Metodología De Evaluación Del Desempeño De Métodos De Imputación Mediante Una Métrica Tradicional Complementada Con Un Nuevo Indicador

Carlos Roberto Primorac,
Departamento de Informática, Universidad Nacional del Nordeste, Argentina
David Luís La Red Martínez,
Mirta Eve Giovannini,
Universidad Tecnológica Nacional,
Facultad Regional Resistencia, Argentina

Doi:10.19044/esj.2020.v16n18p61 URL:http://dx.doi.org/10.19044/esj.2020.v16n18p61

Resumen

Los valores faltantes (MV: Missing Values), valores no observados en el conjunto de datos (dataset), constituyen un obstáculo común que enfrentan investigadores en contextos del mundo real. Las técnicas de imputación de datos permiten estimarlos utilizando diferentes algoritmos, mediante los cuales se puede imputar una característica importante para una instancia en particular. La mayoría de los artículos publicados en este campo tratan sobre nuevos métodos de imputación, sin embargo, pocos estudios abordan la evaluación de los métodos existentes con el objeto de aportar pautas más adecuadas para la imputación de datos. El objetivo de este trabajo es mostrar una metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador, basado en el promedio normalizado de la raíz cuadrada del error cuadrático medio (RMSE: Root Mean Squared Error). A partir de un conjunto de datos completo, se generaron 63 conjuntos de datos con MV. Estos fueron imputados mediante los métodos de imputación por medias, k-NN, k-Means y hot-deck. El desempeño de los métodos de imputación fue evaluado utilizando la métrica tradicional complementada con un nuevo indicador propuesto. Los resultados muestran que el error para el método de imputación k-Means es el más bajo considerando la totalidad de conjuntos de datos. El entorno de trabajo desarrollado para realizar los experimentos de amputación y posterior imputación resultó apropiado y permite la incorporación a futuro de otros mecanismos de amputación y otros métodos de imputación, siendo

parte esencial de la metodología propuesta.

Parablas clave: Amputación De Datos, Imputación De Datos, Evaluación De Desempeño De Métodos De Imputación

Methodology for Evaluating the Performance of Imputation Methods Using a Traditional Metric Complemented with a New Indicator

Carlos Roberto Primorac,
Departamento de Informática, Universidad Nacional del Nordeste, Argentina
David Luís La Red Martínez,
Mirta Eve Giovannini,
Universidad Tecnológica Nacional,
Facultad Regional Resistencia, Argentina

Abstract

Missing Values (MV), values not observed in the dataset, constitute a common obstacle faced by researchers in real-world contexts. Data imputation techniques allow estimating them using different algorithms, through which an important characteristic can be imputed to a particular instance. Most of the articles published in this field deal with new imputation methods, however, few studies address the evaluation of existing methods in order to provide more appropriate guidelines for imputation of data. The objective of this work is to show a methodology for evaluating the performance of imputation methods using a traditional metric complemented with a new indicator, based on the normalized average of the Root Mean Squared Error (RMSE). From a complete data set, 63 data sets were generated with MV. These were imputed using the methods of imputation by means, k-NN, k-Means and hot-deck. The performance of the imputation methods was evaluated using the traditional metric complemented with a new proposed indicator. The results show that the error for the k-Means imputation method is the lowest considering all data sets. The work environment developed to perform the amputation and subsequent imputation experiments was appropriate and allows the incorporation of other amputation mechanisms and other imputation methods in the future, being an essential part of the proposed methodology.

Keywords: Data Amputation, Data Imputation, Performance Evaluation Of Imputation Methods

Introducción

Los valores faltantes (MV: Missing Values), valores no observados en el conjunto de datos (dataset), constituyen un obstáculo común que enfrentan investigadores en contextos del mundo real. Ocurren en una variedad de dominios, por diferentes razones y, sin importar cuáles sean, tienen severas implicaciones en el proceso de extracción del conocimiento (Santos et al., 2019). La presencia de estas imperfecciones generalmente requiere de una fase de pre-procesamiento en la cual, con el fin de que resulten útiles y suficientemente claros, los datos se deben preparar y limpiar (Luengo et al., 2012).

En la literatura se proponen dos enfoques generales para tratar con los MV (Farhangfar et al., 2007). En el caso más simple, sencillamente se omiten. Una segunda alternativa consiste en utilizar técnicas de imputación y, a partir de los datos completos, estimarlos utilizando diferentes algoritmos, mediante los cuales se puede imputar una característica importante para una instancia en particular (Aljuaid & Sasi, 2017).

El enfoque clásico para la evaluación del desempeño de los métodos de imputación sigue cuatro pasos (Santos et al., 2019) (Schouten et al., 2018):

1. Recopilación o simulación de conjuntos de datos completos.
2. Generación de conjuntos de datos con MV a partir de los conjuntos de datos completos.
3. Imputación de datos utilizando diferentes estrategias.
4. Evaluación del desempeño de los algoritmos de imputación en términos de la calidad de la estimación.

La mayoría de los artículos publicados en este campo tratan sobre nuevos métodos de imputación, sin embargo, pocos estudios abordan la evaluación de los métodos existentes con el objeto de aportar pautas más adecuadas para la imputación de datos (Jadhav et al., 2019).

El objetivo de este trabajo es mostrar una metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador, basado en el promedio normalizado de la raíz cuadrada del error cuadrático medio (RMSE: Root Mean Squared Error).

El artículo se organiza de la siguiente manera: en la sección 2 se comentan los conceptos fundamentales acerca de los MV y se mencionan las estrategias de imputación generales para enfrentarse a estos. En la sección 3 se describen el conjunto de datos utilizados, los procedimientos para la generación de MV, los métodos de imputación empleados y la medida de

desempeño utilizada, como así también el software utilizado y el diseño de los experimentos efectuados. En la sección 4 se comentan y analizan los resultados obtenidos, considerándose la métrica de evaluación de desempeño propuesta de los métodos de imputación utilizados. Finalmente, en la sección 5 se exponen las conclusiones y líneas futuras de trabajo.

Conocimientos previos

El modelo de datos

Un conjunto de datos multivariados completo Y puede representarse mediante una matriz de datos rectangular de $n \times p$, donde n es el número de casos y p es el número de variables (Figura 1). El conjunto de datos completo es hipotético, puesto que contiene valores no observados, denotados por el símbolo "?", que pueden aparecer en diferentes proporciones, patrones y de acuerdo con algún mecanismo. En consecuencia, se puede considerar como formado por dos componentes $Y = (Y_{obs}, Y_{mis})$.

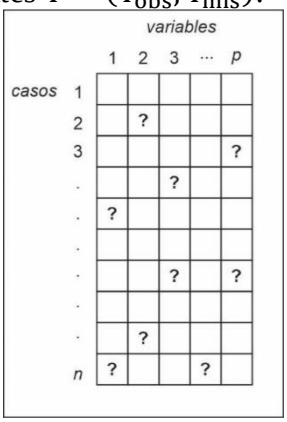

Figura 1. Conjunto de datos hipotéticamente completo (J. L. Schafer, 1997).

Se define la matriz indicadora de respuesta R como una matriz de ceros y unos de $n \times p$. Los elementos de Y y R se denotan por y_{ij} y r_{ij} respectivamente, donde $1 \leq i \leq n$ y $1 \leq j \leq p$. Si y_{ij} es un valor observado, entonces $r_{ij} = 1$, si y_{ij} es un valor no observado, entonces $r_{ij} = 0$.

Los valores observados se denotan colectivamente por Y_{obs}, los valores no observados por Y_{mis} y contienen todos los elementos y_{ij} para los cuales $r_{ij} = 0$. Así, R indica la localización de los valores faltantes en el conjunto de datos (Van Buuren, 2012).

Mecanismos de MV

Rubin (1976), clasificó los problemas de MV en tres categorías: falta completamente aleatoria (MCAR: Missing Completely At Random), falta aleatoria (MAR: Missing At Random) y falta no aleatoria (MNAR: Missing

Not At Random). En esta propuesta, cada dato puntual tiene una probabilidad de faltar. El proceso que gobierna estas probabilidades se denomina mecanismo de datos faltantes o de respuesta. El modelo para el proceso se conoce como modelo de datos faltantes o de respuesta (Van Buuren, 2012).

La definición formal de estos mecanismos, implica diferentes distribuciones de probabilidad para la matriz indicadora de respuesta, éstas describen esencialmente las relaciones entre ésta y los datos. En la práctica, en general no es posible determinar con certeza los parámetros de estas distribuciones. Sin embargo, no es importante conocer en detalle estos parámetros, solamente es necesario comprender si existe una relación entre R y los componentes de Y (Enders, 2010).

Rubin (1976) establece que los valores son MCAR si $P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|\psi)$. Así, la probabilidad de $R = 0$, depende únicamente de algunos parámetros ψ. En otras palabras, la probabilidad de que falte un valor en Y es completamente aleatoria.

Se dice que los valores son MAR si $P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|Y_{obs}, \psi)$. Es decir, la probabilidad de $R = 0$, depende de valores observados a través de algunos parámetros que relacionan a Y_{obs} con R. Dicho de otra manera, la probabilidad de que falten valores en Y está relacionada con otra variable observada Y_{obs} en el modelo de análisis, pero no con los valores de Y.

