**Limpieza de datos**

Los datos faltantes o *missing data* son un problema importante a la hora de realizar análisis de datos, ya que prácticamente todos los métodos estándar en Estadística asumen que la información está completa para todas las variables incluidas en el análisis. Además, la falta de datos es un problema frecuente en todos los tipos de estudio, sin importar que el diseño sea muy estricto o que los investigadores traten de prevenirlo.

Los *missing data* se definen como los valores que no están disponibles pero que de otra manera serían significativos para el análisis si se pudieran observar. A pesar de la falta de datos, el objetivo es hacer inferencias sobre la población objetivo de la muestra completa, aunque no hay ningún método universal para tratar estos datos. La causa de ello, es que, aunque se suela conocer la selección de los sujetos de un estudio, se suele desconocer la causa de que sus observaciones se hayan perdido. Esto obliga a que haya que hacer suposiciones adicionales sobre los datos para realizar el análisis, pero las mismas no se pueden validar únicamente a partir de los datos observados.

Hay que realizar la distinción entre mecanismos y patrones de pérdida de datos. Los patrones describen los valores que son observados y los que son perdidos, mientras que los mecanismos describen el proceso mediante el cual se produce la pérdida de los valores. En prácticamente cualquier referencia sobre missing data se distinguen tres tipos de mecanismos de pérdida de datos:

* MCAR (Missing completely at random). Este mecanismo requiere que la probabilidad de pérdida de los datos de una variable Y no dependa de otras variables medidas, X, ni de la propia variable Y. Entonces, los datos observados de la variable Y son una muestra aleatoria simple de los datos completos que se deberían haber observado en dicha variable. Esto no implica que dicho mecanismo de pérdida no se pueda relacionar con el mecanismo de pérdida de los datos de otra variable. Los datos de este mecanismo son los que se tratan con mayor facilidad.
* MAR (Missing at random). El mecanismo MAR tiene la misma base que el MCAR, pero es menos restrictivo. La diferencia entre ambos mecanismos consiste en que con MAR los datos que faltan no son una muestra aleatoria simple de todos los valores. En este mecanismo la probabilidad de que se den datos faltantes en una variable Y depende de una o más de las demás variables medidas en el análisis, X, pero no depende de los valores de la propia variable Y.
* MNAR (Missing not at random). En este mecanismo la probabilidad de pérdida de datos en una variable Y depende de los valores de dicha variable Y, y también puede depender de los valores observados de las demás variables.

En la práctica, realizar estas distinciones entre los mecanismos de pérdida de datos no suele ser posible, sobre todo con MAR y MNAR, y salvo que haya información adicional, el tipo de mecanismo solamente se puede razonar o especular.

Para comprender mejor lo que ocurre con la pérdida de datos en función del mecanismo, se va a poner un ejemplo extraído de. En dicho ejemplo se considera un modelo de regresión lineal simple para modelar la relación entre las variables x e y, siguiendo la ecuación y = x+e generando una muestra independiente de tamaño 300 y conociendo todos los valores de x e y. Los datos de x van a estar completos para cualquiera de las situaciones que se plantee. Se van a generar otras tres muestras para modelar los diferentes mecanismos de pérdida de datos en y. En la primera de ellas (MCAR), se pierden algunos datos de manera aleatoria. En la segunda (MAR), se pierden los datos en los que x < 0. En la tercera (MNAR), se pierden los datos en los que y < 0 y x < 0. Se muestran los resultados obtenidos de en la Figura 1.

