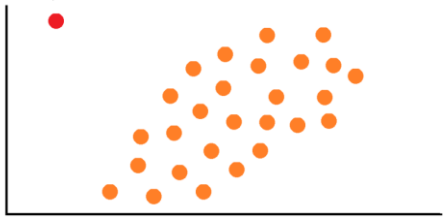
Análisis de valores atípicos

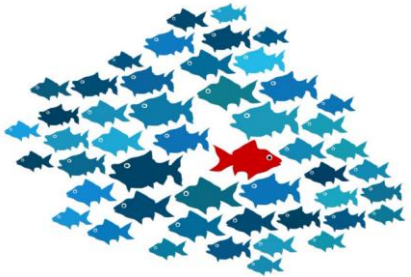
75.06 / 95.58 Organización de Datos
Cat. Ing. Juan M. Rodríguez

Contenidos seleccionados del curso de Posgrado DM- UBA

Análisis de Outliers

- ¿Qué es un “outlier”?
- Tipos de outliers
- Métodos univariados
- Métodos multivariados





Análisis de Outliers

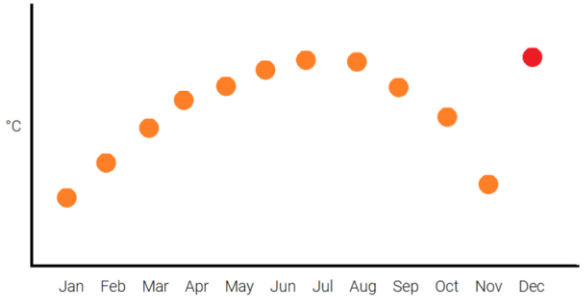
“Un outlier es una observación que se desvía tanto de las otras observaciones como para despertar sospechas que fue generado por un mecanismo diferente”

D. Hawkins. Identification of Outliers (1980)

- Es un concepto subjetivo al problema.
- Son observaciones distantes del resto de los datos
- Pueden deberse a un error de medición, aleatoriedad, que esa instancia pertenezca a una familia distinta del resto, etc

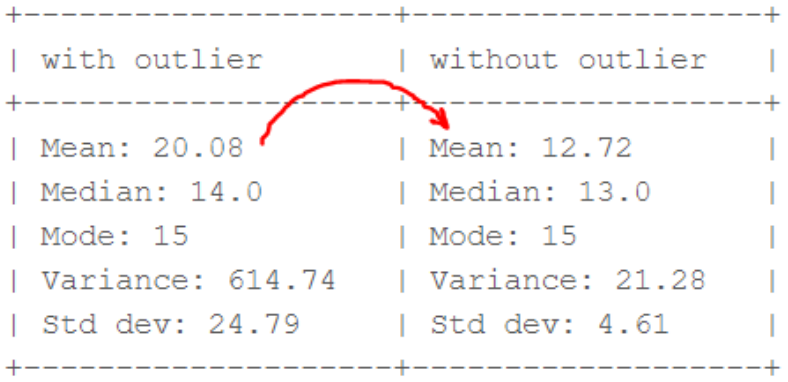
Ejemplo 1: Una persona de 120 años de edad

Ejemplo 2: Una persona de 4 años que mide 1.80mts



Análisis de Outliers

La detección de outliers es importante
su presencia puede influenciar los resultados
de un análisis estadístico clásico.



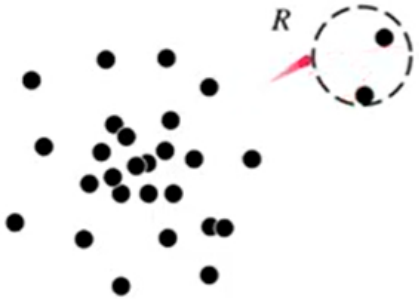
with outlier	without outlier
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61

¿Es necesario eliminarlos?

