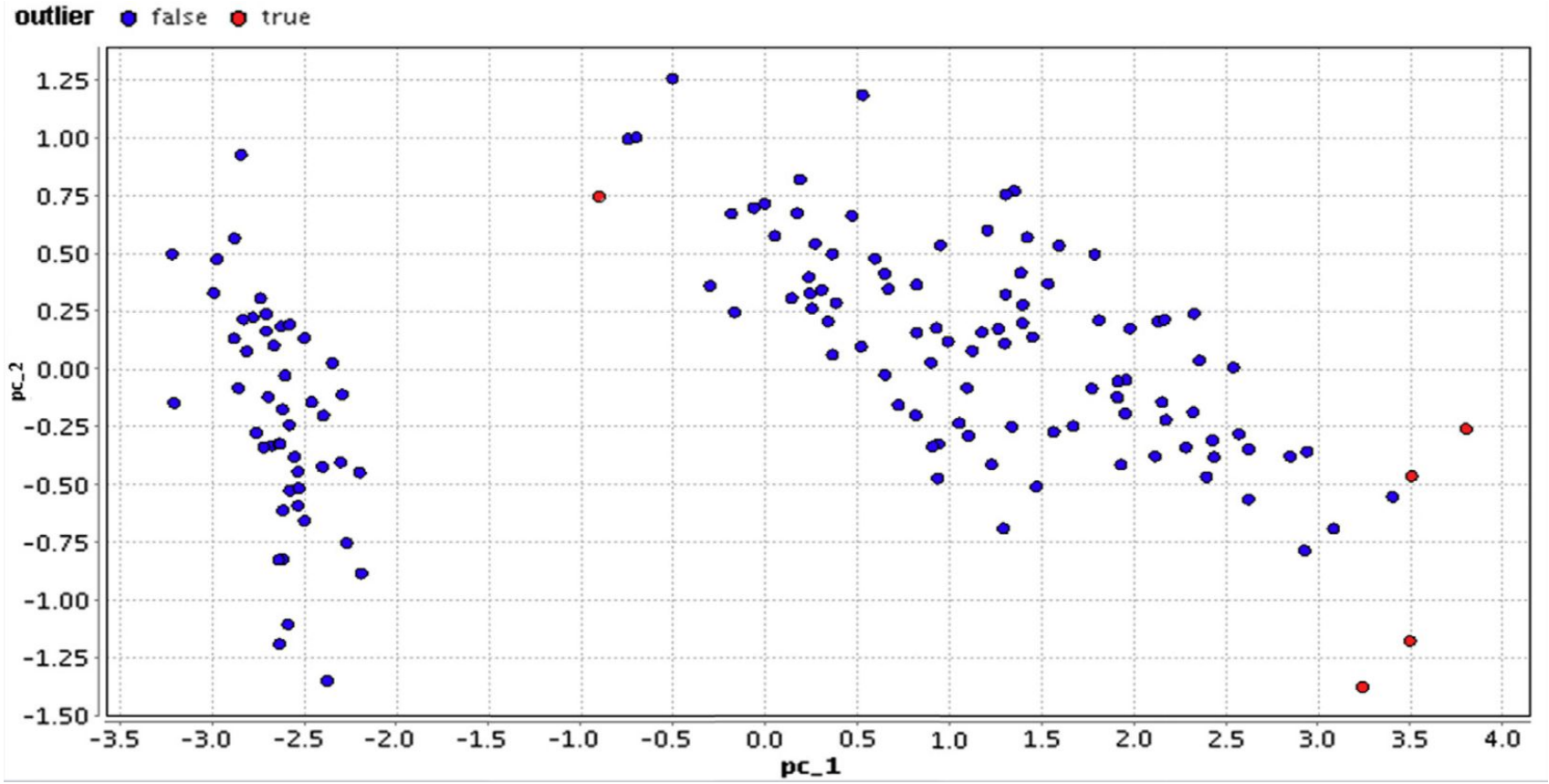
1. Se le asigna un puntaje a cada instancia, cuyo valor es la distancia hasta su k-vecino más cercano.