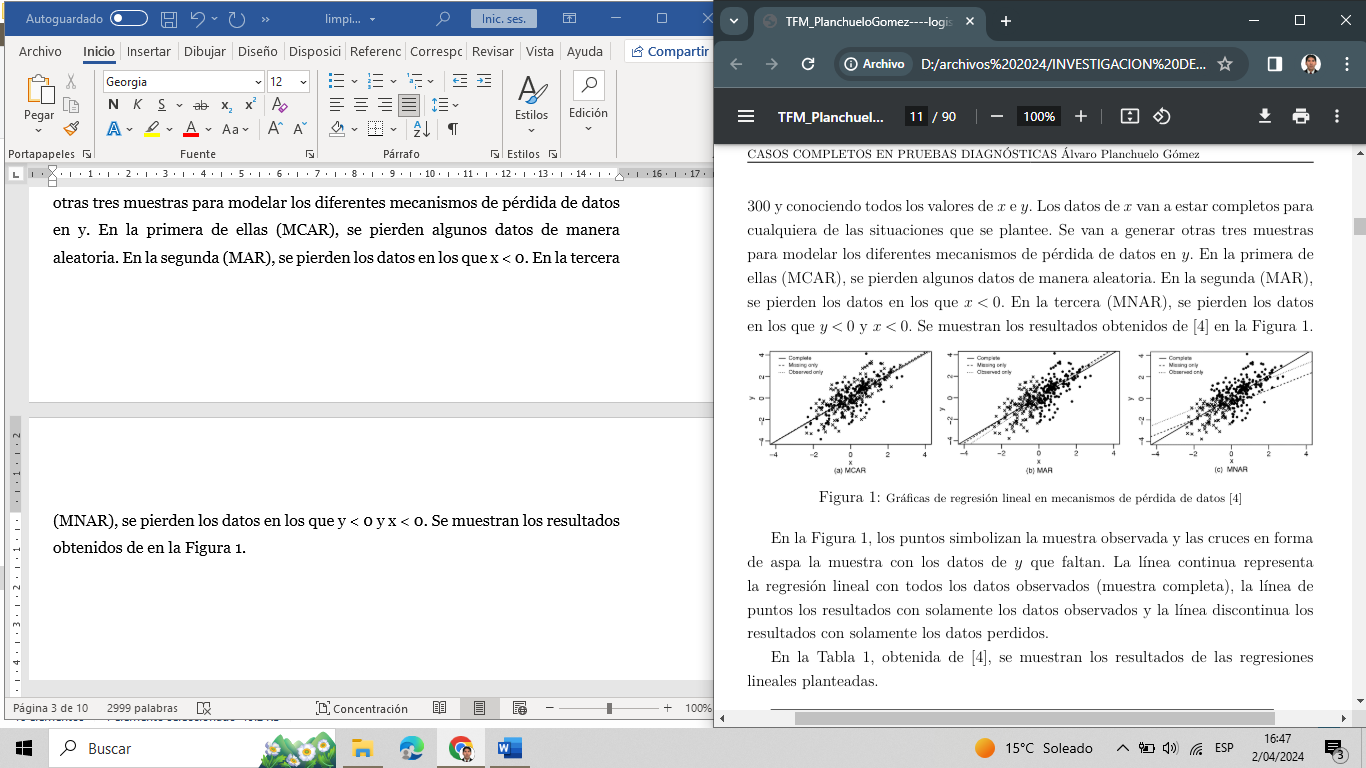


Figura 1: Gráficas de regresión lineal en mecanismos de pérdida de datos

En la Figura 1, los puntos simbolizan la muestra observada y las cruces en forma de aspa la muestra con los datos de y que faltan. La línea continua representa la regresión lineal con todos los datos observados (muestra completa), la línea de puntos los resultados con solamente los datos observados y la línea discontinua los resultados con solamente los datos perdidos.

En la Tabla 1, se muestran los resultados de las regresiones lineales planteadas.