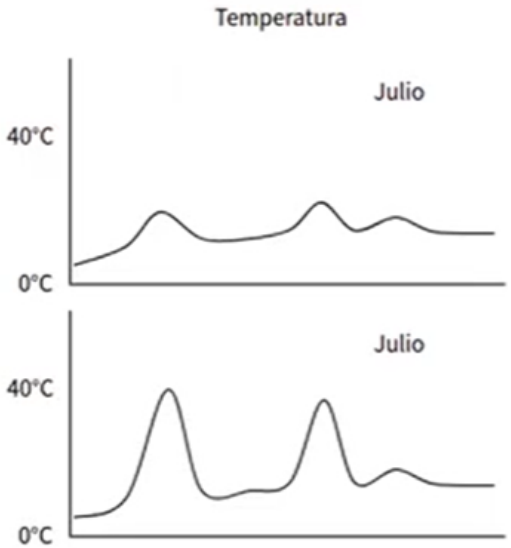
- Deben ser **cuidadosamente inspeccionados**
- Pueden estar alertando anomalías, en algunas situaciones nuestra tarea de interés será encontrarlos :
  - Detección de Fraudes
  - Detección de Fallas
  - Patologías Médicas

Tipos de outlier

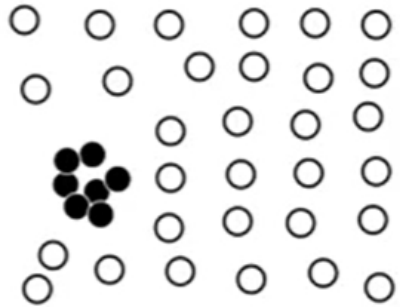
Global Outlier



Contextual Outlier



Collective Outlier



Tipos de outlier

Univariado

- Son valores **atípicos** que podemos encontrar en una simple variable.
- El problema de los enfoques univariados es que son buenos para detección de extremos pero no en otros casos.

Multivariado

- Los valores **atípicos** multivariados se pueden encontrar en un espacio n-dimensional.
- Para detectar valores atípicos en espacios n-dimensionales es necesario ajustar un modelo.

Tipos de outlier

En grandes volúmenes de datos la detección de outliers resulta más eficiente estudiando todas las variables.

Los outliers, en casos multivariados, pueden provocar dos tipos de efectos:

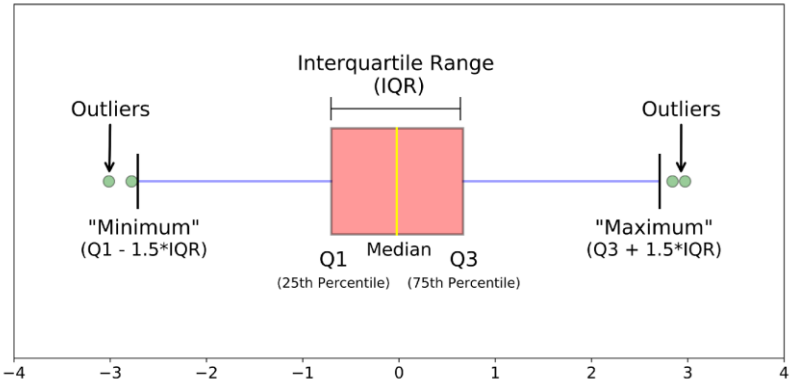
- El **efecto de enmascaramiento** se produce cuando un grupo de outliers esconden a otro/s. Es decir, los outliers enmascarados se harán visibles cuando se elimine/n el o los outliers que los esconden.

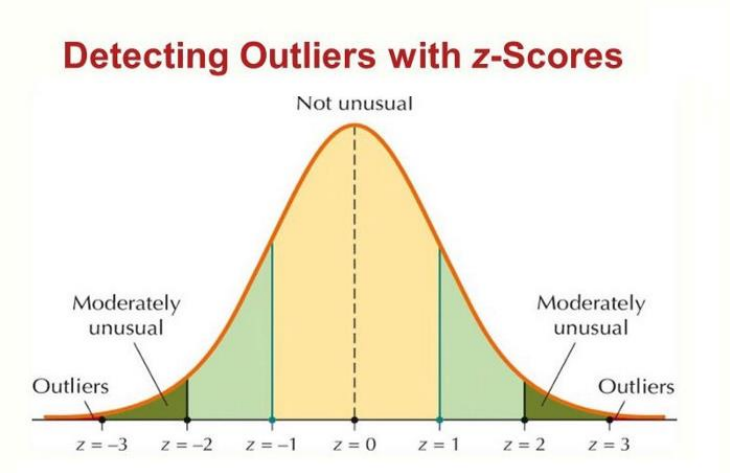
- El **efecto de inundación** ocurre cuando una observación sólo es outlier en presencia de otra/s observación/es. Si se quitara/n la/s última/s, la primera dejaría de ser outlier.