Finalmente, se dice que los valores son MNAR si $P(R = 0|Y_{obs}, Y_{mis}, \psi)$. En este caso no es posible simplificar la expresión general del modelo, puesto que la probabilidad de $R = 0$ puede depender tanto de valores observados Y_{obs} como de valores no observados Y_{mis}. En otras palabras, la probabilidad de que falten datos en Y puede depender de otras variables, por ejemplo observadas Y_{obs} como así también de valores no observados en sí mismos, es decir Y_{mis}.

Patrones de MV

Es importante distinguir entre patrones y mecanismos de MV. Un patrón se refiere a la configuración de valores observados y no observados en el conjunto de datos, mientras que el mecanismo describe las posibles relaciones entre las variables medidas y la probabilidad de MV. El patrón simplemente describe la localización de los MV en el conjunto de datos pero no explica las causas (Van Buuren, 2012).

La Figura 2, ilustra tres patrones típicos que se pueden encontrar en la literatura clásica (Enders, 2010), (Van Buuren, 2012), en color rojo se representa la localización de los MV. El patrón univariado (Figura 2a) presenta MV en una única variable. Es relativamente raro en algunas disciplinas, sin embargo, puede surgir en estudios experimentales. Los patrones monótonos o escalonados (Figura 1b), se asocian con estudios longitudinales en los cuales

participantes abandonan la encuesta y no regresan. Finalmente, el patrón multivariado o general (Figura1c) es quizás la configuración más común, los MV se encuentran distribuidos aleatoriamente en todo el conjunto de datos. El patrón aparentemente aleatorio es engañoso dado que los valores aún pueden faltar sistemáticamente (por ejemplo, puede haber una relación entre los valores de Y_1 y la propensión a los datos faltantes en Y_2).

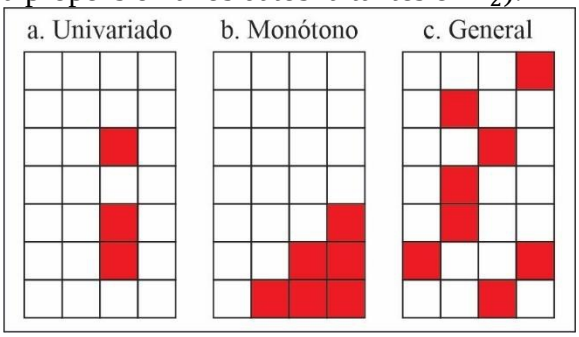

Figura 2. Patrones típicos en la literatura (Elaboración propia en base a (Enders, 2010) y (Van Buuren, 2012)).

Porcentaje de MV

El porcentaje o proporción de MV, conocidos como tasa de datos faltantes (MR: Missing Rates) (Santos et al., 2019), es una medida común de cuánta información se perdió en el conjunto de datos (Madley-Dowd et al., 2019). Las opiniones respecto de MR aceptables difieren (van der Meijs, 2018). Valores menores al 1% se consideran triviales y del 1 al 5% manejables. Sin embargo, pérdidas de entre el 5% y 15% requieren métodos sofisticados para su tratamiento, mientras que tasas de más del 15% pueden afectar severamente cualquier tipo de estudio (Twala, 2009).

Generación de MV

Un aspecto clave en la evaluación de los métodos de imputación de datos es la generación de MV (Santos et al., 2019). El proceso por el cual se generan conjuntos de datos con MV a partir de conjuntos de datos completos se conoce como amputación (Schouten et al., 2018).

En general, los investigadores desarrollan enfoques diseñados *ad-hoc*, en su mayoría, para conjuntos de datos particulares (Santos et al., 2019). Todas estas técnicas de amputación tienen un aspecto en común, los MV se generan en una variable a la vez en un proceso conocido como amputación univariada (Schouten et al., 2018).

El procedimiento de amputación multivariado (Schouten et al., 2018) es un enfoque alternativo que permite generar MV en múltiples variables para cualquier conjunto de datos. La amputación multivariada, facilita la definición de patrones y su ocurrencia relativa, permite manipular la MR y ajustar con precisión los distintos mecanismos de MV. Estas características han sido

determinantes para elegir este método en los ensayos realizados que se describirán más adelante.

Enfoques para enfrentarse a los MV

Los enfoques tradicionales más utilizados que ignoran los MV son el método de eliminación por listas (listwise deletion) o análisis de casos completos (CC: Complete-Case Analysis) y de eliminación por pares (Pairwise Deletion o Pairwise Inclusion) o análisis de casos disponibles (AC: Available-Case Analysis) (Jadhav et al., 2019). En el primer caso, el análisis se realiza con información completa, simplemente omitiendo todos aquellos casos con MV. La principal ventaja de este método es su sencillez, no obstante, reduce el tamaño de la muestra sobre todo si la MR es alta. El análisis AC intenta mitigar la pérdida de información utilizando únicamente los casos con datos completos para estimar diferentes parámetros de forma consistente. Ambos enfoques asumen que los valores son MCAR y si este supuesto no se cumple, las estimaciones podrían estar sesgadas (Enders, 2010).

Los métodos de imputación simple generan un único valor de reemplazo para cada MV y constituyen los enfoques más populares para sustituir valores no observados por estimaciones para obtener un conjunto de datos completo que puede ser analizado por diferentes métodos estadísticos (Jadhav et al., 2019). Una breve clasificación de los más populares incluye a los métodos de imputación por medias, por regresión, por regresión estocástica, hot-deck y cold-deck (Josepn L. Schafer & Graham, 2002).

La imputación por medias, también conocida como sustitución por la media o imputación por medias no condicionadas, es el método más simple para sustituir MV mediante el promedio aritmético de los valores observados para una variable.

El método de imputación por regresión o imputación por medias condicionadas reemplaza los MV con valores predichos a partir de una ecuación de regresión. Las variables tienden a estar correlacionadas, por lo tanto, tiene sentido generar imputaciones a partir de variables observadas. El método de imputación por regresión estocástica es una variante que sustituye MV utilizando una ecuación de regresión que incluye un término de error residual normalmente distribuido para mejorar las predicciones.

Los métodos de imputación hot-deck, consisten en una colección de técnicas que imputan los MV con valores de observaciones similares. En su forma más simple, reemplaza cada MV por un valor calculado a partir de uno o más casos completos (donantes) del mismo conjunto de datos. Estos se pueden elegir aleatoriamente de uno de los donantes o calculando la media de los correspondientes valores de los donantes. El método de imputación cold-deck es similar, excepto que los donantes se obtienen de una fuente distinta a la del conjunto de datos a imputar.

La imputación simple asume que los valores son MCAR y su principal desventaja es que no tiene en cuenta la variabilidad debido a la predicción de los MV, lo que conduce a subestimar el error estándar de los parámetros calculados a partir del conjunto de datos imputado (Enders, 2010).

Uno de los métodos para estimar parámetros desconocidos de un modelo es la estimación por máxima verosimilitud (ML: Maximum Likelihood). Cuando se cuenta con conjuntos de datos completos, las estimaciones se basan en maximizar la verosimilitud de los datos observados. El mismo principio se cumple cuando se tienen conjuntos de datos con MV (Pigott, 2001). El algoritmo EM (Expectation-Maximization) (Dempster et al., 1977), es un enfoque iterativo general de dos pasos, E (expectation) y M (maximization), ampliamente utilizado para obtener estimaciones por ML de parámetros en conjuntos de datos con MV. Estas estimaciones no requieren conocimiento acerca de si los valores son MCAR o MAR, sin embargo pueden resultar sesgadas bajo el supuesto de valores MNAR (Pigott, 2001).

Propuestos por Rubin (1987), los métodos de imputación múltiple (MI: Multiple Imputation) generan un número de copias del conjunto de datos original con diferentes imputaciones, analizando cada uno por separado. Estos, producen múltiples conjuntos de parámetros y errores estándar que se combinan en un único conjunto de resultados. En general, entre 5 y 10 imputaciones son suficientes para producir inferencias altamente eficientes (Pigott, 2001) (Tobias, 2017). Constituye una estrategia robusta, puesto que requiere supuestos menos estrictos acerca del mecanismo de MV. Sin embargo, continúan siendo débiles, ya que en el caso de valores MNAR las estimaciones podrían estar sesgadas (Pigott, 2001) (Tobias, 2017).

En los últimos años, se ha propuesto el uso de algoritmos de aprendizaje automático (ML: Machine Learning) como métodos de imputación (Liu & Gopalakrishnan, 2017). Estas técnicas se basan en la construcción de un modelo predictivo para estimar valores no observados en función de los valores observados en el conjunto de datos (García-Laencina et al., 2010). Tanto los algoritmos de aprendizaje supervisados (clasificación) como los no supervisados (clustering) se pueden adecuar para la imputación (Liu & Gopalakrishnan, 2017). Algoritmos de ML bien conocidos tales como los árboles de decisión (DT: Decision Trees), k-vecinos más cercanos (k-NN: k-Nearest Neighbours), agrupamiento por k-Medias (k-Means Clustering) y redes bayesianas (Bayesian Networks), han sido utilizados como métodos de imputación en diferentes dominios (Liu & Gopalakrishnan, 2017) (Rahman & Davis, n.d.) (García-Laencina et al., 2010) (Jerez et al., 2010) (Nadzurah et al., 2018) (Luengo et al., 2012).

Materiales y métodos

En esta sección se describe el procedimiento seguido para evaluar el desempeño de métodos de imputación mediante una nueva medida, basada en el promedio normalizado de los RMSE.