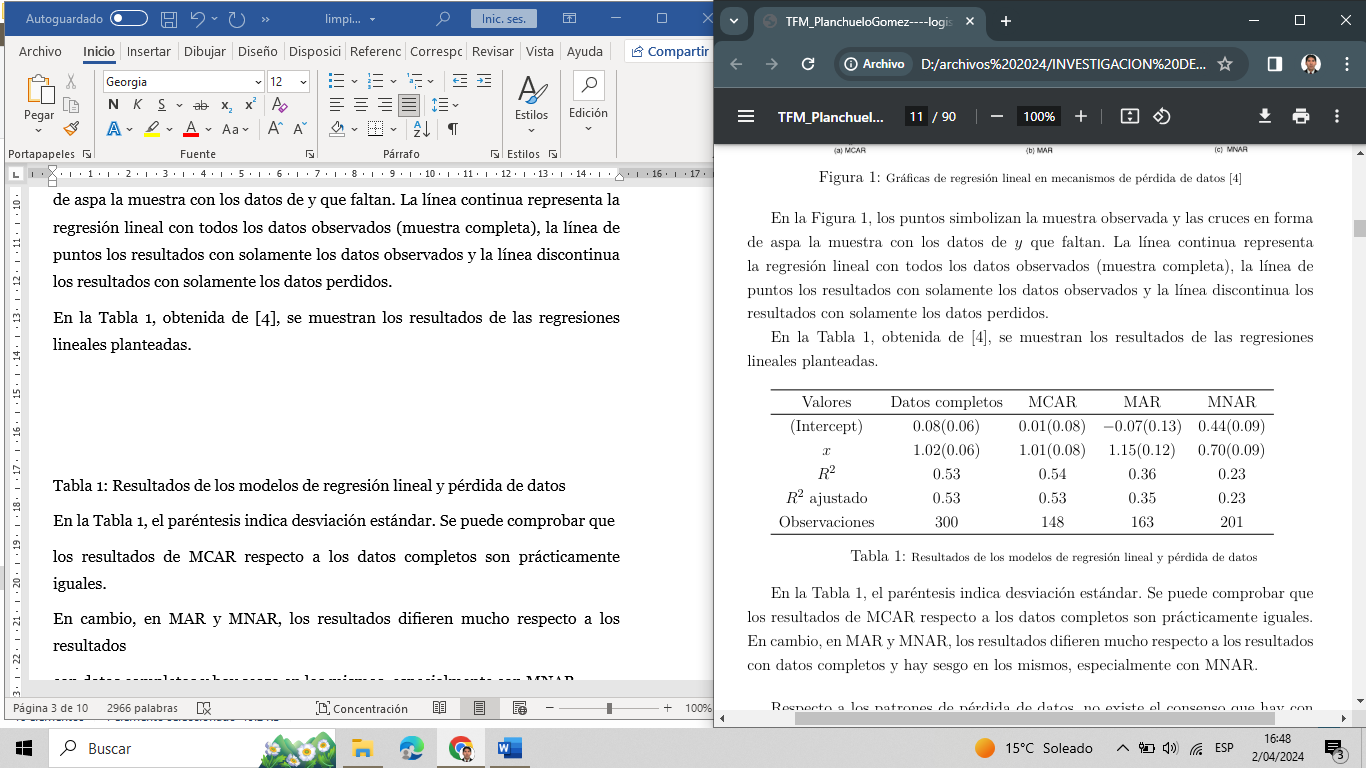


Tabla 1: Resultados de los modelos de regresión lineal y pérdida de datos

En la Tabla 1, el paréntesis indica desviación estándar. Se puede comprobar que los resultados de MCAR respecto a los datos completos son prácticamente iguales. En cambio, en MAR y MNAR, los resultados difieren mucho respecto a los resultados con datos completos y hay sesgo en los mismos, especialmente con MNAR.

Respecto a los patrones de pérdida de datos, no existe el consenso que hay con los mecanismos de pérdida de datos, así que se van a exponer los patrones indicados y explicados.

El primer patrón es el de una variable. Solamente faltan datos de una única variable mientras que el resto está completo.

Otro patrón es el monótono, que es muy común en estudios longitudinales. En este patrón, faltan los datos de varias variables de una parte muy localizada de la estructura de datos; en estudios longitudinales, por ejemplo, faltarían los datos de la parte final del estudio.

El patrón file-matching se da cuando dos o más variables no son medidas conjuntamente. Un ejemplo podría ser que no se pueden realizar dos pruebas molestas para los pacientes el mismo día y que de ambas pruebas se extraigan datos que se incluyen en un estudio. Es común en ensayos clínicos aleatorios cruzados en los que los resultados potenciales de un tratamiento no se pueden medir a la vez que los del placebo, siguiendo este caso una estructura del modelo causal de Rubin.

