Datos Faltantes

1. Definición de dato faltante.
2. Formas de abordar este problema.
3. Técnicas de imputación de datos.

Definición de dato faltante (Missing Value)

Los MV pueden generarse por distintas razones como:

- Fallas en capturar datos por efectos ambientales o fallas de hardware (redes de sensores).
- No dar respuesta por decisión propia (redes sociales).
- Prohibición por protocolos (hospitales).
- Falta de recursos: dinero, tiempo o máquinas (hospitales, observatorios, etc.)

ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	**?**	60	No	1
7	Yellow	100	**?**	2
8	**?**	40	**?**	1

```
df
```

	column_a	column_b	column_c	column_d
0	1.0	1.2	a	True
1	2.0	1.4	?	True
2	4.0	NaN	c	NaN
3	4.0	6.2	d	None
4	NaN	NaN	--	False
5	NaN	1.1	NaN	True
6	6.0	4.3	d	False

Se supone que hay un modelo detrás de la generación de MV.

- Missing completely at random (MCAR): ocurre cuando la probabilidad de que una variable tenga valor faltante es independiente de la misma variable y de cualquier otra influencia externa. Lo que significa que el los MV no dependen de los datos.

- Missing at Random (MAR): ocurre cuando la probabilidad de que una variable tenga valor faltante es independiente de las variables con valores faltantes pero dependiente de las otras variables con valores observables. Esto sugiere un supuesto menos restringido.

- Not Missing at Random (NMAR): ocurre cuando la probabilidad de que una variable tenga valor faltante no es al azar, por lo tanto depende de las variables faltantes.

Trabajaremos bajo el supuesto MAR

Formas de abordar el problema de Missing Value (MV)

¿Como abordar los MV?

Existen 3 grandes enfoques:

- Eliminar los ejemplos (o variables) con MV: sencillo, pero no se puede aplicar siempre. Solo cuando el porcentaje de MV es pequeño para cuando queramos eliminar ejemplos. Si queremos eliminar una variable, debe tener muchos datos faltantes esa variable.

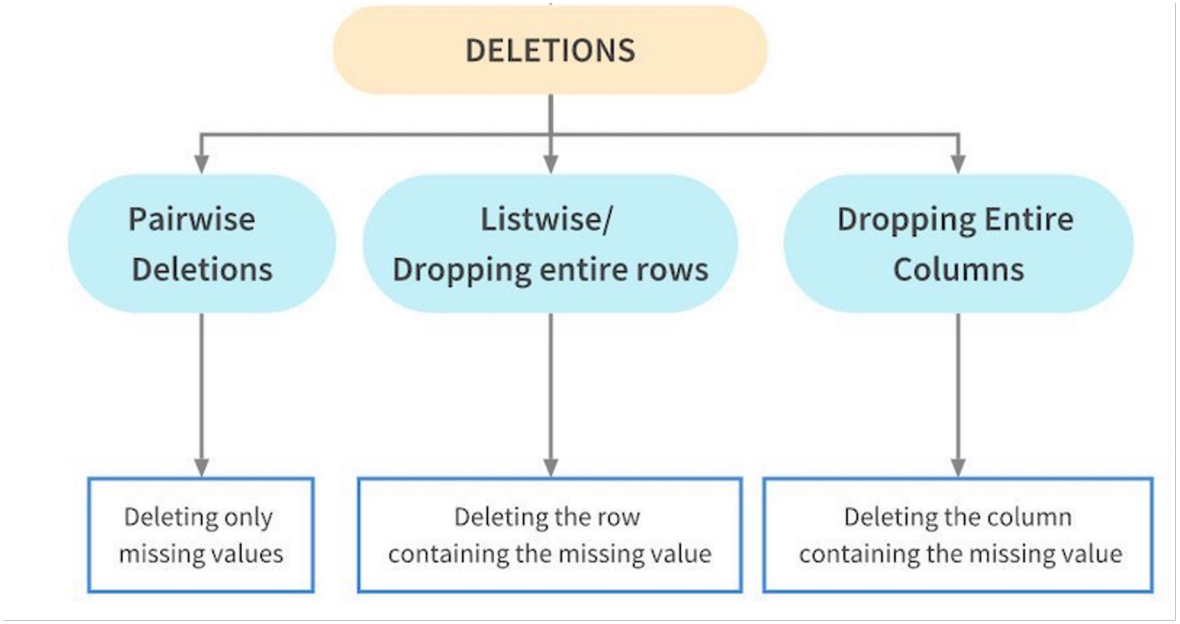
- Imputación del MV: el MV se estima con alguna técnica de estimación de datos. Puede ocurrir que la estimación sea tan mala que sea peor el remedio que la enfermedad.