2. Se seleccionan las n instancias con mayor puntaje y se dice que son probables outliers.

Distancias - Parámetros

1. Número de **vecinos**
2. Número de **outliers**
3. Función de **distancia**

Distancias - Salida



Densidad

Densidad

¿Cómo detectar un outlier?

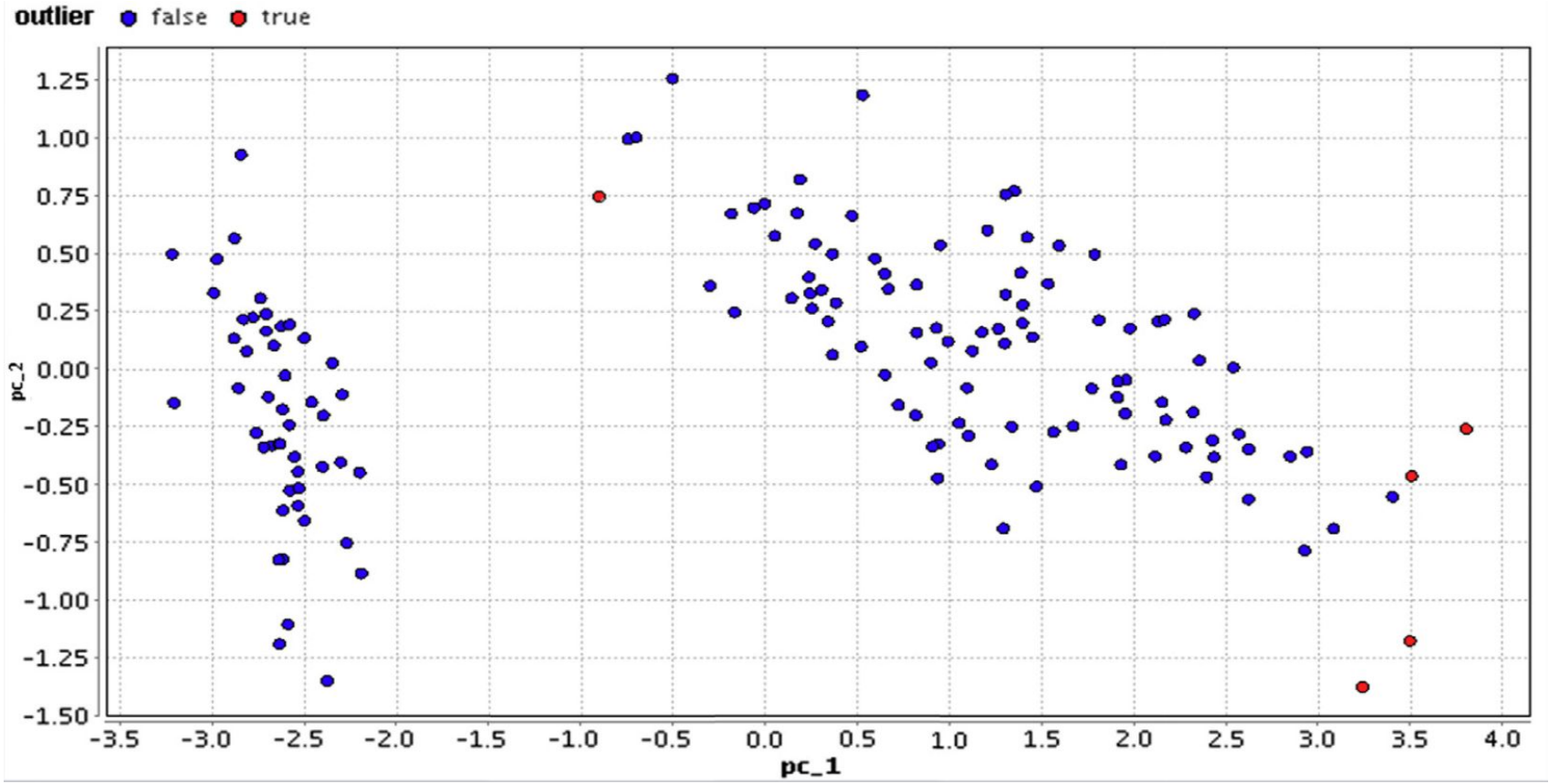
- Los valores atípicos ocurren con menos frecuencia que datos normales → En el espacio de los datos atípicos ocupan áreas de baja densidad y puntos de datos normal ocupan zonas de alta densidad.