Y por último, en el patrón general no hay ninguna estructura de falta de datos que se pueda determinar con claridad. Es común en cuestionarios que no son completados por pacientes.

**¿Qué es Limpieza de Datos?**

Uno de los aspectos más importantes al trabajar con datos, consiste en el pre-procesamiento de datos.

La limpieza de datos es un paso necesario antes de realizar cualquier tarea relevante en análisis/ciencia de datos, desde elaborar un simple análisis gráfico, hasta implementar un modelo de aprendizaje automático.

En la práctica es probablemente sea la tarea que más tiempo te tomará hacer.

Denominado en Inglés Data Cleaning o Data Cleansing, este proceso involucra detectar, eliminar, corregir o transformar cualquier anomalía, perturbación o irrelevancia de los datos.

La limpieza de datos es el proceso de identificar, corregir o eliminar información incorrecta, incompleta, duplicada o irrelevante en un conjunto de datos. Este proceso es fundamental para garantizar la calidad y la precisión de los datos antes de su uso en análisis o aplicaciones.

La importancia de estudiar y tratar los datos faltantes radica en que, en caso de no hacerlo, la ausencia de unas pocas observaciones en algunas variables puede reducir de manera drástica el tamaño de la muestra, afectando de manera muy significativa a los intervalos de confianza, a la potencia del estudio y provocando el posible sesgo de los parámetros estimados.

**Razones para efectuar Limpieza de Datos**

La limpieza de datos suele involucrar varias tareas, como la eliminación de duplicados, la corrección de errores tipográficos, la eliminación de valores atípicos, la normalización de los datos y la eliminación de datos irrelevantes o incompletos. En algunos casos, puede ser necesario realizar una limpieza manual, especialmente cuando se trata de datos no estructurados.

Este proceso ayuda a obtener resultados confiables, un error peligroso a la hora de realizar un análisis es abstraerse en exceso en el modelo a implementar, y no empezar por lo más simple. Tu fuente de información y su tratamiento. Los datos en el mundo real rara vez son homogéneos y directamente intuitivos. Asimismo, realizar este proceso puede reducir posibles sesgos en el análisis. He aquí varias de las razones más comunes:

**Remover Outliers**: Por outlier nos referimos a cualquier dato que difiere significativamente del promedio o rango observado del resto de nuestros datos. Aunque los outliers son mucho más fáciles de apreciar gráficamente, también pueden detectarse a primera vista en un conjunto de datos, por ejemplo, encontrar la variable edad de una persona en 480. Inmediatamente nos indica que es un error de entrada de datos, conocido en inglés como "data entry error", probablemente la edad sea 48. Eliminación de valores atípicos: Esta técnica consiste en identificar y eliminar los valores que están muy por encima o por debajo del rango normal, lo cual puede ser un error o simplemente un valor atípico que no sigue la tendencia de los demás datos.

**Trabajar datos Nulos:**No se ha introducido información alguna en alguno de los elementos de la variable. Estos pueden deberse a falta de información o bien a no haber podido estimar la variable, un error de entrada, o bien hubo un fenómeno externo que impidió la recolección de datos. Cualquiera sea la razón es importante estar consciente de esta situación, puesto que conociendo la localización y el tipo de datos que se encuentran nulos, podremos extrapolarlos, efectuar una transformación o bien, al menos conocer las limitaciones de nuestra información. Este fenómeno también puede ser conocido como "Missing Data". Imputación de datos faltantes: Esta técnica se utiliza para llenar los valores faltantes en los datos. Esto se puede hacer mediante la asignación de valores promedio o la interpolación de los valores de otros registros similares.

**Eliminar o transformar datos erróneos e irrelevantes:** Estos representan información fuera de contexto en nuestra base de datos, por ejemplo, tener la variable precio de venta en números negativos. Esta técnica consiste en eliminar las variables que no son necesarias para el análisis.