Descripción del conjunto de datos

El conjunto de datos "Iris" se obtuvo del UCI Machine Learning Laboratory (*Iris Data Set*, 2020). Contiene tres clases de 50 instancias cada una, 150 en total sin MV, referidas a tres tipos de flores de la planta iris ("setosa", "versicolor" y "virginica"). El conjunto de datos tiene cinco atributos, cuatro predictivos ("sepal length", "sepal width", "petal length" y "petal width") y uno de clase ("class").

Se realizó un análisis para detectar y eliminar outliers, valores extremos que se desvían de otras observaciones (Ben-Gal, 2005). Se utilizó el método de la desviación estándar para cada variable y se eliminaron todas aquellas observaciones con valores fuera del intervalo $\bar{x} \pm 2\sigma$ (Seo, 2006), obteniendo un conjunto de datos completo con 139 casos.

Generación de los conjuntos de datos con MV

Se adoptó el método de amputación multivariado descripto en (Schouten et al., 2018) implementado en la función "ampute" del paquete "mice" (*Multivariate Imputation by Chained Equations,* 2020) del software R (*The R Project for Statistical Computing*, 2020).

Se definieron tres valores de MR (10%, 15% y 20%) para casos definidos en tres configuraciones de patrones (univariado, multivariado simple y multivariado complejo), generados bajo los supuestos MCAR, MAR y MNAR. Un patrón univariado (univa) tiene MV en una variable, un patrón multivariado simple (multiva2) tiene MV en hasta dos variables y un patrón multivariado complejo (multiva3) en hasta 3 variables. Se definió la frecuencia de ocurrencia de cada patrón en $1/k$, siendo k el número de patrones especificados.

Para la generación de valores MCAR se siguió lo expuesto por (Twala, 2009), considerando que las variables a ser amputadas deben ser aquellas más correlacionadas con la variable de clase t. De esta manera, en la generación de valores MCAR univa se seleccionó "petal width" como variable a ser amputada. En la generación de valores MCAR multiva2, se seleccionaron como variables a ser amputadas "petal width" y "petal length". Finalmente, para la generación de valores MCAR multiva3, se seleccionaron como variables a ser amputadas "petal width", petal length" y "sepal length".

La amputación de valores MAR sigue lo expuesto por (Twala et al., 2006) utilizando una variable observada (y_i^{obs}) determinante, también conocida como causativa (Garciarena & Santana, 2017), que define la

localización de la variable no observada (y_i^{mis}). Se consideraron pares de variables correlacionadas (y_i^{obs}, y_i^{mis}), donde la primera componente (y_i^{obs}) indica la variable causativa y la segunda (y_i^{miss}) la variable a ser amputada, aquella más correlacionada con la variable de clase t.

De esta manera se consideraron los pares {("petal length", "petal width")} para la generación de valores MAR univa; {("peta width", "petal length"), ("petal length", "sepal length)} para la generación de valores MAR multiva2 y {("petal width", "petal length"), ("petal length", "sepal length"), ("sepal width", ("sepal length")} para la generación de valores MAR multiva3.

La generación de valores MNAR univa, MNAR multiva2 y MNAR multiva3 sigue un enfoque similar al propuesto para la generación de valores MAR, pero en este caso la variable a ser amputada está determinada por la componente (y_i^{mis}) en sí misma.

En la generación de valores MAR y MNAR los casos candidatos a contener MV reciben una probabilidad basada en un puntaje de suma ponderada (WSS: weighted sum score). La asignación de estas probabilidades se realiza en base a una función con distribución logística. Los WSS se obtienen multiplicando cada variable por un peso elegido arbitrariamente (Schouten et al., 2018). Para cada patrón definido, es posible ponderar cada variable de manera que gobierne el impacto de éstas en la formación de los WSS. Variables con pesos altos tendrán mayor influencia que variables con pesos bajos.

De esta manera, es posible generar valores MAR asignando pesos igual a cero a las variables que serán amputadas y pesos distintos de cero a las variables determinantes. Por el contrario, si se asignan pesos distintos de cero a las variables que serán amputadas, se obtendrán valores MNAR.

Para la generación de valores MAR y MNAR, las variables causativas fueron ponderadas con pesos igual a 8 y se consideraron tres distribuciones logísticas: de cola derecha (RIGHT), centrada (MID) y de cola izquierda (LEFT) (Schouten et al., 2018).

Imputación de los conjuntos de datos amputados

En este trabajo se utilizaron los lenguajes de programación R (*The R Project for Statistical Computing*, 2020) y Python (*Python*, 2020). Se utilizó la librería "rpy2" (Foundation, 2020) como interfaz entre ambos lenguajes para la ejecución de experimentos de amputación e imputación.

Se utilizaron los métodos de imputación por medias, hot-deck, k-NN y k-Means. La clase "SimpleImputer" (*sklearn.impute.SimpleImputer*, 2020) de la librería "scikit-learn" (*scikit-learn*, 2007) de Python (*Python*, 2020) implementó la estrategia de imputación simple por media. La clase "k-

NNImputer", de la librería "missingpy" (*missingpy 0.2.0*, 2020) de Python, soportó la imputación utilizando el enfoque k-NN. La función "hot.deck" del paquete "hot.deck-package" (*Multiple Hot-Deck Imputation*, 2020) del software R (*The R Project for Statistical Computing*, 2020) proporcionó una implementación del método de imputación Hot Deck. Finalmente, el método de imputación por k-Means se desarrolló en dos pasos (Gajawada & Toshniwal, 2012), utilizando la clase "sklearn.cluster.k-Means" (*sklearn.cluster.KMeans*, 2020) de la librería "scikit-learn" (*scikit-learn*, 2007) de Python (*Python*, 2020). En el primer paso, se calculó la cantidad óptima de clusters. Seguidamente, la información de los clusters obtenidos se utilizó para imputar los MV.

Evaluación del desempeño de los algoritmos de imputación

Uno de los indicadores de desempeño más representativo y ampliamente utilizado para evaluar el desempeño de los métodos de imputación es el RMSE (Jadhav et al., 2019). En este estudio se propone un nuevo indicador basado en esta métrica tradicional.

Sea Y el conjunto de datos completo, representado por la matriz de la Tabla 1, con n casos y p variables donde y_{ij}, con $1 \leq i \leq n$ y $1 \leq j \leq p$, son valores observados.

Y_1	Y_2	…	Y_j	…	Y_p
y_{11}	y_{12}	…	y_{1j}	…	y_{1p}
y_{21}	y_{22}	…	y_{2j}	…	y_{2p}
…	…	…	…	…	…
y_{i1}	y_{i2}	…	y_{ij}	…	y_{ip}
…	…	…	…	…	…
y_{n1}	y_{n2}	…	y_{nj}	…	y_{np}

Tabla 1. Conjunto de datos completo **Y** (Elaboración propia).

Sean a_r y m_s, $1 \leq r \leq l$ y $1 \leq s \leq q$, con l indicando la cantidad de procedimientos de amputación y q representando el número de métodos de imputación.

Sea el conjunto $Y^{a_r m_s}$, representado en la matriz de la Tabla 2, que contiene a los elementos $y_{ij}^{a_r m_s}$ que son valores imputados por el método m_s luego de haber sido el valor original amputado por el procedimiento a_r.

$Y_1^{a_r m_s}$	$Y_2^{a_r m_s}$	…	$Y_j^{a_r m_s}$	…	$Y_p^{a_r m_s}$
$y_{11}^{a_r m_s}$	$y_{12}^{a_r m_s}$	…	$y_{1j}^{a_r m_s}$	…	$y_{1p}^{a_r m_s}$
$y_{21}^{a_r m_s}$	$y_{22}^{a_r m_s}$	…	$y_{2j}^{a_r m_s}$	…	$y_{2p}^{a_r m_s}$
⋮	⋮	…	…	…	⋮
$y_{i1}^{a_r m_s}$	$y_{i2}^{a_r m_s}$	…	$y_{ij}^{a_r m_s}$	…	$y_{ip}^{a_r m_s}$
⋮	⋮	…	⋮	…	⋮
$y_{n1}^{a_r m_s}$	$y_{n2}^{a_r m_s}$	…	$y_{nj}^{a_r m_s}$	…	$y_{np}^{a_r m_s}$

Tabla 2. Conjunto de datos con elementos $\mathbf{y_{ij}^{a_r m_s}}$ imputados por el método $\mathbf{m_s}$ luego de haber sido amputados por el método $\mathbf{a_r}$ (Elaboración propia).

Sea $F_j^{a_r m_s} \subset \{1, \dots, n\}$ el conjunto de índices correspondientes a los valores de la variable $Y_j^{a_r m_s}$ imputada por el método m_s luego de haber sido el valor original amputado por el método a_r.

A continuación, se describe la metodología propuesta para obtener el indicador.

