GESTIÓN DE OUTLIERS

Tipos de Outliers

Univariados: En base a una sola variable

Multivariados: En base a la relación entre al menos 2 variables

Detección Univariados

Visualizaciones

- Boxplots
- Histogramas

Estadísticas Descriptivas

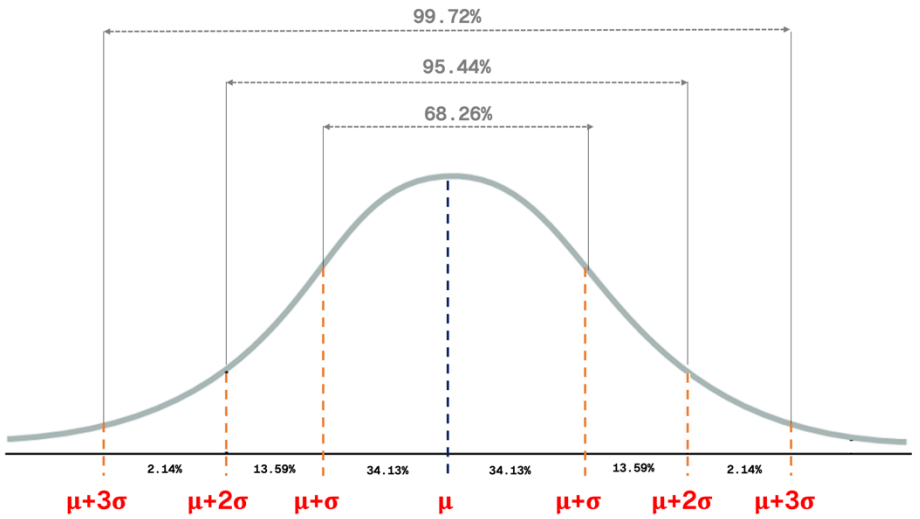
- Desviación Estándar
- Media vs Mediana

Métodos Estadísticos Descriptivos

- Z-score: cuando los datos siguen una dist. normal.
- IQR: si los datos no siguen una dist. normal.

Z-score

- El **z-score** mide cuántas desviaciones estándar se aleja un dato de la media (promedio) del conjunto de datos.
- Si el z-score de un valor es 0, significa que está justo en la media.
- Valores de z-score mayores o menores indican qué tan lejos está un dato de la media:
  - Un z-score de 1 significa que el valor está a 1 desviación estándar por encima de la media.
  - Un z-score de -2 significa que el valor está a 2 desviaciones estándar por debajo de la media.
- **Usaremos el z-score** cuando los datos siguen una distribución aproximadamente normal (en forma de campana).



Un umbral que se suele usar comúnmente es 3. Este valor establece que los datos dentro de 3 veces la desviación estándar de la media representan el 99.7% de los datos en la distribución. Podemos concluir que los puntos de datos que caen más allá de este umbral son *outliers* porque difieren del 99.7% de los datos.

IQR

- El **IQR** es una medida que indica el rango en el que se encuentra la mitad central de los datos.
- **Usaremos el IQR** cuando los datos tienen valores extremos o no siguen una distribución normal, ya que el IQR es menos sensible a estos valores extremos.

$$Limite\ Inferior = Q1 - (k\ * IQR)$$

$$Limite\ Superior = Q3 + (k\ * IQR)$$

Donde:
- Q1: el valor del primer cuartil
- Q2: el valor del tercer cuartil
- IQR: rango intercuartílico (Q3- Q1)
- k: factor para multiplicar por el IQR para determinar los límites.