Corrección de errores tipográficos: En esta técnica se buscan y corrigen errores de escritura o de formato que puedan estar presentes en los datos.

Normalización de los datos: Esta técnica consiste en transformar los datos a un formato común, por ejemplo, cambiar todas las fechas a un formato específico para facilitar el análisis.

Eliminación de datos duplicados: Se busca y elimina los registros que son iguales o muy similares.

Análisis de valores extremos: Identificar valores que son muy extremos y analizar la causa.

Establecimiento de reglas de validación: Se pueden establecer reglas para garantizar la calidad de los datos al momento de su ingreso.

**Metodología para el tratamiento de datos faltantes (missing)**

A continuación, se van a explicar los diferentes métodos de tratamiento de datos

faltantes a partir de las divisiones realizadas.

**Métodos que eliminan observaciones**

Estos métodos consisten en eliminar la información de cualquier individuo en el que haya pérdida de datos, por lo menos, en una de las variables de estudio. Estos métodos son los más sencillos de implementar y los que menos recursos computacionales necesitan. Además, estos métodos reducen el tamaño de la muestra, lo que disminuye la eficiencia y conlleva que los errores estándar sean mayores en los parámetros de interés .

Los métodos más importantes son los de casos completos (Complete-Case), que

pueden ser ponderados o no ponderados.

En el caso de no hacer ponderación, simplemente consiste en eliminar las observaciones en las que falte algún dato de alguna de las variables. La desventaja de este método es el sesgo que provoca en datos que no son MCAR, como ya se comentó a raíz de los resultados que se vieron en la Tabla 1, y la pérdida de precisión al tener un tamaño muestral menor. Se debería optar por este método en caso de que la pérdida de precisión y el sesgo sean despreciables o muy pequeños, pero son difíciles de cuantificar debido a que dependen de las diferencias en los parámetros de interés en los casos completos e incompletos. Aunque no haya una regla universal establecida respecto a un porcentaje de fallo para aplicar este método de manera apropiada, llegaron a la conclusión de que este método era más recomendable que la imputación múltiple (posteriormente se desarrollará) cuando el porcentaje de datos faltantes era aproximadamente del 10% o menor.

El método ponderado se emplea especialmente con encuestas. En caso de que falten datos solamente en una variable, se crea un modelo para predecir la ausencia de respuesta de la variable como una función del resto de variables disponibles y el inverso de las probabilidades predichas se puede utilizar como peso para hacer los casos completos más representativos. En caso de que falten datos en dos o más variables, el método es más complicado y da problemas con los errores estándar cuando las probabilidades predichas son próximas a 0.

El último método consiste en eliminar las variables con una gran cantidad de datos faltantes, pero este método elimina información (variables) que puede ser muy importante para realizar modelos provocando sesgo y pérdida de precisión. Es un método que normalmente no se plantea aplicar.

**Métodos que imputan los datos faltantes**

La idea de estos métodos es que la información de los valores que faltan se puede extraer de las variables observadas. Se van a distinguir dos clases de métodos en esta subsección: imputación simple e imputación múltiple.

* Imputación simple

Los métodos de imputación simple reemplazan los datos faltantes por un único valor. En general, para cualquier enfoque de este tipo de métodos, se subestiman los errores estándar de las variables en las que faltan datos originalmente al tratar a los datos obtenidos como una muestra completa y no tener en cuenta las consecuencias del método.

El método más sencillo consiste simplemente en sustituir los datos faltantes de una variable por la media de los valores observados en dicha variable. En caso de que reemplazar por la media no tuviera sentido (por ejemplo, con el sexo), se podría reemplazar por la moda. Este método conlleva varios problemas aparte de la subestimación de varianzas y covarianzas. Uno de ellos es que se modifican las relaciones entre las variables debido a que las correlaciones tienden a ser nulas. El sesgo no se elimina si los datos son MAR o MNAR.

Se incrementa la potencia estadística de manera espuria, así como la tasa del error de tipo I. 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 y Método de variables indicadoras, que pueden verse con detalle en.