- La densidad es un recuento de los puntos de datos en una unidad normalizada espacio y es inversamente proporcional a la distancia entre puntos de datos

Densidad - Parámetros

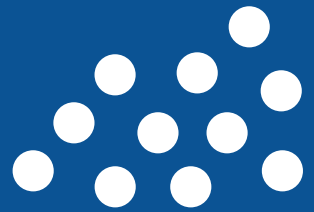
1. Distancia
2. Proporción
3. Función de **distancia**

Densidad - Salida



LOF

Local Outlier Factor



LOF - Descripción

¿Cómo detectar un outlier?

Densidad del punto **<** Densidad del vecindario

Densidad del punto

(**Promedio** de la **distancia a todos los puntos** del vecindario)$^{-1}$

Densidad del vecindario

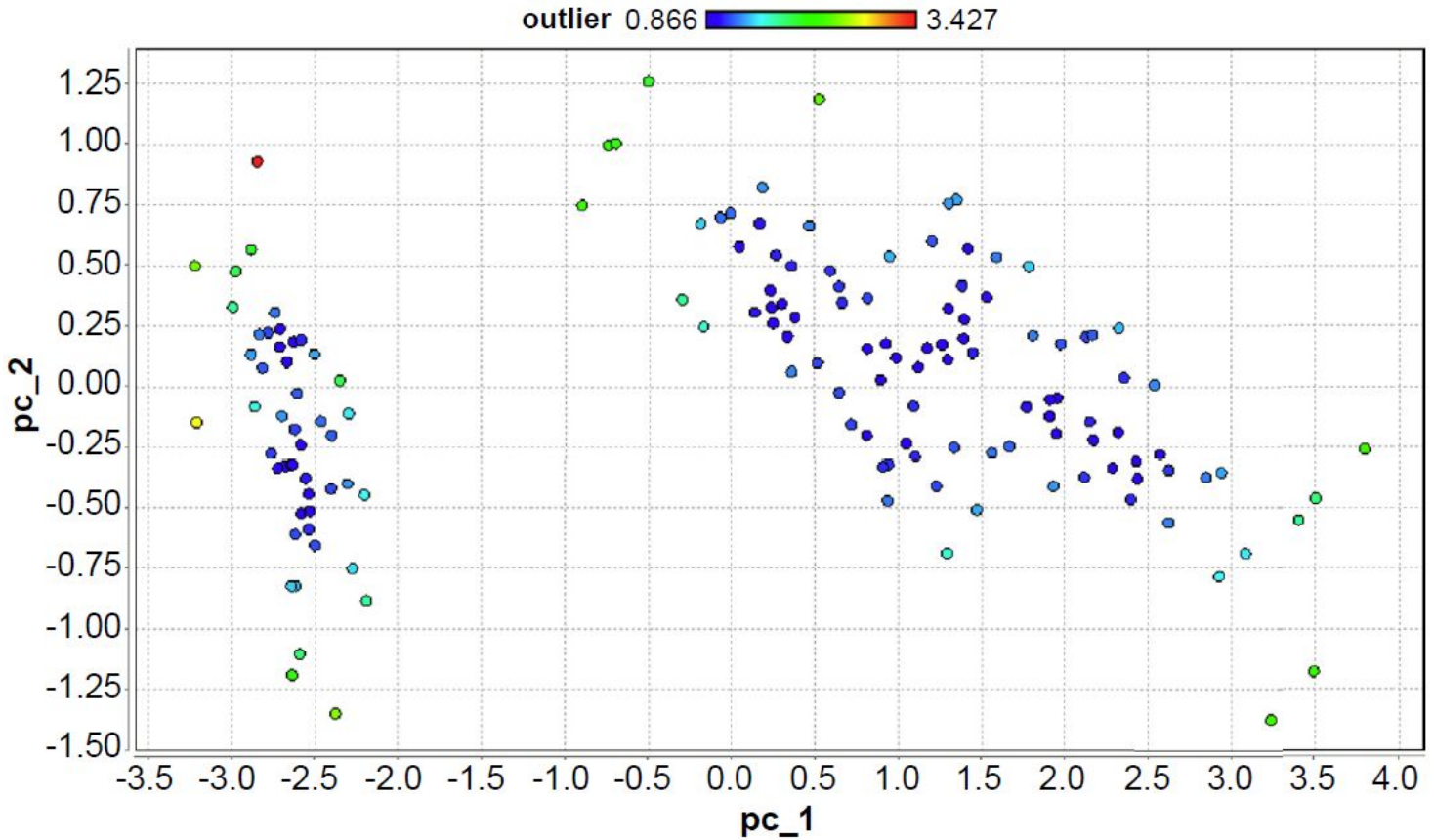
Promedio de las **densidades** de los puntos del vecindario