Paso 1: Se calcula el RMSE para la variable $Y_j^{a_r m_s}$ imputada por el método m_s luego de haber sido amputada por el procedimiento a_r:

$$\mathrm{RMSE}\left(Y_j^{a_r m_s}\right) = \sqrt{\frac{\sum_{i \in F_j^{a_r m_s}} \left(y_{ij} - y_{ij}^{a_r m_s}\right)^2}{\#\left(F_j^{a_r m_s}\right)}}; \begin{array}{l} 1 \le j \le p \\ \text{con } 1 \le r \le l \\ 1 \le s \le q \end{array}$$

Estos resultados se pueden sintetizar en la Tabla 3:

$Y_j^{a_r m_s}$	$\mathrm{RMSE}\left(Y_j^{a_r m_s}\right)$
$Y_1^{a_1 m_1}$	$\mathrm{RMSE}\left(Y_1^{a_1 m_1}\right)$
$Y_1^{a_1 m_2}$	$\mathrm{RMSE}\left(Y_1^{a_1 m_2}\right)$
⋮	⋮
$Y_j^{a_r m_s}$	$\mathrm{RMSE}\left(Y_j^{a_r m_s}\right)$
$Y_p^{a_l m_q}$	$\mathrm{RMSE}\left(Y_p^{a_l m_q}\right)$

Tabla 3. RMSE para la variable $\mathbf{Y_j^{a_r m_s}}$ (Elaboración propia)

Paso 2: Se normaliza el RMSE para cada variable $Y_j^{a_r m_s}$:

$$\mathrm{RMSEN}\left(Y_j^{a_r m_s}\right) = \frac{\mathrm{RMSE}\left(Y_j^{a_r m_s}\right) - \min\left\{\mathrm{RMSE}\left(Y_j^{a_t m_u}\right)\right\}}{\max\left\{\mathrm{RMSE}\left(Y_j^{a_t m_u}\right)\right\} - \min\left\{\mathrm{RMSE}\left(Y_j^{a_t m_u}\right)\right\}}; \text{con } \begin{matrix} 1 \le j \le p \\ 1 \le r \le l \\ 1 \le s \le q \\ 1 \le t \le l \\ 1 \le u \le q \end{matrix}$$

Esto facilita comparar los RMSE para cada variable, dado que éstas pueden estar en diferentes escalas. Estos resultados se pueden sintetizar en la Tabla 4.

$Y_j^{a_r m_s}$	$\mathrm{RMSE}\left(Y_j^{a_r m_s}\right)$	$\mathrm{RMSEN}\left(Y_j^{a_r m_s}\right)$
$Y_1^{a_1 m_1}$	$\mathrm{RMSE}\left(Y_1^{a_1 m_1}\right)$	$\mathrm{RMSEN}\left(Y_1^{a_1 m_1}\right)$
$Y_1^{a_1 m_2}$	$\mathrm{RMSE}\left(Y_1^{a_1 m_2}\right)$	$\mathrm{RMSEN}\left(Y_1^{a_1 m_2}\right)$
⋮	⋮	⋮
$Y_j^{a_r m_s}$	$\mathrm{RMSE}\left(Y_j^{a_r m_s}\right)$	$\mathrm{RMSEN}\left(Y_j^{a_r m_s}\right)$
$Y_p^{a_l m_q}$	$\mathrm{RMSE}\left(Y_p^{a_l m_q}\right)$	$\mathrm{RMSEN}\left(Y_p^{a_l m_q}\right)$

Tabla 4. RMSE Normalizado para cada variable $\mathbf{Y_j^{a_r m_s}}$ (Elaboración propia).

Paso 3: Se calcula el promedio de los RMSEN para cada uno de los métodos de imputación m_s y cada una de las variables $Y_j^{a_r m_s}$, para todos los métodos de amputación a_r:

$$\mathrm{PRMSEN}\left(Y_j^{m_s}\right) = \frac{\sum_{r=1}^{l} \mathrm{RMSEN}\left(Y_j^{a_r m_s}\right)}{l}; \text{con } \begin{matrix} 1 \le j \le p \\ 1 \le s \le q \end{matrix}$$

Se puede sintetizar en la Tabla 5.

Paso 4: Finalmente, se calcula el promedio de errores producidos por el método m_s sobre todas las variables Y_j que da un valor representativo del error producido por cada método de imputación para la totalidad de conjuntos de datos imputados. Errores bajos, indican mejor desempeño del método de imputación:

$$E(m_s) = \frac{\sum_{j=1}^{p} \mathrm{PRMSEN}\left(Y_j^{m_s}\right)}{p}; \text{ con } s = \{1, \dots, q\}$$

Se puede sintetizar en la Tabla 6.

	m_1	m_2	$\cdots$	m_s	$\cdots$	m_q
Y_1	$\mathrm{PRMSEN}(\mathrm{Y}_1^{\mathrm{m}_1})$	$\mathrm{PRMSEN}(\mathrm{Y}_1^{\mathrm{m}_2})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_1^{\mathrm{m}_s})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_1^{\mathrm{m}_q})$
Y_2	$\mathrm{PRMSEN}(\mathrm{Y}_2^{\mathrm{m}_1})$	$\mathrm{PRMSEN}(\mathrm{Y}_2^{\mathrm{m}_2})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_2^{\mathrm{m}_s})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_2^{\mathrm{m}_q})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_j	$\mathrm{PRMSEN}(\mathrm{Y}_\mathrm{j}^{\mathrm{m}_1})$	$\mathrm{PRMSEN}(\mathrm{Y}_\mathrm{j}^{\mathrm{m}_2})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_\mathrm{j}^{\mathrm{m}_s})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_\mathrm{j}^{\mathrm{m}_q})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_p	$\mathrm{PRMSEN}(\mathrm{Y}_p^{\mathrm{m}_1})$	$\mathrm{PRMSEN}(\mathrm{Y}_p^{\mathrm{m}_2})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_p^{\mathrm{m}_s})$	$\vdots$	$\mathrm{PRMSEN}(\mathrm{Y}_p^{\mathrm{m}_q})$

Tabla 5. Promedio de los RMSEN para cada uno de los métodos de imputación $\mathbf{m_s}$ y cada una de las variables $\mathbf{Y_j^{a_r m_s}}$, para todos los métodos de amputación $\mathbf{a_r}$ (Elaboración propia)

m_1	m_2	$\cdots$	m_s	$\cdots$	m_q
$\mathrm{E}(\mathrm{m}_1)$	$\mathrm{E}(\mathrm{m}_2)$	$\cdots$	$\mathrm{E}(\mathrm{m}_\mathrm{s})$	$\cdots$	$\mathrm{E}(\mathrm{m}_q)$

Tabla 6. Promedio de errores producidos por el método $\mathbf{m_s}$ sobre todas las variables $\mathbf{Y_j}$ (Elaboración propia).

Resultados y discusiones

A partir del conjunto de datos original sin outliers se ejecutaron los procedimientos de amputación y se obtuvieron en total 63 conjuntos de datos con MV. De estos, 9 (3 patrones x 3 MR) corresponden a conjuntos de datos amputados bajo el supuesto MCAR, 27 (3 patrones x 3 MR x 3 tipos) a conjuntos de datos amputados bajo el supuesto MAR y 27 (3 patrones x 3 MR x 3 tipos) a conjuntos de datos amputados bajo el supuesto MNAR.

Los 63 conjuntos de datos amputados fueron imputados mediante los métodos de imputación por medias, k-NN, k-Means y hot-deck, de los cuales se obtuvieron en total 252 conjuntos de datos imputados. De estos, 36 (9 conjuntos de datos amputados x 4 métodos de imputación) corresponden a conjuntos de datos imputados luego de haber sido amputados bajo el supuesto MCAR, 108 (27 conjuntos de datos amputados x 4 métodos de imputación) corresponden a conjuntos de datos imputados luego de haber sido amputados bajo el supuesto MAR y 108 (27 conjuntos de datos amputados x 4 métodos de imputación) corresponden a conjuntos imputados luego de haber sido amputados bajo el supuesto MNAR.

Luego de imputados los conjuntos de datos amputados, se calculó el RMSE para las variables “petal width”, “petal length” y “sepal length” imputadas por los cuatro métodos de imputación luego de haber sido amputadas por los 63 procedimientos de amputación.

En las Tablas 7 a 15 se presentan los RMSE obtenidos para las variables “petal width”, “petal length” y “sepal length” del conjunto de datos “Iris”

imputadas por los métodos media, k-NN, k-Means y hot-deck, luego de haber sido amputadas de acuerdo a los 63 procedimientos definidos.

Mecanismo	Tipo	Variable Amputada	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	-	petal.width	0,75998962	0,20600966	0,21376225	0,42817442
MAR	LEFT	petal.width	0,87816775	0,13264133	0,14071314	0,28047579
	MID	petal.width	0,70369456	0,16350208	0,181797	0,38271476
	RIGHT	petal.width	0,80500398	0,21232206	0,21689876	0,33431229
MNAR	LEFT	petal.width	0,87816775	0,13264133	0,14084575	0,28047579
	MID	petal.width	0,68333333	0,16816116	0,19449282	0,38729833
	RIGHT	petal.width	0,8687979	0,23444956	0,24426747	0,53412931

Tabla 7. RMSE para la variable "petal width" imputada por medias, k-NN, k-Means y hot-deck luego de haber sido amputada bajo los supuestos MCAR, MAR y MNAR en un 10% de los casos (Elaboración propia).

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	-	petal.length	1,51135976	0,39006255	0,38046459	0,53851648
		petal.width	0,60373904	0,14059874	0,14378355	0,41533119
MAR	LEFT	petal.width	0,89477605	0,11571469	0,11026825	0,20916501
		petal.length	2,10045108	0,13822873	0,13360983	1,2228592
	MID	petal.width	0,63383913	0,18054184	0,22082303	0,49613894
		petal.length	1,55739935	0,33772313	0,37746644	0,67352533
	RIGHT	petal.width	0,62339739	0,18886354	0,22661633	0,62021502
		petal.length	1,61011056	0,34872428	0,42082671	0,65737574
MNAR	LEFT	petal.length	2,10045108	0,13822873	0,13734795	1,2228592
		petal.width	0,83908077	0,1173637	0,12527034	0,23048861
	MID	petal.length	1,50904663	0,35247836	0,30774712	0,50695167
		petal.width	0,55003687	0,12138589	0,18573439	0,4330127
	RIGHT	petal.length	1,61011056	0,34872428	0,47619264	0,65737574
		petal.width	0,62339739	0,18886354	0,21572414	0,62021502

Tabla 8. RMSE para las variables "petal width" y "petal length" imputadas por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 10% de los casos (Elaboración propia).

