Limpieza de datos - Datos Faltantes

Tipos de datos faltantes

* MCAR: Missing completely at random. La razón de la falta de datos es ajena a los datos mismos. No existen relaciones con la variable misma donde se encuentren los datos faltantes o con las restantes variables en el dataset que expliquen por qué faltan.
* MNAR: Missing not at random. La razón depende del dato en sí.
* MAR: Missing at random. La causa de los datos faltantes no depende de estos mismos datos faltantes pero puede estar relacionada con otras variables.

Estrategias para trabajar con datos faltantes

***Eliminar registros o variables*** -> Si la eliminación de un subconjunto disminuye significativamente la utilidad de los datos, la eliminación del caso puede no ser efectiva. No se recomienda en situaciones que no sean MCAR.

***Imputar datos*** -> Vamos a utilizar diferentes métodos para completar esos valores que faltan. Muchas veces se van a poder lograr mediante métodos de aprendizaje o consultando a expertos del dominio.

### Imputación de datos

#### Sustitución de casos

Se reemplaza con valores no observados. Debería ser realizado por un experto en esos datos.

#### Sustitución por Media o Mediana

Se reemplaza utilizando la medida calculada de los valores presentes. Algunas desventajas:

* La varianza estimada de la nueva variable no es válida porque está atenuada por los valores repetidos.
* Se distorsiona la distribución.
* Las correlaciones que se observen estarán deprimidas debido a la repetición de un sólo valor constante.

#### Imputación Cold Deck

Selecciona valores o usa relaciones obtenidas de fuentes distintas de la base de datos actual.

#### Imputación Hot Deck

Se reemplazan los faltantes con valores obtenidos de registros que son los más similares. Hay que definir qué es “similar”.

#### Imputación por regresión

El dato faltante es reemplazado con el valor predicho por un modelo de regresión.

#### Imputación MICE - Multivariate Imputation by Chained Equations

Trabaja bajo el supuesto de que el origen de los faltantes es Missing at Random (MAR).

Es un proceso de imputación de datos faltantes iterativo con el cual en cada iteración cada valor faltante de cada variable se predice en función de las variables restantes.

Esta iteración se repite hasta que se encuentre convergencia en los valores. Por lo general 10 iteraciones es suficiente.

En cada iteración se genera un dataset.

Transformación de Datos - Feature Engineering

Esta etapa incluye cualquier proceso de modificación de la forma de los datos (es común que los datos sufran algún tipo de transformación). El objetivo principal de esta etapa es mejorar el rendimiento de los modelos creados mediante la transformación de los datos que utilizan.

Algunas técnicas son:

* Normalización
* Discretización
* Lograr normalidad
* Imaginación (Generación de nuevas variables)

Normalización

Se aplica sobre valores numéricos. Consiste en escalar los features de manera que puedan ser mapeados a un rango más pequeño.

Es principalmente utilizada cuando:

* Las unidades de medidas dificultan la comparación de features.
* Se quiere evitar que atributos con mayores magnitudes tengan pesos muy diferentes al resto.

### Min Max

Funciona al ver cuánto más grande es el valor actual del valor mínimo del feature y escala esta diferencia por el rango.

### Z score

Los valores para un atributo se normalizan en base a su media y desvío estándar. Es útil cuando el verdadero mínimo y máximo del atributo no son conocidos, o cuando hay valores atípicos que dominan la normalización min max.

### Decimal scaling

Asegura que cada valor normalizado se encuentre entre -1 y 1.

Discretización

Es una técnica que permite dividir el rango de una variable continua en intervalos. Se reducen los valores de una variable contínua a un número reducido de etiquetas.

### Binning

Se divide a la variable en un número específico de bins. Los criterios de agrupamiento pueden ser por ejemplo:

* Igual-Frecuencia: la misma cantidad de observaciones en un bin.
* Igual-Ancho: definimos rangos o intervalos de clases para cada bin.
* Cuantiles: separar en intervalos utilizando Mediana, Cuantiles, Percentiles.

A su vez para cada uno de los agrupamientos podemos hacer:

* Reemplazo por media o mediana
* Reemplazo por una etiqueta o valor entero

Variables Dummies - One Hot Encoding

Algunos métodos analíticos requieren que las variables predictoras sean numéricas. Cuando tenemos categóricos, podemos recodificar la variable categórica en una o más variables Dummies.

Generación de nuevas variables

Por ejemplo: sumar fuentes de información para calcular la distancia desde un inmueble en venta al espacio verde más cercano.

Análisis de valores atípicos

Análisis de Outliers

**Outlier**: observación que se desvía tanto de las otras como para despertar sospechas que fue generado por un mecanismo diferente. 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 del resto, etc.

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

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

### 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.

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. Los outliers enmascarados se harán visibles cuando se elimine/n el o los outliers que los esconden.
* El efecto inundación → Ocurre cuando una observación solo es outlier en presencia de otra/s observación/es. Si se quitara/n la/s última/s, la primera dejaría de ser outlier.

### 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.

### 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

Es una métrica que indica cuántas desviaciones estándar tiene una observación de la media muestral asumiendo una distribución gaussiana. 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.

La regla de oro es que valores mayores a 3.5 son considerados outliers.

MAD → Median Absolute Deviation: 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 los valores extremos univariados. Las estadísticas de una distribución univariada se resumen en términos de cinco cantidades:

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

Generalmente la regla de decisión es:

* outliers moderados → +/- 1.5\*IQR
* outliers severos → +/- 3\*IQR

# Métodos Multivariados

## Distancia de Mahalanobis

Es una medida de distancia entre un punto *x* y un conjunto de observaciones con media *u* y una matriz de covarianza *S*.

## LOF - Local Outlier Factor

Es un método que valora puntos en un conjunto de datos multivariados. Está 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.

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.

## Isolation Forest

Es un algoritmo no supervisado y no paramétrico basado en árboles de decisión.

Los datos anómalos se pueden aislar de los datos normales mediante particiones recursivas del conjunto de datos.

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.