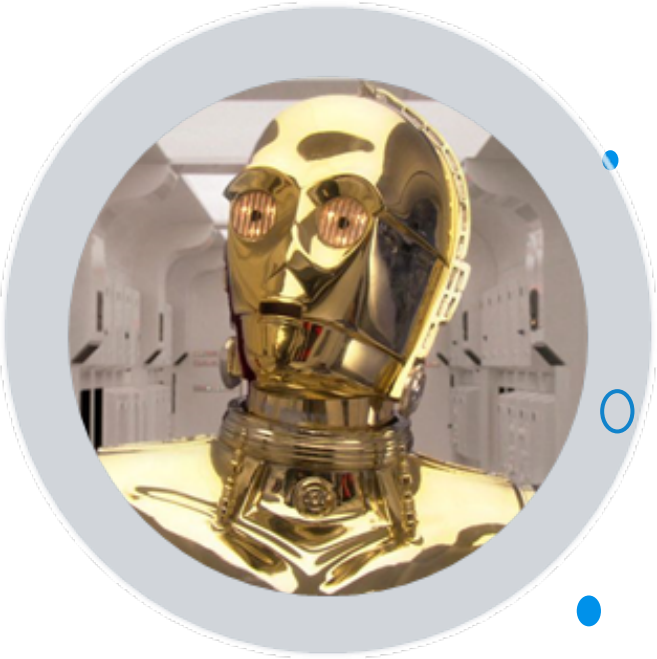
NO MÁS PERDIDO CON LOS DATOS PERDIDOS



by Estephani Rivera Jaramillo

Cox, Gertrude Mary

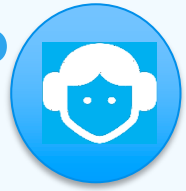
“ Lo mejor que se puede hacer con los datos que faltan es no tener ninguno ”

by Estephani Rivera Jaramillo

Introducción

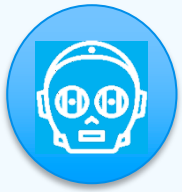
Qué son los datos perdidos y su importancia

Tipos de datos perdidos



Métodos de imputación adecuado

Otras estrategias para tratar datos perdidos



Código (Rstudio)

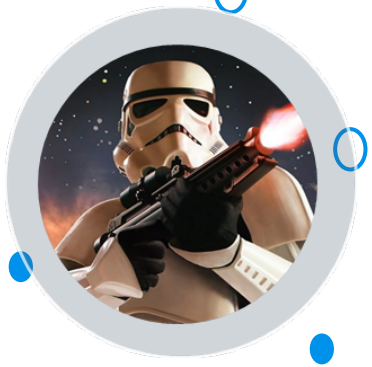
Kahoot!

by Estephani Rivera Jaramillo

Introducción

Trabajar con datos del mundo real significa trabajar con datos faltantes.

Comprender cómo funcionan los datos faltantes es importante, ya que pueden tener efectos inesperados en su análisis.	La elección del método para imputar los valores perdidos influye en gran medida en la capacidad predictiva del modelo.	En la mayoría de los métodos de análisis estadístico, la eliminación por lista es el método predeterminado utilizado para imputar los valores perdidos.	Pero no es tan bueno ya que conduce a la pérdida de información.



by Estephani Rivera Jaramillo

Que son los datos perdidos

Son aquellos que no constan debido a cualquier acontecimiento, como por ejemplo errores en la transcripción de los datos o la ausencia de disposición a responder a ciertas cuestiones de una encuesta.

Los datos pueden faltar de manera aleatoria o no aleatoria.

Datos faltantes aleatorios

- Pueden perturbar el análisis de datos dado que disminuyen el tamaño de las muestras y en consecuencia la potencia de las pruebas de contraste de hipótesis.

Dato faltantes no aleatorios

- Ocasionan además, disminución de la representación de la muestra.