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		sepal.length	0,83415793	0,39451103	0,61816921	0,82689177
		petal.length	1,82488162	0,40069025	0,60819554	0,37416574
		petal.width	0,86073605	0,08680657	0,10861836	0,41833001
MAR	LEFT	sepal.length	0,657843	0,65228683	0,48348767	0,66426651
		petal.width	0,85527846	0,64236804	0,14869181	0,51168172
		petal.length	1,92298402	1,53841493	0,24964251	1,33416641
	MID	sepal.length	0,65879911	0,46812125	0,29064356	0,55827114
		petal.width	0,83563914	0,60244256	0,16078283	0,5574668
		petal.length	1,70539063	1,36818077	0,29112868	1,57733953
	RIGHT	sepal.length	0,76830527	0,50607325	0,24214167	0,61933144
		petal.width	0,89988784	0,5934947	0,16216515	0,47659507

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
		petal.length	1,81496443	1,23560417	0,22952416	1,00540208
MNAR	LEFT	sepal.length	0,76931965	0,3226127	0,20702394	0,37859389
		petal.length	1,87752365	1,28922912	0,23754197	1,28208814
		petal.width	0,91029731	0,50361666	0,12817453	0,30759614
	MID	sepal.length	0,56406208	0,5219328	0,40992191	0,70071392
		petal.length	1,66604789	1,36934187	0,29025662	1,21119775
		petal.width	0,80094205	0,58150378	0,14650257	0,65574385
	RIGHT	sepal.length	0,85847202	0,61583399	0,51217763	0,72407577
		petal.length	1,67617684	1,17525807	0,39229043	1,45602198
		petal.width	0,85262422	0,68060124	0,2245507	0,7358183

Tabla 9. RMSE para la variables "petal width", "petal length" y "sepal length" imputadas por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 10% de los casos (Elaboración propia).

Mecanismo	Tipo	Variable Amputada	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	-	petal.width	0,7680641	0,19001203	0,20900838	0,39210968
MAR	LEFT	petal.width	0,90944566	0,12692397	0,16056235	0,34525353
	MID	petal.width	0,63498906	0,15247905	0,17656363	0,43497126
	RIGHT	petal.width	0,74718899	0,18332058	0,1849556	0,49923018
MNAR	LEFT	petal.width	0,86227642	0,16297443	0,18649692	0,26608269
	MID	petal.width	0,61578352	0,14795206	0,16491915	0,44762749
	RIGHT	petal.width	0,80813934	0,22588165	0,23502721	0,53932325

Tabla 10. RMSE para la variable "petalwidth" imputada por medias, k-NN, k-Means y hot-deck luego de haber sido amputada bajo los supuestos MCAR, MAR y MNAR en un 15% de los casos (Elaboración propia).

Mecanismo	Tipo	Variables amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		petal.length	1,56544233	0,42211887	0,37153778	0,59314263
		petal.width	0,79553551	0,22709779	0,25083229	0,72663608
MAR	LEFT	petal.width	0,84539283	0,11264719	0,15052016	0,37300192
		petal.length	1,98383497	0,25981295	0,32813083	0,52153619
	MID	petal.width	0,60572066	0,21235013	0,22848066	0,49043482
		petal.length	1,48476277	0,35832018	0,40639975	0,74386379
	RIGHT	petal.width	0,61097156	0,24957643	0,23800877	0,41298372
		petal.length	1,71901134	0,33012211	0,42251514	0,67564784
MNAR	LEFT	petal.length	2,0125638	0,17216239	0,18588964	0,39791121
		petal.width	0,82824478	0,1128208	0,14152516	0,26371472
	MID	petal.length	1,48476277	0,35832018	0,42241703	0,74386379
		petal.width	0,60572066	0,21235013	0,21955916	0,49043482
	RIGHT	petal.length	1,63114496	0,32653501	0,43540756	0,63683244
		petal.width	0,61097156	0,24957643	0,22361811	0,48476799

Tabla 11. RMSE para la variables "petal width" y "petal length" imputadas por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 15% de los casos (Elaboración propia).

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		sepal.length	0,83327037	0,41459919	0,65690257	0,59254629
		petal.length	1,82488162	0,40069025	0,67687212	0,49598387
		petal.width	0,8353124	0,09659472	0,13806278	0,3681787
MAR	LEFT	sepal.length	0,60500411	0,44274609	0,26201438	0,35237291
		petal.width	0,88125913	0,53724261	0,16921492	0,3009788
		petal.length	1,90296024	1,39551121	0,28299172	1,72336879
MAR	MID	sepal.length	0,59968742	0,452271	0,37446203	0,70663522
		petal.width	0,78019709	0,52924687	0,15414335	0,42426407
		petal.length	1,7427121	1,30499616	0,24032491	1,24863562
MAR	RIGHT	sepal.length	0,72772327	0,488311	0,42106299	0,41499665
		petal.width	0,88096843	0,60709988	0,10809613	0,49874843
		petal.length	1,6549804	1,15463726	0,27506058	0,76696499
MNAR	LEFT	sepal.length	0,75785352	0,42346785	0,21597954	0,35683797
		petal.length	1,86965868	1,4103862	0,20724565	1,3065221
		petal.width	0,92443202	0,53285353	0,11157257	0,24614678
MNAR	MID	sepal.length	0,71610226	0,50145523	0,47648655	0,5501196
		petal.length	1,62839416	1,17785461	0,29828095	1,12527774
		petal.width	0,76948034	0,56533179	0,17597654	0,53655823
MNAR	RIGHT	sepal.length	0,76988998	0,5655635	0,46716343	0,62573034
		petal.length	1,66141886	1,16645937	0,29888906	1,42628388
		petal.width	0,81603254	0,63705327	0,17340868	0,55884496

Tabla 12. RMSE para la variables "petal width", "petal length" y "sepal length" imputadas por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 10% de los casos (Elaboración propia).

Mecanismo	Tipo	Variable Amputada	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		petal.width	0,76454799	0,17382086	0,20110172	0,43643578
MAR	LEFT	petal.width	0,95410812	0,12518338	0,14650572	0,20615528
	MID	petal.width	0,63832558	0,16309546	0,18592667	0,37807562
	RIGHT	petal.width	0,76593097	0,19976403	0,20248819	0,53650521
MNAR	LEFT	petal.width	0,90220617	0,14831944	0,16587806	0,20670576
	MID	petal.width	0,59988332	0,1600255	0,18151841	0,41560471
	RIGHT	petal.width	0,82274455	0,21313111	0,18621013	0,49292289

Tabla 13. RMSE para la variable "petal width" imputada por medias, k-NN, k-Means y hot-deck luego de haber sido amputada bajo los supuestos MCAR, MAR y MNAR en un 20% de los casos (Elaboración propia).

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		petal.length	1,64215332	0,37890625	0,42543633	0,44158804
		petal.width	0,79877191	0,20799593	0,24167141	0,45064057
MAR	LEFT	petal.width	0,83247721	0,15198745	0,17821218	0,2
		petal.length	1,86501693	0,27650897	0,34328586	0,42919754
	MID	petal.width	0,58881473	0,23273521	0,21517664	0,37352886
		petal.length	1,50082571	0,3391131	0,3525311	0,71530879

	RIGHT	petal.width	0,6806401	0,23418031	0,2084347	0,42732739
		petal.length	1,72391083	0,34596083	0,37951163	0,69448376
MNAR	LEFT	petal.length	1,82885333	0,33928391	0,36162865	0,40804412
		petal.width	0,82109096	0,11562293	0,12686536	0,22263247
	MID	petal.length	1,50082571	0,3391131	0,37532179	0,71530879
		petal.width	0,58881473	0,23273521	0,21267394	0,37352886
	RIGHT	petal.length	1,71729466	0,34524254	0,35979056	0,61456676
		petal.width	0,6806401	0,23418031	0,20674634	0,42732739

Tabla 14. RMSE para la variables "petal width" y "petal length" imputadas por medias, k-NN, k-Means y hot-deckluego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 20% de los casos (Elaboración propia)

Mecanismo	Tipo	Variables Amputadas	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR		sepal.length	0,84811146	0,57416592	0,41799813	0,52915026
		petal.length	1,81206129	1,17614638	0,38203534	1,29614814
		petal.width	0,83221105	0,41072312	0,11638496	0,29720924
MAR	LEFT	sepal.length	0,73548435	0,53378759	0,33892501	0,61682752
		petal.width	0,88315099	0,53222368	0,13307307	0,52076866
		petal.length	1,94514044	1,37765771	0,26647523	1,08685326
	MID	sepal.length	0,62693823	0,45906289	0,44461159	0,65223973
		petal.width	0,78778812	0,55366926	0,11728388	0,60710084
		petal.length	1,60955303	1,15289386	0,25027271	0,58166428
	RIGHT	sepal.length	0,75454702	0,51259964	0,40818065	0,56391489
		petal.width	0,81431974	0,58331597	0,18996966	0,57817447
		petal.length	1,77065566	1,13329839	0,30636616	0,9334686
MNAR	LEFT	sepal.length	0,83053394	0,50719968	0,25710726	0,53229065
		petal.length	2,09428142	1,32250656	0,24679042	1,12866686
		petal.width	0,9620648	0,54236161	0,10420529	0,47740619
	MID	sepal.length	0,6736324	0,51594412	0,40109061	0,63874878
		petal.length	1,55707079	1,088963	0,22166146	0,88078912
		petal.width	0,7488772	0,52020687	0,15493296	0,4969472
	RIGHT	sepal.length	0,80276464	0,54123884	0,51927269	0,674698
		petal.length	1,69512549	1,07489514	0,26474349	0,65087354
		petal.width	0,72255587	0,58669878	0,13363914	0,48666426

Tabla 15. RMSE para la variables "petal width", "petal length" y "sepal length" imputadas por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 10% de los casos (Elaboración propia).