Outliers

Métodos Univariados

- IQR: Analizar los valores que están por fuera del IRQ
- Z-score y Z-score Modificado
- Identificar valores extremos a partir de 1, 2 o 3 desvíos de la media.

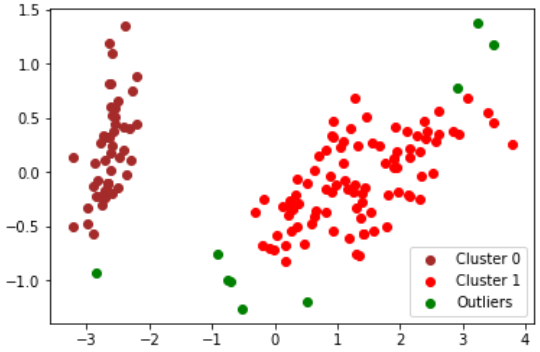


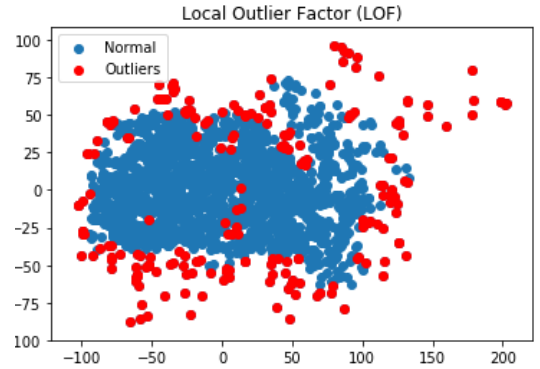


Outliers

Métodos Multivariados

- Análisis globales: Clustering.
  - Utilizando medidas de distancia como Mahalanobis. Los valores similares son agrupados y los que quedan aislados pueden ser considerados outliers.
- Local Outlier Factor (LOF)
  - Es un método de detección de outliers basado en distancias.
  - Calcula un score de *outlier* a partir de una distancia que se normaliza por densidad.
- Métodos basados en árboles de búsqueda: IsolationForest





Métodos Univariados

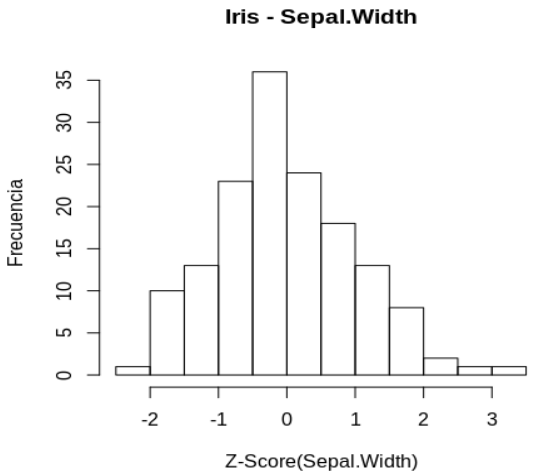
Z-Score

Z-Score es una métrica que indica cuántas desviaciones estándar tiene una observación de la media muestral, asumiendo una distribución gaussiana.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Cuando calculamos Z-Score para cada muestra debemos fijar un umbral:

- Un valor como “regla de oro” es Z > 3



Z-Score Modificado

La media de la muestra y la desviación estándar de la muestra, pueden verse afectados por los valores extremos presentes en los datos

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Regla de oro:
valores mayores a 3.5 son considerados outliers

Median Absolute Deviation

$$MAD = median\{|x_i - \tilde{x}|\}$$

Es la mediana de los desvíos absolutos respecto de la mediana.