LOF - Parámetros

rapidminer

1. **Límite inferior** (cant. mínima de puntos)
2. **Límite superior** (cant. mínima de puntos)
3. Función de **distancia**

LOF - Salida

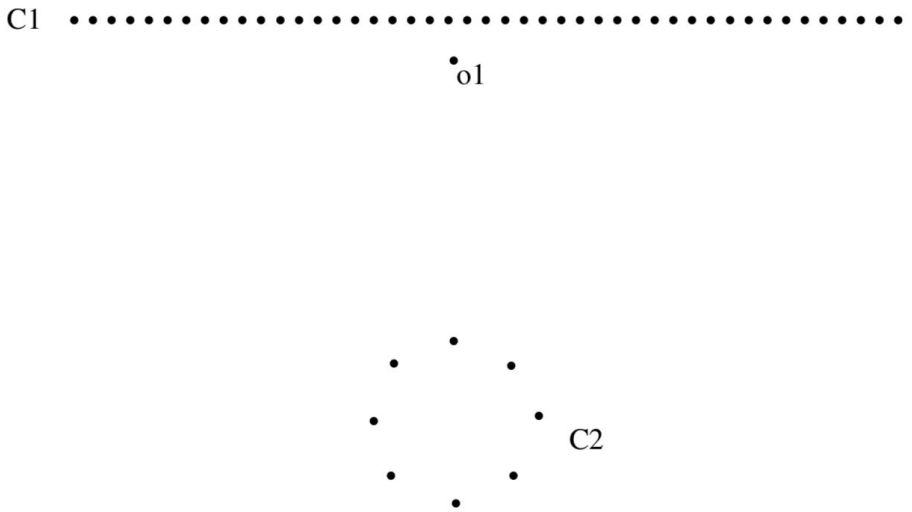

outlier 0.866
3.427
pc_2
pc_1
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

COF

COF - Descripción

Densidad del punto

Aislamiento



Ejemplo
Hipotiroidismo

El dataset

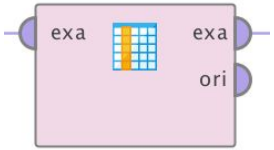
Thyroid Disease Data Set

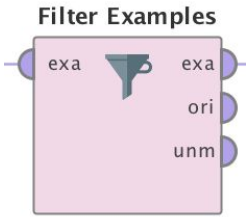
http://archive.ics.uci.edu/ml/datasets/thyroid+disease

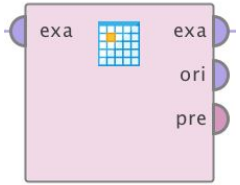
2213 instancias
30 atributos

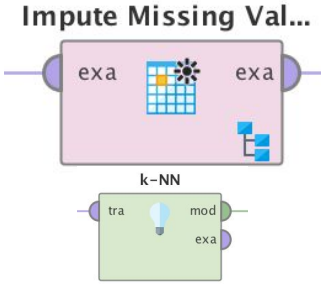
Preprocesamiento del dataset

Preprocesamiento









Atributos utilizados (23)

- age
- sex
- on thyroxine
- query on thyroxine
- on antithyroid med
- sick
- pregnant
- thyroid surgery
- I131 treatment
- query hypothyroid
- query hyperthyroid
- lithium
- goitre
- tumor
- hypopituitary
- psych
- TSH
- T3
- TT4
- T4U
- FTI
- referral source
- class