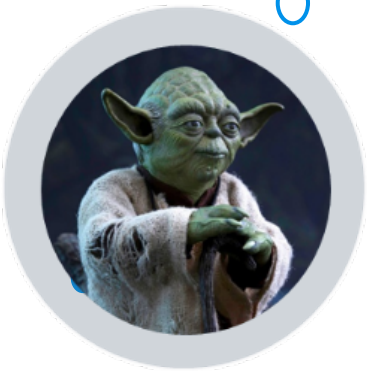
by Estephani Rivera Jaramillo

by Estephani Rivera Jaramillo

Porque son importantes los datos perdidos?

3 motivos más importantes por los que se suelen tratar los valores perdidos son:

- Pueden introducir un sesgo considerable (una diferencia notable entre los datos observados y los no observados).
- Hacen el análisis y el manejo de los datos más complicado
- Pérdida de información que estos producen.



by Estephani Rivera Jaramillo

Tipos de datos perdidos

MAR

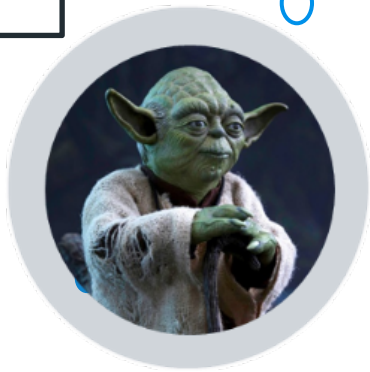
Missing At Random , La ausencia de datos está asociada a variables observables presente en el conjunto de datos.

MNAR

Los patrones de pérdida no ignorables, son los que ocurren cuando la ausencia de los datos depende la variable perdida, aquí debes estudiar el patrón de perdida de datos ausentes para luego imputar.

MCAR

Missing Completely At Random, Es decir la ausencia de la información no ha sido originada por ninguna variable presente en el conjunto de datos



by Estephani Rivera Jaramillo

by Estephani Rivera Jaramillo

Tipos de datos perdidos - Ejemplos

MAR

Por ejemplo si observamos genero, raza, educación y edad para todos los encuestados, entonces ingresos es MAR si la probabilidad de no-respuesta para esta pregunta depende únicamente de estas variables completamente observada.

MNAR

Es común que en ensayos clínicos si un tratamiento particular causa molestias los pacientes son más propensos a abandonar el estudio, y dado que se busca medir la eficacia de cada tratamiento tenemos faltantes no aleatorios.

MCAR

Si todos los encuestados deciden si contestar la pregunta de ingresos lanzando un dado y negándose a contestar si observa un 6. En el caso MCAR eliminar las observaciones con faltantes no genera un sesgo en la inferencia.

by Estephani Rivera Jaramillo

by Estephani Rivera Jaramillo

Un ejemplo MAS!

- ✓ Cada renglón representa a un empleado
- ✓ IQ es una medición que se le hizo al empleado cuando fue contratado.
- ✓ Performance es una evaluación de desempeño a 6 meses de su contratación.
- ✓ Los datos completos están dados por la segunda columna (Performance).

IQ	Performance
78	9
84	13
84	10
85	8
87	7
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

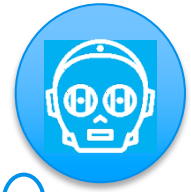
by Estephani Rivera Jaramillo

Un ejemplo MAS!

Ahora consideramos tres escenarios que podrían resultar en datos faltantes de la variable performance:

- ✓ En un accidente, los archivos de desempeño de empleados cuyo apellido que comienza con las letras A-C se pierden.
- ✓ Los empleados con medidas más bajas de IQ no se contratan, así que no se observa su desempeño.
- ✓ Los empleados con peor desempeño (apreciativo, que correlaciona con la evaluación formal) son despedidos antes de los 6 meses de la evaluación de desempeño formal.