Para hacer MAD comparable con la desviación estándar, se normaliza por 0.6745

Análisis de Box-Plot

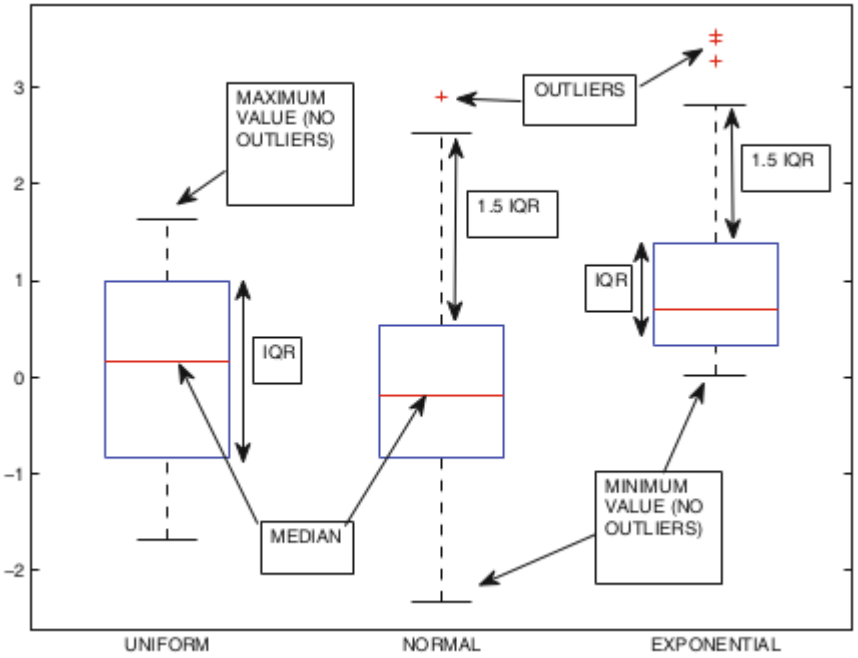
Los Box-Plots permiten visualizar valores extremos univariados.

Las estadísticas de una distribución univariada se resumen en términos de cinco cantidades:

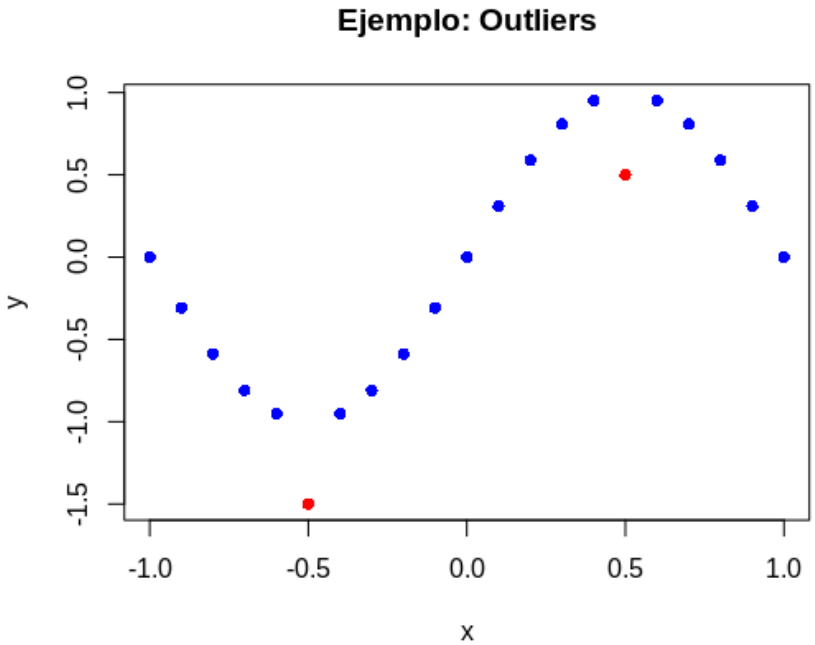
- Mínimo/máximo (bigotes)
- Primer y tercer cuantil (caja)
- Mediana (línea media de la caja)
- IQR = Q3 - Q1

Generalmente la regla de decisión:

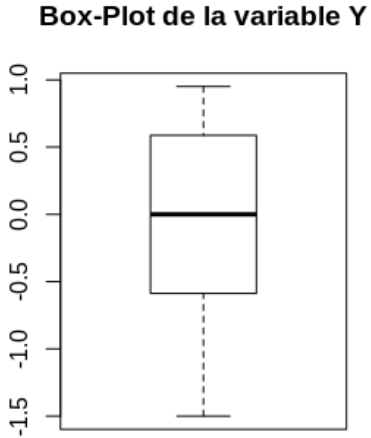
+/- 1.5*IQR Outliers moderados
+/- 3 *IQR Outliers severos



Análisis de Box-Plot



¿Qué pasa con el box-plot?