Viendo los problemas de subestimación del error de los métodos anteriores, una estrategia para compensar dicho efecto puede ser añadir error para una mejor estimación de las varianzas y covarianzas y además poder obtener estimaciones precisas de los datos faltantes. El error añadido serviría para captar las diferentes fuentes de incertidumbre que tienen influencia en los datos reales. Los problemas de este método son que las estimaciones son más imprecisas que en otros métodos ya comentados y que sigue habiendo subestimación de la variabilidad de los datos debido a que la varianza y las covarianzas de una muestra son parámetros que también deben ser estimados respecto a la población real.

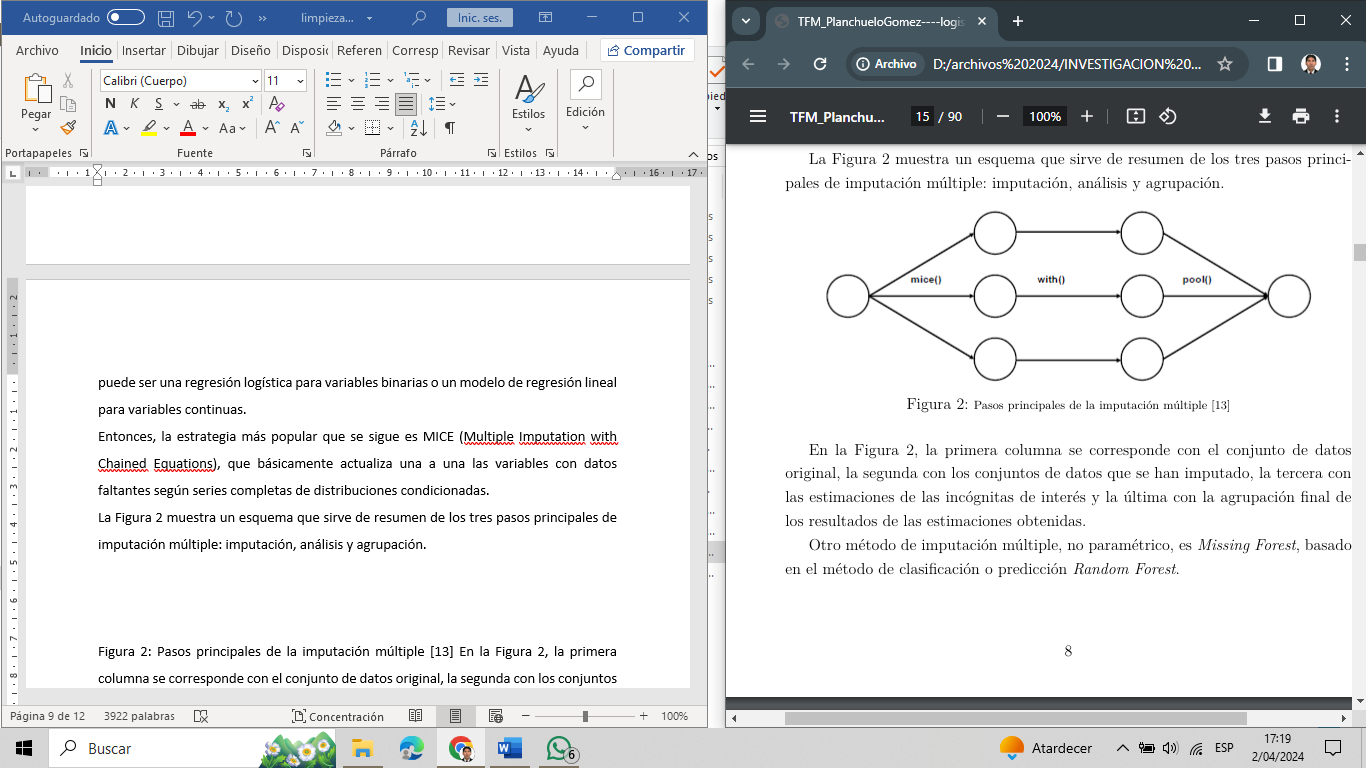
* Imputación múltiple

Los métodos de imputación múltiple consisten en reemplazar cada valor perdido por un conjunto de m valores obteniéndose así m conjuntos completos de datos, lo que da lugar a m estimaciones con sus respectivas varianzas o errores estándar [4]. Entonces, se combinan las estimaciones dando lugar a estimaciones e intervalos de confianza que incorporan la incertidumbre causada por la pérdida de datos [4]. Estos métodos permiten minimizar el sesgo o la pérdida de potencia estadística causada por la pérdida de datos con datos MCAR o MAR.

A priori, una desventaja que podría ser importante de estos métodos respecto a los de imputación simple es que son más difíciles de implementar, lo que además podría conllevar mayor gasto computacional. Para disminuir la dificultad que puede suponer la implementación de métodos complicados, ya existen paquetes en softwares como R para facilitar dicho trabajo, destacando, por ejemplo, el paquete mice. En caso de que solamente haya una variable en la que haya pérdida de datos, el modelo de imputación puede ser una regresión logística para variables binarias o un modelo de regresión lineal para variables continuas.

Entonces, la estrategia más popular que se sigue es MICE (Multiple Imputation with Chained Equations), que básicamente actualiza una a una las variables con datos faltantes según series completas de distribuciones condicionadas.

La Figura 2 muestra un esquema que sirve de resumen de los tres pasos principales de imputación múltiple: imputación, análisis y agrupación.

Figura 2: Pasos principales de la imputación múltiple [13]