- Trabajar con los datos tal cual vienen (sin eliminar, sin imputar): aquí todo el trabajo se lo lleva el clasificador, regresor u otro algoritmo. Añade complejidad al algoritmo y además, puede producir sesgo.

¿Como abordar los MV?

Eliminar los ejemplos (o variables) con MV: sencillo, pero no se puede aplicar siempre. Solo cuando el porcentaje de MV es pequeño.

Si el porcentaje de MV es grande, debemos pensar en otra solución. Es bueno ver como están distribuido los MV en el dataset completo.



Fuente: https://www.kaggle.com/parulpandey/a-guide-to-handling-missing-values-in-python

¿Como abordar los MV?

Trabajar con los datos tal cual vienen (sin eliminar, sin imputar):
aquí todo el trabajo se lo lleva el clasificador, regresor u otro algoritmo. Añade complejidad al algoritmo y además, puede producir sesgo.

Existen algunos algoritmos que pueden manejar MV de manera natural, aunque son los menos. Uno de ellos es: *Árboles de decisión*

También existen versiones mejoradas de algunos algoritmos que permiten lidiar con MV. En este caso, podría ser una solución interesante y que entregue buenos resultados, pero no asegura que sea la mejor solución.

¿Como abordar los MV?

Imputación del MV: el MV se estima con alguna técnica de estimación de datos. Puede ocurrir que la estimación sea tan mala que el remedio sea peor que la enfermedad.

Por esto ultimo, es importante que la imputación de datos sea la mejor posible.

Existen muchos esfuerzos para mejorar las técnicas de imputación. Aquí hablaremos de algunas, ya que en muchos casos es lo mejor que podemos hacer.

Destacamos que el proceso de manejo de MV pertenece a la fase de preprocesamiento de datos. Incluso, es una de las primeras cosas que se deben hacer.

Técnicas de imputación de datos

Técnicas de imputación

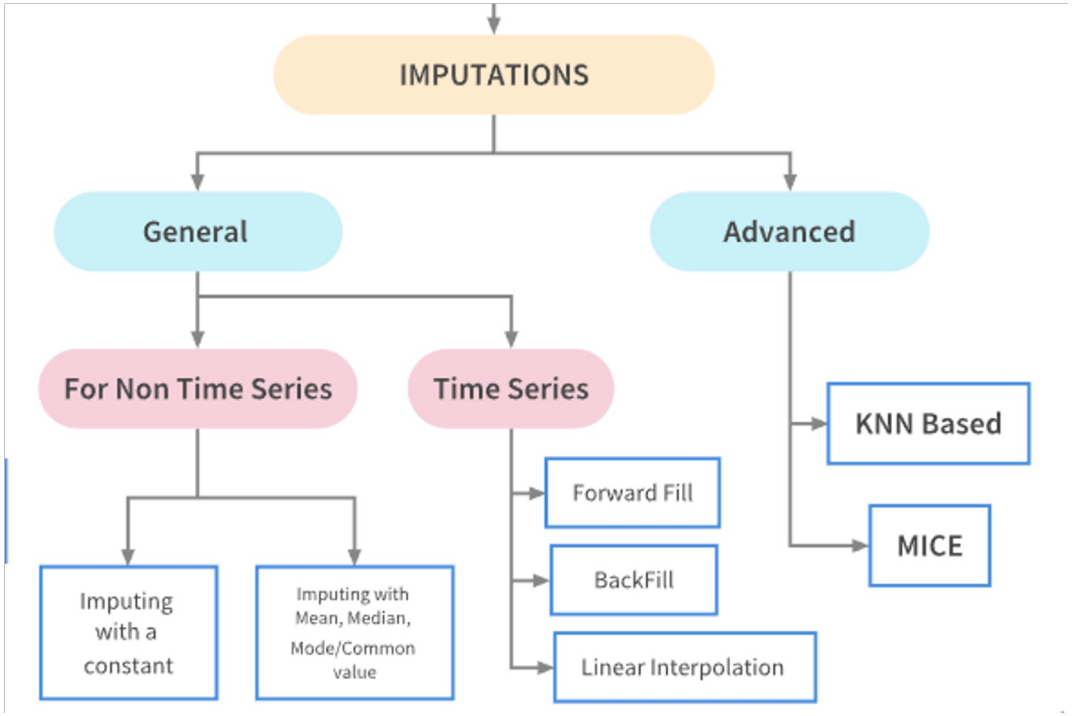
En muchos casos, la mejor opción es imputar datos, ya que esto permite trabajar con todos los datos aunque produzca un sesgo.