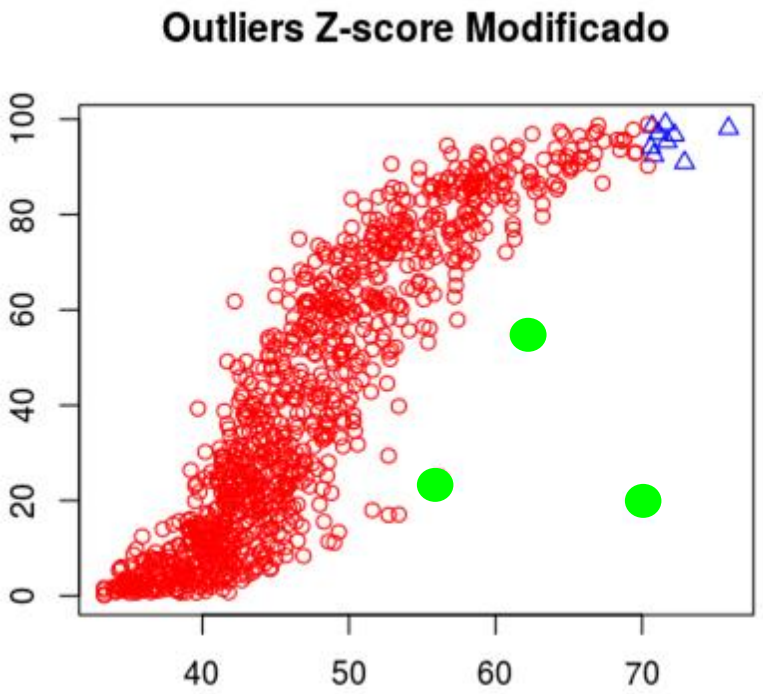
En el scatter se observan dos valores atípicos.

Problema

Una forma de tratar valores atípicos es eliminar los valores más altos y más bajos de una variable.

Esto puede funcionar bastante bien, pero no tiene en cuenta las **combinaciones de variables**.

¿Qué ocurre con los casos ● ?



Métodos Multivariados

Distancia de Mahalanobis

Es una medida de distancia entre el punto $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$

y un conjunto de observaciones con media $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$

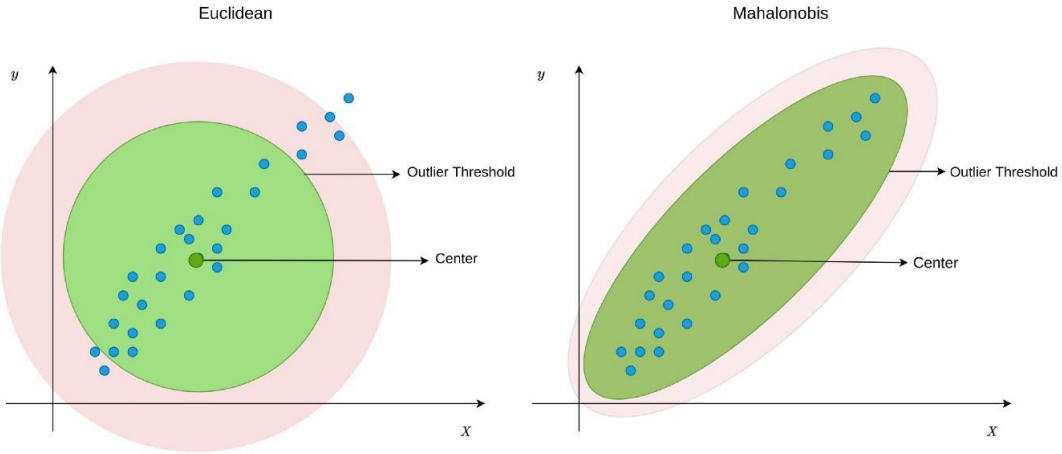
y una matriz de covarianza S.