by Estephani Rivera Jaramillo



	IQ	performance	performance.MCAR	performance.MAR	performance.MNAR
1	78	9	9	NA	NA
2	84	13	13	NA	13
3	84	10	NA	NA	10
4	85	8	8	NA	NA
5	87	7	7	NA	NA
6	91	7	7	NA	NA
7	92	9	NA	NA	NA
8	94	9	9	9	NA
9	94	11	11	11	11
10	96	7	7	7	NA
11	99	7	7	7	NA
12	105	10	10	10	10
13	105	11	NA	11	11
14	106	15	NA	15	15
15	108	10	10	10	10
16	112	10	10	10	10
17	113	12	12	12	12
18	115	14	14	14	14
19	118	16	16	16	16
20	134	12	NA	12	12

Consideramos cómo es el mecanismo de censura para la variable performance bajo los distintos escenarios:

MCAR: En el primer escenario, la letra del primer apellido no correlaciona con IQ o performance. Los faltantes de performance tienen una probabilidad fija de ocurrir, que no depende de ninguna otra variable. Este es el caso de la columna peformance.MCAR.

MAR: En el segundo escenario, la aparición de performance depende del IQ, que siempre es observado. Pero una vez que condicionamos a IQ, no está relacionada con los valores que toma performance.MAR.

MNAR En el último caso, los faltantes de performance.MNAR dependen tanto de performance y de IQ. Este es el caso MNAR.

by Estephani Rivera Jaramillo

¿ Que es Imputar?

El término "imputación" se refiere al proceso de reemplazar los valores faltantes de instancias en un conjunto de datos incompleto dado con sus valores potenciales o reales de acuerdo con una estrategia específica

Por lo tanto, se han propuesto y empleado varios métodos estadísticos y de aprendizaje automático con el fin de aproximar los valores faltantes en conjuntos de datos incompletos de la forma más eficaz posible.

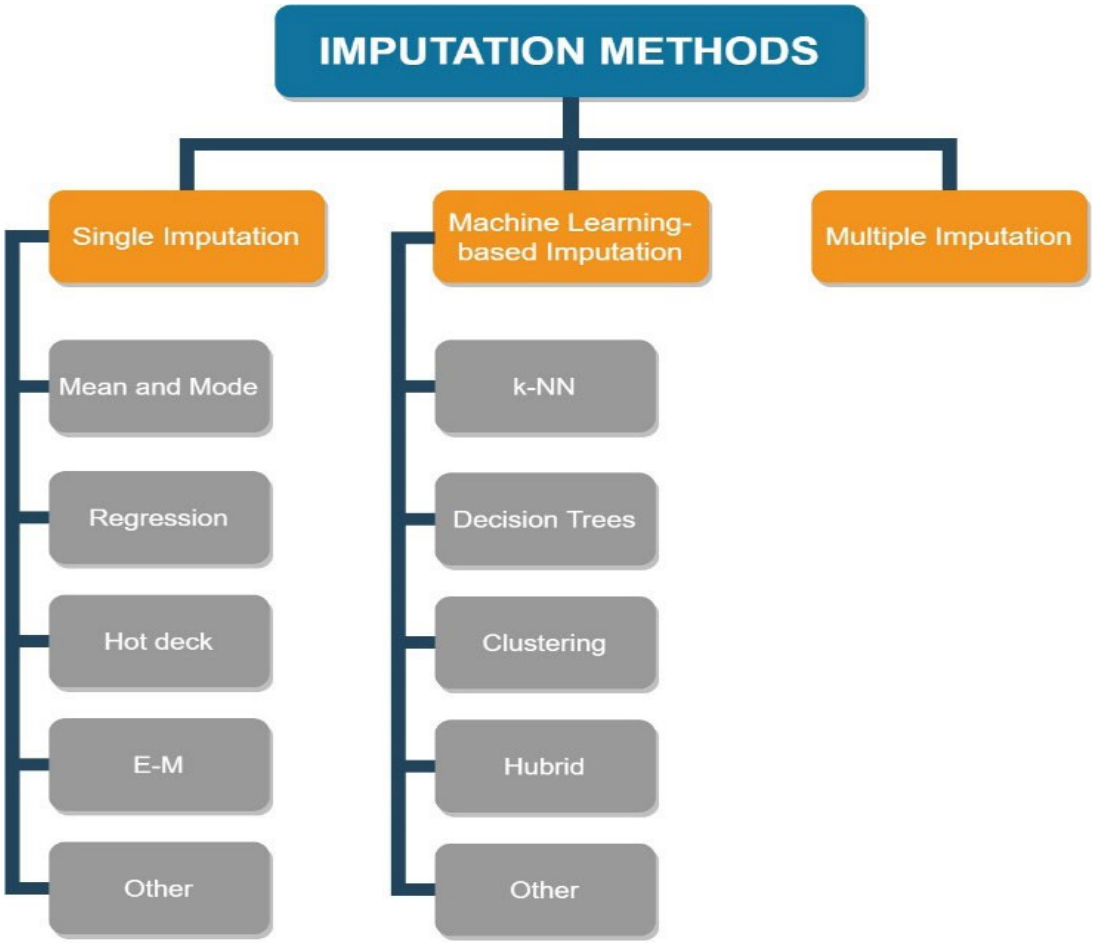
Los métodos de imputación generalmente se clasifican en tres clases principales: simple, basado en aprendizaje automático y múltiple