A continuación, se normalizaron los RMSE para las variables petal width", "petal length" y "sepal length". A menor valor del RMSE Normalizado, mejor es la estimación dada por el método de imputación.

En las Tablas 16 a 24, se presentan los RMSE Normalizados para las variables "petal width", "petal length" y "sepal length" para los distintos porcentajes de casos imputados. Se utilizó un mapa de calor en escala de grises, donde valores bajos se representan con colores claros y valores altos con colores oscuros.

En la Tabla 16 se presentan los RMSE Normalizados para la variable "petal width" imputada por los cuatro métodos luego de haber sido amputada en un 10% de los casos, en patrones univa, multiva2 y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 16, el desempeño de los métodos de imputación k-NN y k-Means resultaron ser los mejores en todos los escenarios.

El método k-NN, resultó ser el mejor para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MCAR en patrones univa, multiva2 y multiva3.

En el escenario para la variable "petal width" amputada bajo el supuesto MAR en patrón univa y considerando las tres distribuciones (LEFT, MID y RIGHT), k-NN resultó ser el mejor método de imputación.

Bajo el supuesto MAR en patrón multiva2 y distribución LEFT, el método k-Means resultó el mejor. Sin embargo, k-NN fue el mejor en el caso de las distribuciones MID y RIGHT.

La situación cambió para la variable "petal width" amputada bajo el supuesto MAR en patrón multiva3 y considerando las tres distribuciones (LEFT, MID y RIGHT), resultando en este caso k-Means ser el mejor método de imputación.

Finalmente, k-NN resultó ser el mejor método para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MNAR en patrones univa y multiva2 y las tres distribuciones (LEFT, MID y RIGHT). Mientras que, k-Means resultó ser el mejor en el supuesto MNAR en patrón multiva3 y distribuciones LEFT, MID y RIGHT.

En síntesis, para la variable "petal width" amputada en el 10% de los casos considerando 21 escenarios diferentes, en 14 k-NN resultó ser el mejor método de imputación, mientras que en los 7 restantes, el mejor resultó ser k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	univa	-	0,76912507	0,13619192	0,1450494	0,39001958
	multiva2	-	0,59060567	0,06145863	0,06509734	0,37534594
	multiva3	-	0,88422988	0	0,02492041	0,37877215
MAR	univa	LEFT	0,90414594	0,05236713	0,06158934	0,22127095
		MID	0,70480684	0,08762616	0,10852847	0,33808102
		RIGHT	0,82055487	0,14340396	0,14863293	0,28278023
	multiva2	LEFT	0,92312125	0,03302811	0,02680544	0,13979696
		MID	0,62499562	0,10709442	0,15311649	0,4676704
		RIGHT	0,61306572	0,11660213	0,15973544	0,6094298
	multiva3	LEFT	0,87799447	0,63474007	0,07070512	0,48542834
		MID	0,85555616	0,58912441	0,08451936	0,53773871
		RIGHT	0,92896158	0,5789013	0,08609868	0,44534114
MNAR	univa	LEFT	0,90414594	0,05236713	0,06174084	0,22127095

		MID	0,68154374	0,09294924	0,12303369	0,34331784
		RIGHT	0,8934407	0,16868507	0,17990223	0,51107516
	multiva2	LEFT	0,85948828	0,03491214	0,04394563	0,1641596
		MID	0,52924987	0,03950757	0,11302701	0,39554742
		RIGHT	0,61306572	0,11660213	0,1472909	0,6094298
	multiva3	LEFT	0,94085461	0,47621385	0,04726373	0,25225649
		MID	0,81591404	0,56520144	0,06820387	0,6500222
		RIGHT	0,87496195	0,67842226	0,15737542	0,74150886

Tabla 16. RMSE Normalizado para la variable "petal width" amputada en un 10% de los casos (Elaboración propia).

En la Tabla 17 se presentan los RMSE Normalizados para la variable "petal length" imputada por los cuatro métodos luego de haber sido amputada en un 10% de los casos, en patrones multiva2 y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 17, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, k-Means resultó ser el mejor método de imputación para imputar la variable "petal length" luego de haber sido amputada bajo el supuesto MCAR en patrón multiva2. Sin embargo, bajo el supuesto MCAR multiva3, k-NN resultó ser el mejor método para imputar la misma variable.

k-Means resultó ser el mejor método para imputar la variable "petal length" amputada bajo el supuesto MAR en patrón multiva2 y distribución LEFT. Por el contrario, bajo el supuesto MAR en patrón multiva2 y distribuciones MID y RIGHT el mejor método resultó ser k-NN.

Para la variable "petal length" amputada bajo el supuesto MAR en patron multiva3 y las tres distribuciones (LEFT, MID y RIGHT) el mejor método de imputación resulto ser k-Means.

En el escenario para el cual la variable "petal width" fue amputada bajo el supuesto MNAR en patrón multiva2 y distribuciones LEFT y MID, k-Means resultó ser el mejor método para imputar la variable "petal length". Por el contrario, en el caso de distribución RIGHT, k-NN resulto ser el mejor para imputar la misma variable.

Finalmente, k-Means resultó ser el mejor método para imputar la variable "petal length" luego de haber sido amputada bajo el supuesto MNAR en patrón multiva3 y las tres distribuciones (LEFT, MID y RIGHT).

En síntesis, para la variable "petal length" amputada en el 10% de los casos y considerando 14 escenarios de amputación diferentes, en 4 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 10, el mejor fue k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva2		0,70048863	0,13038812	0,12550823	0,20586646
	multiva3		0,85989237	0,13579155	0,24129335	0,12230571
MAR	multiva2	LEFT	1	0,00234839	0	0,55380645
		MID	0,72389651	0,10377721	0,12398388	0,27450894
		RIGHT	0,75069644	0,10937052	0,14602952	0,26629801
	multiva3	LEFT	0,90977052	0,71424428	0,05899444	0,61039831
		MID	0,79913964	0,62769222	0,08008722	0,73403469
		RIGHT	0,85485018	0,56028637	0,04876567	0,44324485
MNAR	multiva2	LEFT	1	0,00234839	0,00190057	0,55380645
		MID	0,69931257	0,11127921	0,08853653	0,18981799
		RIGHT	0,75069644	0,10937052	0,17417919	0,26629801
	multiva3	LEFT	0,88665713	0,58755087	0,05284216	0,58392019
		MID	0,77913663	0,62828256	0,07964384	0,54787743
		RIGHT	0,78428648	0,52960464	0,13152084	0,67235327

Tabla 17. RMSE Normalizado para la variable "petal length" amputada en un 10% de los casos (Elaboración propia).

En la Tabla 18 se presentan los RMSE Normalizados para la variable "sepal length" imputada por los cuatro métodos luego de haber sido amputada en un 10% de los casos, en patrón multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 18, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

k-NN resultó ser el mejor método de imputación para imputar la variable "sepal length" luego de haber sido amputada bajo el supuesto MCAR en patrón multiva3.

k-Means resultó ser el mejor método de imputación para imputar la variable "sepal length" luego de haber sido amputada bajo los supuestos MAR y MNAR en patrón multiva3 y las tres distribuciones (LEFT, MID y RIGHT).

En síntesis, para la variable "sepal length" amputada en el 10% de los casos considerando 7 escenarios diferentes, en 1 k-NN resulta ser el mejor método de imputación, mientras que en los restantes 6, el mejor resulta ser k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva3		0,96267685	0,28780052	0,63112516	0,951523
MAR	multiva3	LEFT	0,69202608	0,68349713	0,42438338	0,70188644
		MID	0,69349375	0,40079528	0,12835961	0,53917912
		RIGHT	0,86159027	0,45905318	0,05390719	0,63290922
MNAR	multiva3	LEFT	0,86314739	0,17743357	0	0,26336704
		MID	0,54806846	0,48339825	0,31145686	0,75783474
		RIGHT	1	0,62754049	0,46842366	0,79369614

Tabla 18. RMSE Normalizado para la variable "sepal length" amputada en un 10% de los casos (Elaboración propia).

En la Tabla 19 se presentan los RMSE Normalizados para la variable "petal width" imputada por los cuatro métodos luego de haber sido amputada en un 15% de los casos, en patrones univa, multiva2 y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 19, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, se puede observar que k-NN resultó ser el mejor método de imputación para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MCAR en patrones univa, multiva2 y multiva3.

k-NN resultó ser el mejor método de imputación para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MAR en patrón univa y distribuciones LEFT, MID y RIGHT.

De igual manera, k-NN resultó ser el mejor método de imputación para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MAR en patrón multiva2 y distribuciones LEFT y MID. Mientras en el caso de una distribución RIGHT, el mejor método de imputación resulto ser k-Means.

k-Means resultó ser el mejor método de imputación para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MAR en patrón multiva3 y distribuciones LEFT, MID y RIGHT.