En la Figura 2, la primera columna se corresponde con el conjunto de datos original, la segunda con los conjuntos de datos que se han imputado, la tercera con las estimaciones de las incógnitas de interés y la última con la agrupación final de los resultados de las estimaciones obtenidas. Otro método de imputación múltiple, no paramétrico, es Missing Forest, basado en el método de clasificación o predicción Random Forest

**Modelos predictivos/estadísticos que imputan los datos que faltan**

Esto debería hacerse junto con algún tipo de esquema de validación cruzada para evitar fugas. Esto puede ser muy efectivo y puede ayudar con el modelo final. Hay muchas opciones para tal modelo predictivo, incluyendo la red neuronal. Aquí estoy enumerando algunas que son muy populares.

Imputación por media: Este método consiste en reemplazar los valores faltantes con la media de los valores observados en la misma variable. Este método es adecuado cuando la distribución de los datos es simétrica y no hay valores extremos.

Imputación por moda: Este método se utiliza para variables categóricas y consiste en reemplazar los valores faltantes con la moda o valor más frecuente en la misma variable.

Regresión lineal: La regresión lineal se puede utilizar para imputar datos faltantes cuando hay una relación lineal entre las variables con valores faltantes y otras variables del conjunto de datos.

K-vecinos más cercanos: Este método se basa en encontrar los K registros más cercanos al registro con valor faltante y utiliza sus valores para imputar el valor faltante.

Análisis de componentes principales (PCA): El análisis de componentes principales se puede utilizar para imputar datos faltantes al reducir la dimensionalidad del conjunto de datos y reconstruir los valores faltantes en base a los componentes principales.

Árboles de decisión: Los árboles de decisión se pueden utilizar para imputar valores faltantes en variables categóricas, construyendo un árbol para predecir el valor faltante en función de las demás variables.

Bosque aleatorio

El bosque aleatorio es un método de imputación no paramétrico aplicable a varios tipos de variables que funciona bien tanto con datos que faltan al azar como con datos que no faltan al azar. El bosque aleatorio utiliza múltiples árboles de decisión para estimar los valores perdidos y produce estimaciones de error de imputación OOB (out of bag). Una advertencia es que el bosque aleatorio funciona mejor con datasets grandes y el uso del bosque aleatorio datasets pequeños corre el riesgo de generar sobreajuste.