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

by Estephani Rivera Jaramillo

Métodos de imputación adecuado



by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación simple:

Los métodos de imputación simple reemplazan los datos faltantes por un único valor. Este método conlleva varios problemas aparte de la subestimación de varianzas y covarianzas.

Otra aproximación similar pero más refinada es la imputación de la media condicionada, o también llamada imputación basada en regresión, en lugar de reemplazar cada valor faltante por un valor de media de la variable con observaciones perdidas, se sustituyen los diferentes valores por la media de la variable condicionada a las demás variables que se han observado completamente (sin datos perdidos).

Otros métodos de imputación simple son:

- ✓ Last Observation Carried Forward (LOCF)
- ✓ Sustitución por observaciones relacionadas,
- ✓ Hot Deck
- ✓ Método de variables indicadoras

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación simple:

Imputación por Media, Moda

- Según la media, los valores faltantes de un atributo numérico único se reemplazan con la media aritmética correspondiente de los observados de ese atributo.
- Según la moda completa los valores faltantes de un atributo discreto o categórico con el valor observado con mayor frecuencia.
- En ambos casos los valores faltantes se completan con valores estimados, lo que inevitablemente introduce un sesgo adicional

Imputación por Regresión

- De acuerdo al método, se construye un modelo de regresión a partir de los datos observados de una instancia específica y, posteriormente, se utiliza para predecir los valores de los valores faltantes de esa instancia.
- Se suele aplicar regresión lineal para estimar los valores perdidos de atributos numéricos, mientras que la regresión logística o la regresión logística multinomial se suele utilizar para estimar los valores perdidos de los categóricos.

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación simple:

Imputación Hot Deck

- La imputación hot deck se basa en casos de datos similares pero completos para reemplazar los valores faltantes de los incompletos.
- Una ventaja considerable de la imputación hot deck es que no altera la distribución de los datos observados después del proceso de imputación, a diferencia de la imputación de la media y la moda.

Expectation –Maximization (E-M)

- Es un método iterativo para imputar valores faltantes en conjuntos de datos numéricos incompletos.
- Cada iteración consta de dos pasos: expectativa y maximización.
- El paso de la expectativa se refiere a la estimación de los valores perdidos dados los datos observados, mientras que, en el paso de maximización, los valores estimados actuales se utilizan para maximizar la probabilidad de todos los datos.
- Los valores estimados se actualizan, los dos pasos se repiten hasta la convergencia de la máxima verosimilitud de los datos y las estimaciones finales se utilizan como valores de imputación

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

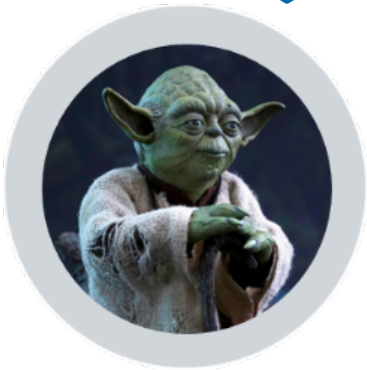
Imputación simple:

Ejemplo N° 1 en Rstudio

Imputación con la media,

data: housing-with-missing-value

Abrir Rstudio debes



by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación simple:

Ejemplo N° 2 en Rstudio

Imputación mediante regresión

data: housing-with-missing-value

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

by Estephani Rivera Jaramillo

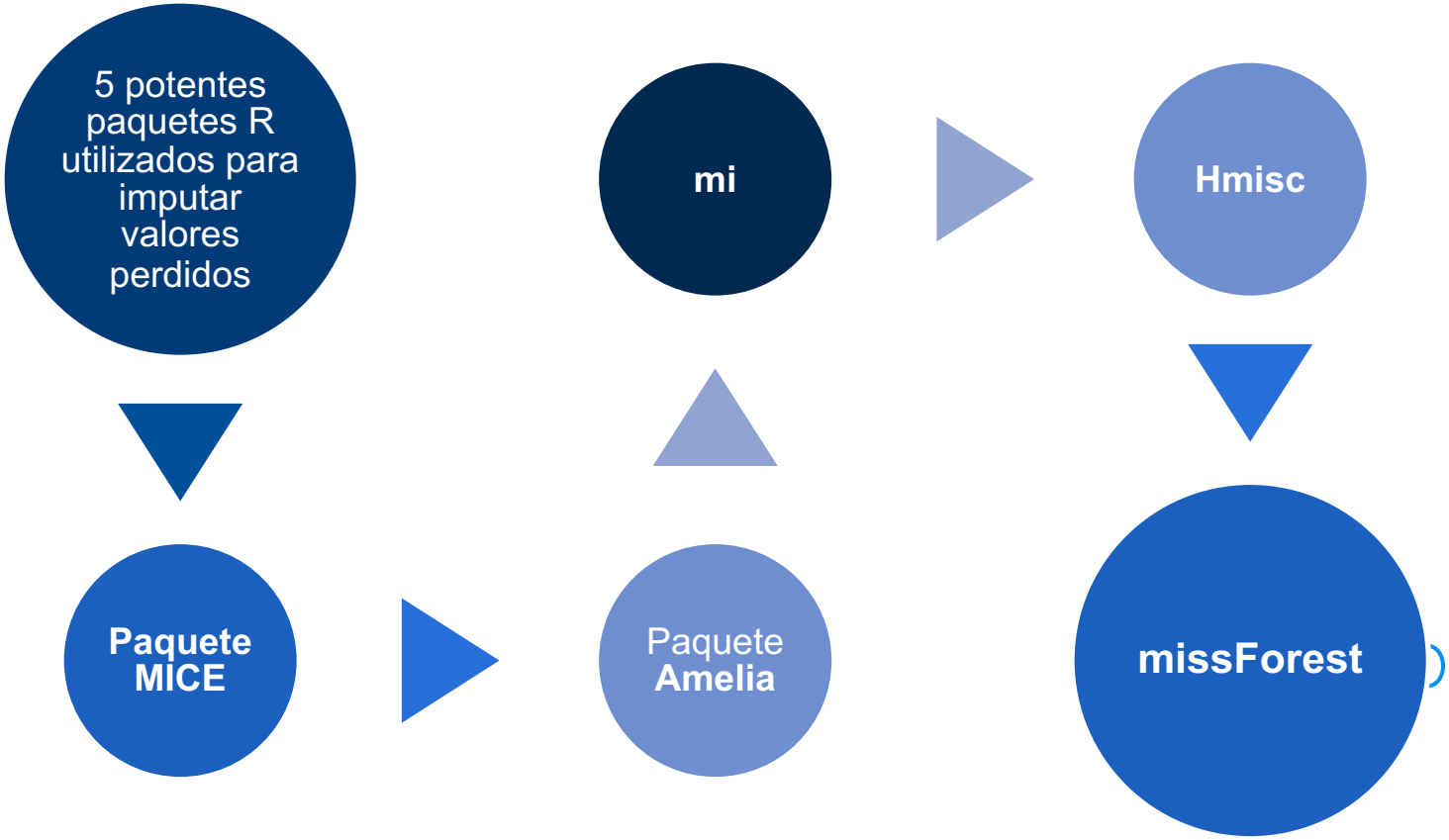
Métodos de imputación adecuado

Donald B. Rubin

Catedrático Emérito de Estadística

by Estephani Rivera Jaramillo

Librerías en Rstudio para Imputación de datos



by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación múltiple:

***MICE*, métodos de imputación**

Método	Descripción	Tipo de dato	*Default*
pmm	Pareamiento por medias predictivas (*predictive mean matching*).	numérico	Sí
norm	Regresión lineal bayesiana.	numérico	
norm.nob	Regresión lineal no bayesiana.	numérico	
mean	Imputación de la media no condicionada.	numérico	
2L.norm	Modelo lineal de dos niveles.	numérico	
logreg	Regresión logística.	categórica, dos niveles	Sí
polyreg	Modelo logístico multinomial.	categórica, > dos niveles	Sí
polr	Modelo logístico ordinal.	ordinal, > dos niveles	Sí
lda	Análisis discriminante lineal.	categórica	