En el escenario en el que la variable "petal width" fue amputada bajo el supuesto MNAR en patrón univa y distribuciones LEFT, MID y RIGHT, el mejor método de imputación resultó ser k-NN.

De igual manera, k-NN resulto ser el mejor método para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MNAR en patrón multiva2 y distribuciones LEFT y MID. Sin embargo, en el caso de una distribución RIGHT, el mejor método fue k-Means.

Finalmente, k-Means resultó ser el mejor método para imputar la variable "petal width" luego de haber sido amputada bajo el supuesto MNAR en patrón multiva3 y las tres distribuciones (LEFT, MID y RIGHT).

En síntesis, para la variable "petal width" amputada en el 15% de los casos considerando 21 escenarios diferentes, en 13 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 8, el mejor resultó ser k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	univa		0,77835032	0,1179143	0,13961801	0,3488149
	multiva2		0,80973696	0,16028553	0,18740266	0,73101799
	multiva3		0,85518286	0,01118316	0,05856124	0,32147328
MAR	univa	LEFT	0,93988158	0,04583494	0,08426746	0,29528081
		MID	0,62630944	0,07503213	0,10254924	0,39778511
		RIGHT	0,7545001	0,11026919	0,11213723	0,47120221
	multiva2	LEFT	0,86669994	0,02952343	0,07279406	0,32698391
		MID	0,5928697	0,14343603	0,16186548	0,46115334
		RIGHT	0,59886896	0,18596782	0,17275153	0,37266391
	multiva3	LEFT	0,9076779	0,51463217	0,09415319	0,24469605
		MID	0,79221251	0,50549688	0,07693362	0,38555193
		RIGHT	0,90734577	0,59444549	0,02432375	0,4706518
MNAR	univa	LEFT	0,88598978	0,08702331	0,11389822	0,20482655
		MID	0,60436672	0,06985995	0,08924519	0,41224511
		RIGHT	0,82413709	0,15889606	0,16934504	0,51700934
	multiva2	LEFT	0,84710795	0,02972178	0,06251708	0,2021211
		MID	0,5928697	0,14343603	0,15167249	0,46115334
		RIGHT	0,59886896	0,18596782	0,15630992	0,45467886
	multiva3	LEFT	0,95700379	0,50961756	0,02829566	0,18204937
		MID	0,77996841	0,54672461	0,10187848	0,51385025
		RIGHT	0,83315523	0,62866784	0,09894464	0,53931329

Tabla 19. RMSE Normalizado para la variable "petal width" amputada en un 15% de los casos (Elaboración propia).

En la Tabla 20 se presentan los RMSE Normalizados para la variable “petal length” imputada por los cuatro métodos luego de haber sido amputada en un 15% de los casos, en patrones multiva2, y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 20, el desempeño de los métodos k-NN y k-Means resulta mejor en todos los escenarios.

Entre estos dos, se puede observar que k-Means resulta ser el mejor método para imputar la variable “petal length” luego de haber sido amputada bajo el supuesto MCAR en patrón multiva2. Sin embargo, bajo el supuesto MCAR multiva3, k-NN resulta ser el mejor método para imputar la misma variable.

k-NN resulta ser el mejor método para imputar la variable “petal length” amputada bajo el supuesto MAR en patrón multiva2 y distribuciones LEFT, MID y RIGHT. Por el contrario, bajo el supuesto MAR en patrón multiva3 y distribuciones LEFT, MID y RIGHT el mejor método resulta ser k-Means.

En el escenario para el cual la variable “petal width” fue amputada bajo el supuesto MNAR en patrón multiva2 y considerando las tres distribuciones

(LEFT, MID y RIGHT), k-NN resulta ser el mejor método para imputar la variable "petal length". Por el contrario, en el caso de la variable "petal length" amputada bajo el supuesto MNAR en patrón multiva3 y considerando las tres distribuciones (LEFT, MID y RIGHT), k-Means resulta ser el mejor método de imputación.

En síntesis, para la variable "petal length" amputada en el 15% de los casos y considerando 14 escenarios de amputación diferentes, en 7 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 7, el mejor fue k-Means.

Mecanismo	Patrón	Tipo	Metodo de Imputacion			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva2		0,7279858	0,14668649	0,12096958	0,23364001
	multiva3		0,85989237	0,13579155	0,27621055	0,18424163
MAR	multiva2	LEFT	0,94070894	0,06416539	0,09890021	0,19723319
	multiva2	MID	0,68696594	0,11424936	0,13869443	0,31027108
	multiva2	RIGHT	0,80606481	0,09991263	0,14688797	0,27558808
	multiva3	LEFT	0,89958984	0,64158782	0,07595015	0,80828026
	multiva3	MID	0,81811497	0,5955673	0,05425709	0,56691194
	multiva3	RIGHT	0,77350959	0,51912041	0,07191773	0,32201641
MNAR	multiva2	LEFT	0,95531552	0,01960126	0,0265806	0,13437861
	multiva2	MID	0,68696594	0,11424936	0,14683808	0,31027108
	multiva2	RIGHT	0,76139095	0,09808885	0,15344285	0,25585319
	multiva3	LEFT	0,88265835	0,6491507	0,03743862	0,59634313
	multiva3	MID	0,75999236	0,53092479	0,08372365	0,50419317
	multiva3	RIGHT	0,7767831	0,52513112	0,08403283	0,65723355

Tabla 20. RMSE Normalizado para la variable "petallength" amputada en un 15% de los casos (Elaboración propia).

En la Tabla 21 se presentan los RMSE Normalizados para la variable "sepal length" imputada por los cuatro métodos luego de haber sido amputada en un 15% de los casos, en patrón multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 21, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, se puede observar que k-NN resultó ser el mejor método para imputar la variable "sepal length" luego de haber sido amputada bajo el supuesto MCAR en patrón multiva3.

k-Means resultó ser el mejor método para imputar la variable "sepal length" amputada bajo el supuesto MAR en patrón multiva3 y distribuciones LEFT, MID y RIGHT.

En el escenario para el cual la variable "sepal length" fue amputada bajo el supuesto MNAR en patrón multiva3 y considerando las tres distribuciones (LEFT, MID y RIGHT), k-Means resultó ser el mejor método para imputar la variable "sepal length".

En síntesis, para la variable “sepal length” amputada en el 15% de los casos y considerando 7 escenarios de amputación diferentes, en 1 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 6, el mejor fue k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva3		0,96131442	0,31863667	0,69058248	0,59179291
MAR		LEFT	0,61091618	0,36184335	0,08441262	0,22311674
		MID	0,60275484	0,37646447	0,25702447	0,76692417
		RIGHT	0,79929522	0,43178739	0,32855888	0,3192468
MNAR		LEFT	0,84554641	0,33225044	0,01374722	0,22997079
		MID	0,78145648	0,45196432	0,41363636	0,52666617
		RIGHT	0,86402288	0,5503732	0,39932499	0,64273181

Tabla 21. RMSE Normalizado para la variable "sepal length"amputada en un 15% de los casos (Elaboración propia).

En la Tabla 22 se presentan los RMSE Normalizados para la variable “petal width” imputada por los cuatro métodos luego de haber sido amputada en un 15% de los casos, en patrones univa, multiva2, y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 22, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, se puede observar que k-NN resultó ser el mejor método para imputar la variable “petal width” luego de haber sido amputada bajo el supuesto MCAR en patrones univa y multiva2. Sin embargo, bajo el supuesto MCAR en patrón multiva3, k-Means resultó ser el mejor método para imputar la misma variable.

k-NN resultó ser el mejor método para imputar la variable “petal witdh” amputada bajo el supuesto MAR en patrón univa y distribuciones LEFT, MID y RIGHT.

Para la variable “petal width” amputada bajo el supuesto MAR en patrón multiva2 y distribución LEFT, k-NN resulto ser el mejor método de imputación. Sin embargo, en el caso de las distribuciones MID y RIGHT, k-Means fue el mejor.

En el escenario para el cual la variable “petal width” fue amputada bajo el supuesto MNAR en patrón univa y distribuciones LEFT y MID, k-NN resulto ser el mejor método de imputación. Sin embargo, en el caso de la distribución RIGHT, k-Means fue el mejor.

k-NN k-NN resultó ser el mejor método para imputar la variable “petal witdh” amputada bajo el supuesto MNAR en patrón multiva2 y distribución LEFT. Por el contrario, en el caso de distribuciones MID y RIGHT, k-Means fue el mejor.

Finalmente, k-Means resulto ser el mjor método para imputar la variable "petal width" amputada bajo el supuesto MNAR en patrón multiva3 y distribuciones LEFT, MID y RIGHT.