Máxima probabilidad

Hay varias estrategias que utilizan el método de máxima probabilidad para manejar los datos faltantes. En ellas, la suposición de que los datos observados son una muestra extraída de una distribución normal multivariante es relativamente fácil de entender. Una vez los parámetros son estimados usando los datos disponibles, los datos faltantes se estiman sobre la base de los parámetros que acaban de ser estimados.

Expectativa-Maximización

La maximización de la expectativa (Expectation Maximization) es un tipo de método de máxima probabilidad que puede utilizarse para crear un nuevo dataset, en el que todos los valores que faltan son imputados con valores estimados por los métodos de máxima probabilidad. Este enfoque comienza con el paso de expectativas, durante el cual se estiman los parámetros (por ejemplo, varianzas, covarianzas y medias), tal vez utilizando la eliminación de la lista.

Esas estimaciones se utilizan luego para crear una ecuación de regresión para predecir los datos que faltan. En el paso de maximización se utilizan esas ecuaciones para completar los datos que faltan. El paso de expectativa se repite luego con los nuevos parámetros, donde se determinan las nuevas ecuaciones de regresión para "rellenar" los datos que faltan. Los pasos de expectativa y maximización se repiten hasta que el sistema se estabiliza.

Análisis de sensibilidad

El análisis de sensibilidad se define como el estudio de cómo la incertidumbre en la salida de un modelo puede asignarse a las diferentes fuentes de incertidumbre en sus entradas. Al analizar los datos que faltan, se hacen supuestos adicionales sobre las razones de los datos que faltan, y estos supuestos suelen ser aplicables al análisis primario. Sin embargo, los supuestos no pueden validarse definitivamente para su corrección. Por lo tanto, el Consejo Nacional de Investigación ha propuesto que se realice el análisis de sensibilidad para evaluar la solidez de los resultados frente a las desviaciones de la suposición de datos faltantes al azar

Algoritmos que funcionan con valores faltantes

No todos los algoritmos fallan cuando faltan datos. Hay algoritmos que son robustos a los datos faltantes, como k-Vecinos más cercanos que puede ignorar una columna de una medida de distancia cuando falta un valor. También hay algoritmos que pueden utilizar el valor que falta como un valor único y diferente cuando se construye el modelo de predicción, como los árboles de clasificación y regresión. Un algoritmo como XGBoost toma en consideración cualquier dato que falte. Si su imputación no funciona bien, pruebe un modelo que sea robusto con datos faltantes

**Outliers**

Los outliers o valores atípicos son valores extremos o fuera de lo común en un conjunto de datos. Estos valores pueden ser el resultado de errores en la medición o la entrada de datos, o pueden indicar la presencia de valores verdaderamente inusuales en el conjunto de datos. Es importante detectar y tratar los outliers para evitar que distorsionen las estadísticas y los modelos que se basan en los datos.

Existen diversas técnicas para detectar outliers, como el método de las tres desviaciones estándar, el método de Tukey, el análisis de componentes principales, entre otros. Una vez detectados, los outliers pueden ser tratados de diferentes maneras, dependiendo del contexto y de la naturaleza de los datos. Algunas técnicas para tratar los outliers incluyen:

Eliminar los valores atípicos: Este enfoque consiste en simplemente eliminar los valores atípicos del conjunto de datos. Esto puede ser apropiado si se sospecha que los outliers son el resultado de errores o mediciones inexactas.

Transformar los datos: A veces, transformar los datos a una escala diferente puede ayudar a reducir el efecto de los outliers. Por ejemplo, si los datos están sesgados hacia valores altos, se pueden tomar logaritmos para "aplanar" la distribución y reducir la influencia de los valores atípicos.

Tratar los valores atípicos como valores perdidos: En algunos casos, puede ser apropiado tratar los valores atípicos como valores perdidos y utilizar técnicas de imputación de datos para estimar los valores faltantes.

Utilizar métodos robustos: En lugar de eliminar o transformar los datos, los métodos robustos están diseñados para minimizar el efecto de los valores atípicos en los resultados. Por ejemplo, los métodos de regresión robustos pueden ser menos sensibles a los outliers que los métodos de regresión estándar.