Existen muchas técnicas para lidiar con esto. La taxonomía no está 100% definida, pero acá entregamos un ejemplo:

No veremos técnicas para series de tiempo (Time Series).

Existe un grupo de técnicas simples que ven el problema solo como una variable. Estas son tecnicas Univariadas.

Técnicas más avanzadas como SVD, KNN, EM son multivariadas. Veremos más adelante la técnica KNN por ser una estrategia general para resolver muchos problemas.

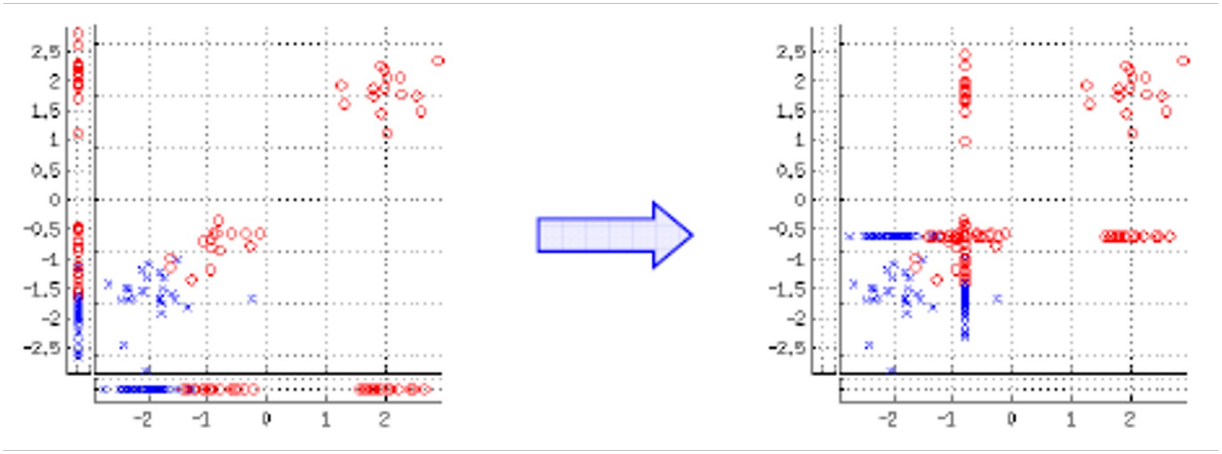


Fuente: https://www.kaggle.com/parulpandey/a-guide-to-handling-missing-values-in-python

» Zero: imputa con cero si hay MV. La técnica más simple (y peor) de imputación. No lo haga a no ser que los datos esten estandarizados….

¿por que?

» Media (mediana sigue la misma estrategia): es mejor, pero cuidado con el efecto que puede producir.



- Las técnicas de imputación están enfocadas a variables numéricas.

- Para variables categóricas se podría imputar con la moda, pero es preferible realizar un encoding antes, identificando si es una variable categórica nominal u ordinal.

- Existe una mejora a la técnica de la media cuando hay presencia de datos atípicos o Outliers.