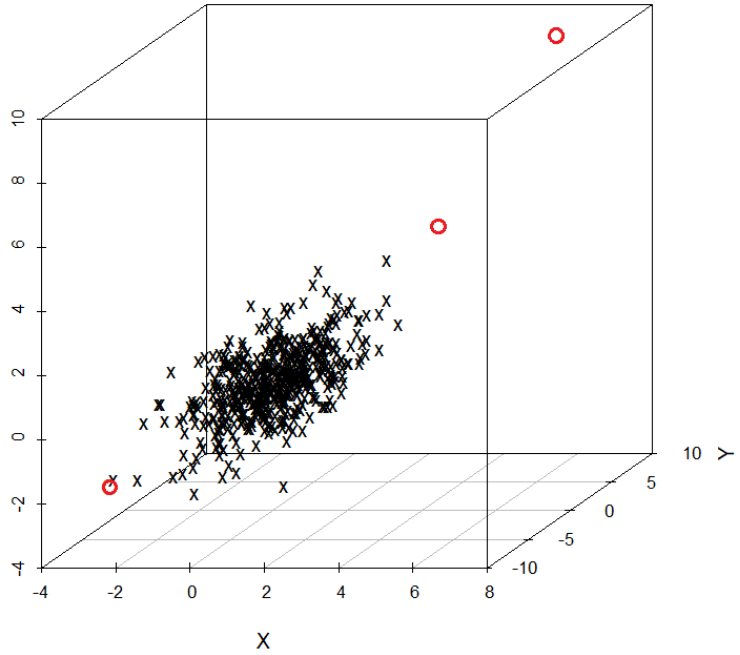
$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

Matriz de distancias con respecto a la media

Inversa de la matriz de covarianzas

Distancia de Mahalanobis





LOF – Local Outlier Factor

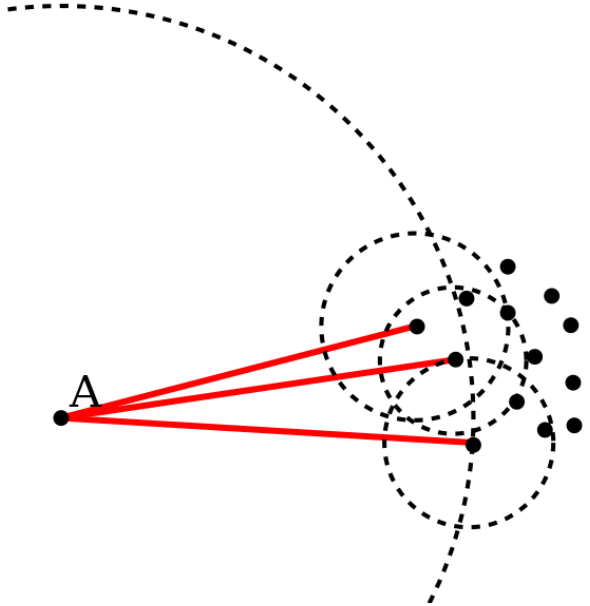
El método LOF valora puntos en un conjunto de datos multivariados.

Es un **método basado en densidad** que utiliza la búsqueda de vecinos más cercanos.

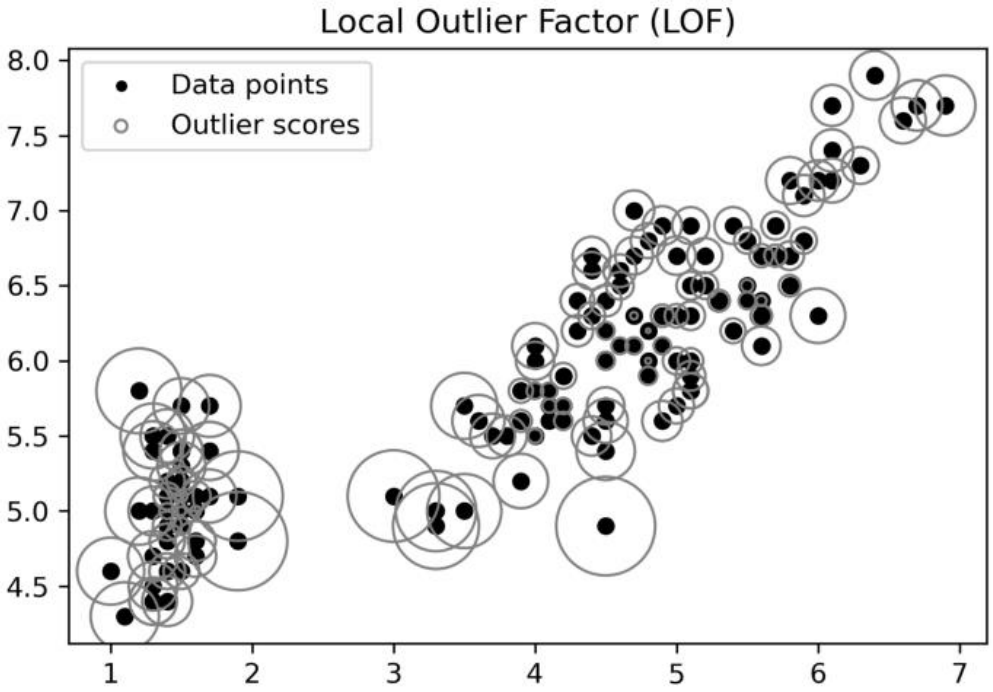
- Se compara la densidad de cualquier punto de datos con la densidad de sus vecinos
- Parámetro k (cantidad de vecinos) y métrica de distancia

