Imputacion por KNN

Para cada observacion con valores faltantes:

1. Encuentra otras k observaciones (donantes, vecinos) que sean mas similares a esa observacion
2. Reemplaza los valores faltantes con los valores agregados de los k vecinos. Si los vecinos mas cercanos tienen valores de 1 y 2, entonces el valor a imputar es el promedio de ambos que es 1.5

Para determinar cuales son los vecinos mas similares se utilizan las metricas de distancia la cual cuantifica la distancia que hay entre dos observaciones. Entre los distintos tipos de metrica de distancia tenemos:

* **Distancia Euclidiana:** util para las variables numericas, se traza una linea recta entre dos puntos, porque se considera el camino mas corto en un espacio euclidiano, y esa distancia se utiliza como parametro para definir si se encuentra lejos o cerca de otro punto.

Untitled.png

* **Distancia Manhattan:** es muy util para variables de tipo factor (dias de la semana, condiciones como lento, rapido), se trata de la distancia absoluta que debemos recorrer para llegar del punto a al punto b.

Untitled (1).png

* **Distancia de Hamming:** es util para variables categoricas, cuantifica la diferencia entre los textos.

Untitled (2).png

* **Distancia de Gower:** util para conjuntos de datos con variables mixtas, se define como la distancia de un vecino a otros vecinos para determinar la distancia global referente a la fila que contiene el valor faltante.

Untitled (3).png

**Ventajas**

* Sencillo de implementar.
* buen rendimiento con conjunto de datos pequenos.
* Excelente para datos numericos, pero tambien funciona para datos mixtos

**Desventajas**

* Su escalabilidad puede ser comprometedora dependiendo del numero de variables y el tamano de las observaciones.
* Requiere transformaciones especiales para las variables categoricas ya que no solo estas introduciendo numeros y se necesita cuntificar estas distancias.
* Posee sensibilidad a valores atipicos