Detección Multivariados

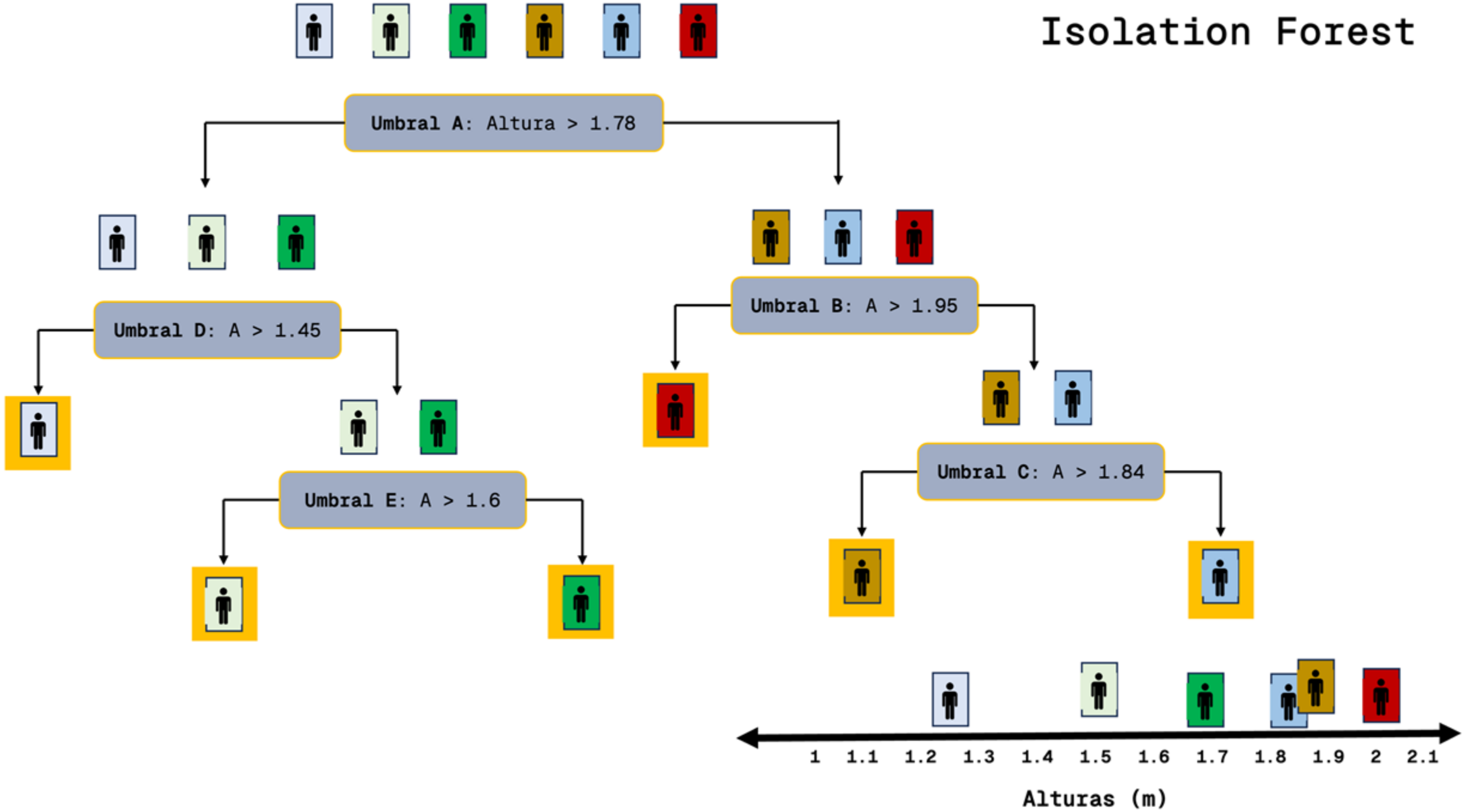
Isolation Forest

- **Utilidad**: Detecta valores atípicos que son raros y claramente diferentes al resto de los datos.
- **Cuando usarlo**: Conjuntos de datos en los que los outliers son escasos y muy distintos.
- **Funcionamiento**: Crea árboles de decisión que dividen los datos aleatoriamente; los puntos que requieren pocas divisiones para ser aislados suelen ser outliers.

Local Outlier Factor

- **Utilidad**: Encuentra anomalías en zonas con diferentes densidades, ideal para detectar outliers locales.
- **Cuando usarlo**: Conjuntos de datos donde algunos puntos están agrupados y otros más dispersos.
- **Funcionamiento**: Compara la densidad de un punto con la de sus vecinos cercanos; los puntos con densidad baja comparada a sus vecinos son considerados outliers.

Isolation Forest



Local Outlier Factor (LOF) – Como Funciona

Calcular la distancia entre todos los puntos

Buscamos el vecino más cercano de cada punto (si K = 2, buscamos el 2º punto más próximo a cada uno de los puntos)

Buscamos los k-vecinos más cercanos. Si K = 2, tendremos que buscar solo 1 punto.

Calcular el LDR, es la distancia estimada a la que un punto puede ser encontrado por sus vecinos. Nos ayuda a entender como de accesible es cada punto.

Calculamos el Factor Local de Anomalía (LOF), que corresponde con la suma de todos los LDR dividido por el número de datos.

Evaluamos si se trata de un Outlier:

- Si el LOF es menor que 1: el punto es normal
- SI el LOF es mayor que 1: el punto puede ser considerado outlier

Local Outlier Factor (LOF) – Parámetros (K)

Datos Muy Similares (Pocos Vecinos)	Datos Variados (Más Vecinos):	Probar y observar
• Por ejemplo, alturas de estudiantes en una escuela (la mayoría mide entre 1.60 m y 1.70 m) • Usar pocos vecinos (k entre 5 y 20). • LOF podrá detectar pequeñas diferencias y encontrar anomalías con mayor precisión (por ejemplo, un estudiante que mide 2 m).	• Por ejemplo, alturas de personas de todas las edades (niños, adultos, personas mayores). • Usar más vecinos (k entre 20 y 50). • LOF detectará anomalías más evidentes y no se enfocará en diferencias pequeñas dentro de cada grupo.	• Por ejemplo, ingresos mensuales de empleados en una empresa con salarios muy diferentes según el área. • Comienza con un valor bajo de k (como 10) y observa los resultados. • Aumenta k si detectas muchas “anomalías” para concentrarte solo en los valores realmente atípicos.

Documentación Adicional

- LOF: https://doedotdev.medium.com/local-outlier-factor-example-by-hand-b57cedb10bd1