Atributos eliminados (7)

Por ser irrelevantes

- ❌ TSH measured
- ❌ T3 measured
- ❌ TT4 measured
- ❌ FTI measured
- ❌ T4U measured

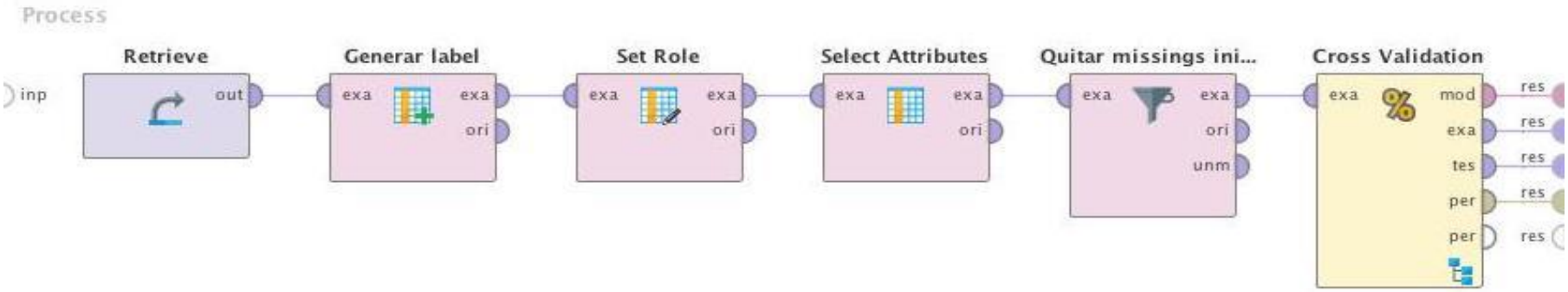
Por falta de información

- ❌ TBG measured
- ❌ TBG

Modelo
Detección de outliers

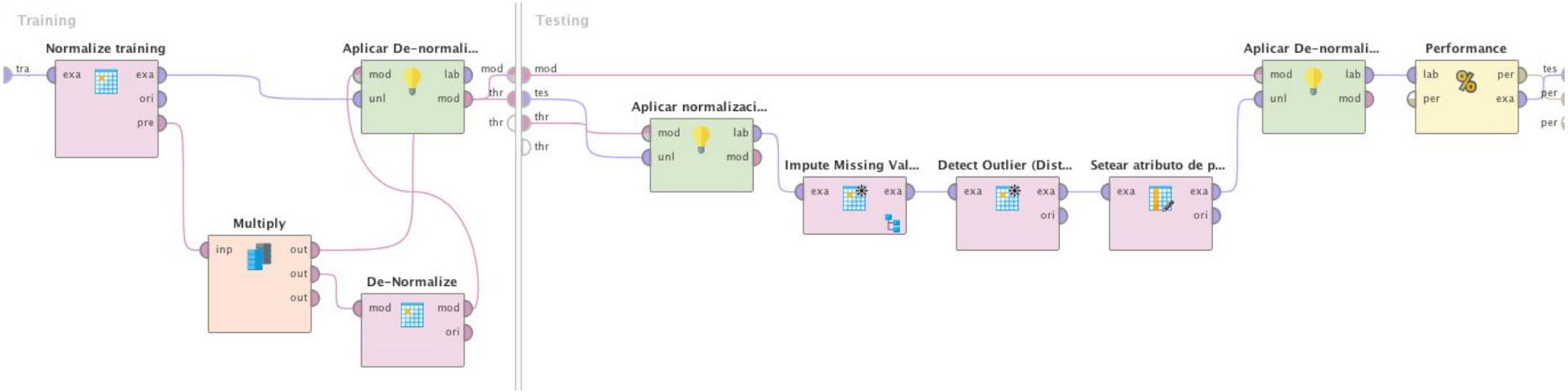
Modelo

Vista de diseño



Modelo

Vista de diseño > cross-validation



Conclusiones
Detección de outliers

Preguntas?

Muchas gracias!