sample	Muestra aleatoria a partir de los datos observados.	cualquiera	
rf	Bosques aleatorios.	cualquiera	

Fuente: Van Buuren y Groothuis-Oudshoorn (2011).

by Estephani Rivera Jaramillo

Métodos de imputación adecuado

Imputación múltiple:

Cuadro 3

***Mi*, tipos de variables y funciones de regresión correspondientes**

Tipo de variable	Descripción	Función de regresión
binary	Variable que contiene dos valores únicos.	*mi.binary*
continuous	Variable numérica continua sin transformación.	*mi.continuous*
count	Variable especificada por el usuario.	*mi.count*
fixed	Variable que contiene un valor único.	*mi.fixed*
log-continuous	Variable continua *log*-escalada.	*mi.continuous*
nonnegative	Variable numérica no negativa con más de cinco valores únicos.	*mi.continuous*
ordered-categorical	Variables que tienen atributo de ordenación.	*mi.polr*
unordered-categorical	Variable factor o carácter.	*mi.categorical*
positive-continuous	Variable positiva con más de cinco valores.	*mi.continuous*
proportion	Variable numérica cuyos valores están entre 0 y 1, sin incluirlos.	*mi.continuous*
predictive-mean-matching	No es un tipo, solo se usa para invocar la función.	*mi.pmm*

Fuente: Su *et al.* (2011).