El método calcula los ***scores*** para cada punto, se debe definir un umbral de corte (depende del dominio)

- Si el score del punto X es 5, significa que la densidad promedio de los vecinos de X es 5 veces mayor que su densidad local

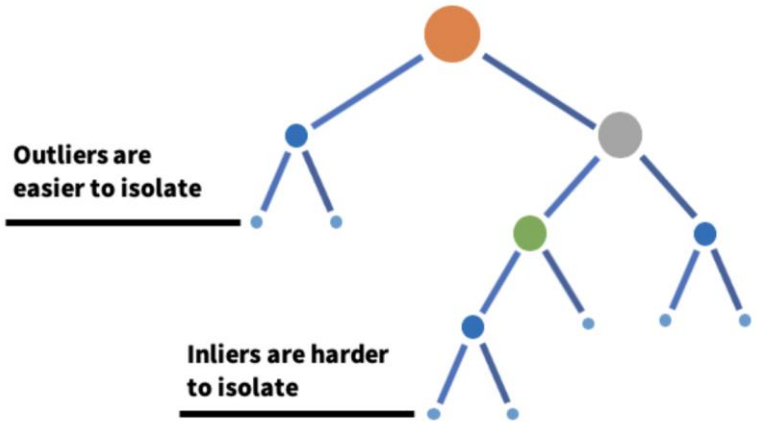


LOF – Local Outlier Factor

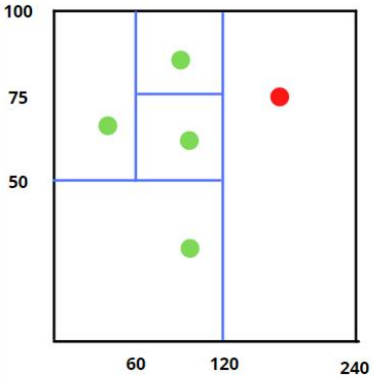


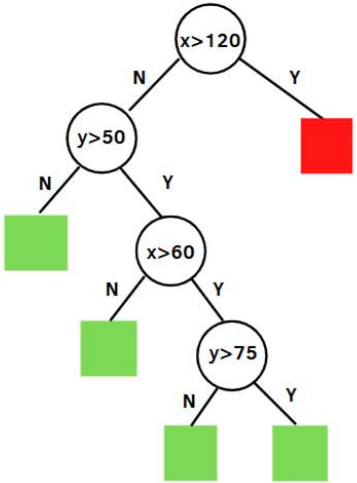
Isolation Forest

- Es un algoritmo no supervisado y no paramétrico basado en árboles de decisión.
- Idea Principal: los datos anómalos se pueden aislar los datos normales mediante particiones recursivas del conjunto de datos.



Isolation Forest





Tomar una muestra de los datos y construir un árbol de aislamiento:

1. Seleccionar aleatoriamente **n características** .

2. Dividir los puntos de datos seleccionando aleatoriamente un **valor** entre el mínimo y el máximo de las características seleccionadas.

La partición de observaciones se repite recursivamente hasta que todas las observaciones estén aisladas.

Isolation Forest

Isolation Forest identifica anomalías como las observaciones con longitudes de ruta promedio cortas en los árboles de aislamiento.

- Utiliza la altura del árbol (cantidad de aristas)