Métodos de imputación adecuado

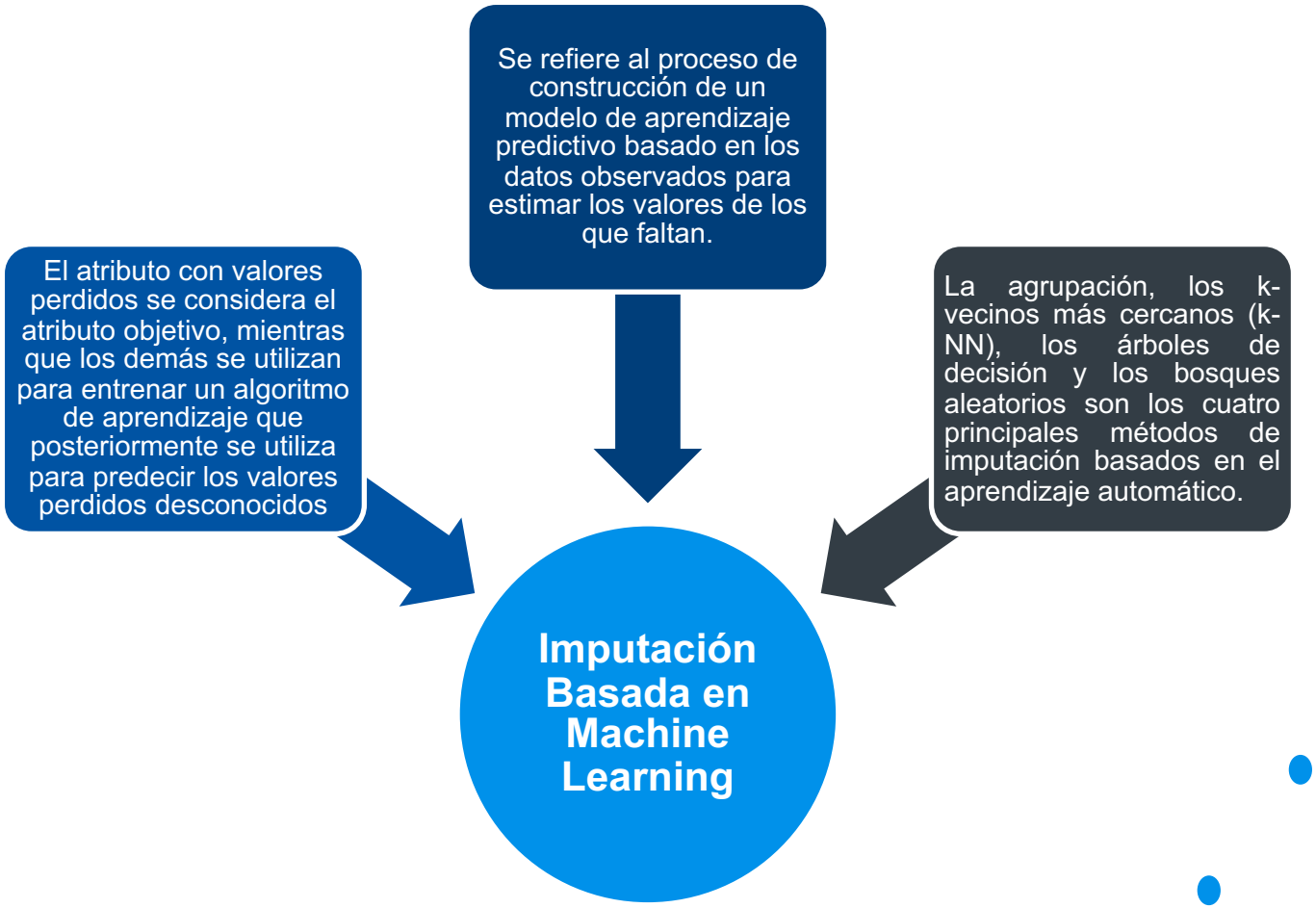
Imputación múltiple

data: housing-with-missing-value

Código (Rstudio)

by Estephani Rivera Jaramillo

Otras estrategias para tratar datos perdidos



by Estephani Rivera Jaramillo

Otras estrategias para tratar datos perdidos

Imputación con KNN

Es un enfoque de imputación basado en similitudes simple y bastante efectivo que se basa en la técnica k-NN.

Para cada valor faltante de una instancia específica, las k instancias más similares se seleccionan de acuerdo con los valores no perdidos compartidos y una medida de similitud predefinida (por ejemplo, distancia euclidiana, distancia de Manhattan o norma de Minkowski).

Código (Rstudio)

data: census

by Estephani Rivera Jaramillo

Un pequeño cuestionario

Kahoot!

by Estephani Rivera Jaramillo

Y recuerda, la estadística y la ciencia de datos nunca será aburrida… porque siempe hay un tema por investigar!

ER

Gracias!

Estephani Rivera Jaramillo

Mis redes :

Linkedin : https://www.linkedin.com/in/estephani-rivera-jaramillo-83224146/
Twitter : https://twitter.com/estephani_jusep
Github : https://github.com/EstephaniRiveraJaramillo
Facebook :https://www.facebook.com/estephani.riverajaramillo
Correo : estephani.rivera.j@gmail.com