En síntesis, para la variable "petal width" amputada en el 20% de los casos y considerando 21 escenarios de amputación diferentes, en 9 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 12, el mejor fue k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	univa		0,7743331	0,09941557	0,1305845	0,39945835
	multiva2		0,81343461	0,13846126	0,17693618	0,41568761
	multiva3		0,85163949	0,37008113	0,03379391	0,24038926
MAR	univa	LEFT	0,99090933	0,04384628	0,06820748	0,13635829
		MID	0,63012148	0,08716159	0,1132467	0,33278071
		RIGHT	0,77591318	0,12905616	0,13216857	0,51378967
	multiva2	LEFT	0,85194359	0,07447046	0,10443274	0,12932576
		MID	0,57355434	0,16672639	0,14666537	0,32758594
		RIGHT	0,67846667	0,16837745	0,13896258	0,38905183
	multiva3	LEFT	0,9098394	0,50889795	0,0528604	0,49581036
		MID	0,80088541	0,53339994	0,03482094	0,59444659
		RIGHT	0,83119832	0,5672719	0,11786589	0,56139763
MNAR	univa	LEFT	0,93161031	0,07027969	0,09034076	0,13698723
		MID	0,58620043	0,08365409	0,10821017	0,37565844
		RIGHT	0,84082383	0,14432831	0,11357056	0,463996
	multiva2	LEFT	0,83893458	0,03292327	0,04576797	0,15518381
		MID	0,57355434	0,16672639	0,14380598	0,32758594
		RIGHT	0,67846667	0,16837745	0,13703359	0,38905183
	multiva3	LEFT	1	0,52048073	0,01987839	0,44626786
		MID	0,75642891	0,49516849	0,07783577	0,46859386
		RIGHT	0,72635627	0,57113683	0,05350715	0,4568454

Tabla 22. RMSE Normalizado para la variable "petal width" amputada en el 20% de los casos (Elaboración propia).

En la Tabla 23 se presentan los RMSE Normalizados para la variable "petal length" imputada por los cuatro métodos luego de haber sido amputada en un 20% de los casos, en patrones multiva2, y multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 23, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, se puede observar que k-NN resultó ser el mejor método para imputar la variable "petal length" luego de haber sido amputada bajo el supuesto MCAR en patrón multiva2. Sin embargo, bajo el supuesto MCAR multiva3, k-Means resultó ser el mejor método para imputar la misma variable.

k-NN resultó ser el mejor método para imputar la variable “petal length” amputada bajo el supuesto MAR en patrón multiva2 y distribuciones LEFT, MID y RIGHT. Por el contrario, en patrón multiva3 el mejor método resultó ser k-Means.

En el escenario para el cual la variable “petal length” fue amputada bajo el supuesto MNAR en patrón multiva2 y las tres distribuciones (LEFT, MID y RIGHT), k-NN resultó ser el mejor método para imputar la variable “petal length”. Por el contrario, en el caso de la variable “petal length” amputada bajo el supuesto MNAR en patrón multiva3 y considerando las tres distribuciones (LEFT, MID y RIGHT), k-Means resultó ser el mejor método de imputación

En síntesis, para la variable “petal length” amputada en el 20% de los casos y considerando 14 escenarios de amputación diferentes, en 7 k-NN resultó ser el mejor método de imputación, mientras que en los restantes 7, el mejor fue k-Means.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva2		0,76698793	0,12471592	0,14837319	0,15658519
	multiva3		0,85337414	0,53005628	0,12630684	0,59106871
MAR	multiva2	LEFT	0,88029835	0,07265413	0,10660547	0,1502855
		MID	0,69513281	0,10448391	0,11130602	0,29575288
		RIGHT	0,80855585	0,1079655	0,12502372	0,28516482
	multiva3	LEFT	0,9210355	0,63251057	0,06755268	0,48465703
		MID	0,75041298	0,51823401	0,05931485	0,22780408
		RIGHT	0,8323223	0,5082711	0,08783441	0,40667175
MNAR	multiva2	LEFT	0,86191171	0,10457076	0,11593148	0,13953047
		MID	0,69513281	0,10448391	0,12289348	0,29575288
		RIGHT	0,805192	0,1076003	0,11499694	0,24453267
	multiva3	LEFT	0,99686316	0,6044701	0,05754435	0,50591629
		MID	0,72372946	0,48572968	0,04476805	0,37988795
		RIGHT	0,79392054	0,47857717	0,06667222	0,2629921

Tabla 23. RMSE Normalizado para la variable "petal length" amputada en el 20% de los casos (Elaboración propia).

En la Tabla 24 se presentan los RMSE Normalizados para la variable “sepal length” imputada por los cuatro métodos luego de haber sido amputada en un 20% de los casos, en patrón multiva3, bajo los supuestos MCAR, MAR y MNAR.

Como se puede apreciar en la Tabla 24, el desempeño de los métodos k-NN y k-Means resultaron ser los mejores en todos los escenarios.

Entre estos dos, se puede observar que k-Means resultó ser el mejor método para imputar la variable “sepal length” luego de haber sido amputada bajo los supuestos MCAR, MAR y MNAR en patrón multiva3 y las tres distribuciones consideradas para MAR y MNAR.

En síntesis, para la variable “sepal length” amputada en el 20% de los casos y considerando 7 escenarios de amputación diferentes, en los 7 k-Means resultó ser el mejor método de imputación.

Mecanismo	Patrón	Tipo	Método de Imputación			
			Media	k-NN	k-Means	hot-deck
MCAR	multiva3		0,98409611	0,56357827	0,32385419	0,49447736
MAR		LEFT	0,8112088	0,50159585	0,20247366	0,62906561
		MID	0,64458597	0,38689031	0,36470696	0,68342483
		RIGHT	0,8404708	0,46907145	0,30878395	0,54784251
MNAR		LEFT	0,95711389	0,4607823	0,07687999	0,49929798
		MID	0,71626347	0,47420538	0,29790044	0,66271565
		RIGHT	0,91448684	0,51303383	0,47931488	0,71789921

Tabla 24. RMSE Normalizado para la variable "sepallength" amputada en el 20% de los casos (Elaboración propia).

El gráfico de barras de la Figura 3, resume la cantidad de veces que los métodos de imputación k-NN y k-Means resultaron los mejores para imputar las variables “petal width”, “petal length” y “sepal length” en la totalidad de imputaciones realizadas.

En la Figura 3 se observa que el método de imputación k-NN resultó ser el mejor para imputar la variable “petal width” en 32 conjuntos de imputaciones realizadas, mientras que k-Means resultó el mejor en 27.

En el caso de la variable “petal length”, k-Means resultó ser el mejor en la mayoría de los casos, 24 conjuntos de imputaciones, mientras que k-NN resultó ser el mejor en 18.

Así mismo, k-Means resultó ser el mejor método para imputar la variable “sepal length” en 19 conjuntos de imputaciones mientras que, k-NN resultó ser el mejor en 2.

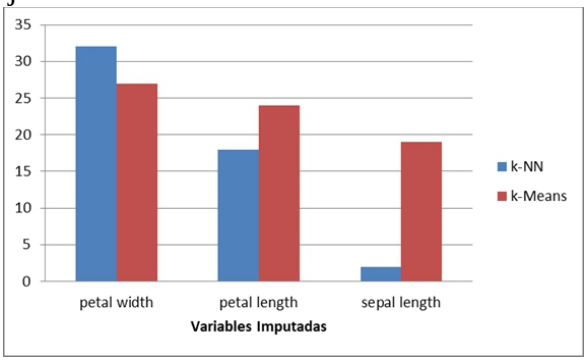

Figura 3. Cantidad de veces que los métodos de imputación k-NN y k-Means resultaron los mejores para imputar las variables “petal width”, “petal length” y “sepal length” en la totalidad de imputaciones realizadas (Elaboración Propia).

En la Tabla 25 se presenta el Promedio de los RMSE Normalizados para las variables “petal width”, “petal length” y “sepal length” del conjunto de datos “Iris”, imputadas por los métodos de imputación por medias, k-NN, k-Means y hot-deck luego de haber sido amputadas por los 63 procedimientos

definidos. Las filas en la tabla indican las variables imputadas y las columnas los métodos de imputación utilizados. Los valores en negrita indican el método con mejor desempeño en la totalidad de los conjuntos de datos amputados. Como se puede observar en la Tabla 25, el Promedio de RMSE Normalizados para k-Means es el más bajo para todas las variables imputadas considerando la totalidad de los ensayos realizados.

Variable	Método de Imputación			
	media	k-NN	k-Means	hot deck
petal width	0,78377057	0,23427899	**0,10121494**	0,39096886
petal length	0,81475313	0,31168774	**0,10057746**	0,3763713
sepal length	0,80735881	0,44342836	**0,29804081**	0,57978896

Tabla 25. Promedio de los RMS

Paso 4: Finalmente, se calcula el promedio de errores producidos por el método m_s sobre todas las variables Y_j que da un valor representativo del error producido por cada método de imputación para la totalidad de conjuntos de datos imputados. Errores bajos, indican mejor desempeño del método de imputación:

$$E(m_s) = \frac{\sum_{j=1}^{p} PRMSEN\left(Y_j^{m_s}\right)}{p};\ con\ s = \{1, \dots, q\}$$

Se puede sintetizar en la siguiente Tabla 6.

E Normalizados (Elaboración propia).

Finalmente, en la Tabla 26, se presenta el Promedio de los Errores producidos por los métodos de imputación medias, k-NN, k-Means y hot-deck y sobre las variables "petal width", "petal length" y "sepal length" del conjunto de datos "Iris". Como se puede observar, el error para el método de imputación k-Means es el más bajo considerando la totalidad de conjuntos de datos.

	media	k-NN	k-Means	hot deck
Error del Método de Imputación	0,80196084	0,32979837	**0,16661107**	0,44904304

Tabla 26. Error promedio de cada método de imputación para la totalidad de conjuntos de datos imputados (Elaboración propia).

Conclusiones

En este trabajo se ha presentado una metodología que incluye una propuesta de un nuevo indicador de desempeño de los métodos de imputación, basado en la conocida métrica RMSE.

El entorno de trabajo implementado para realizar los experimentos de amputación y posterior imputación resultó apropiado, ya que ha permitido parametrizar de manera sencilla los procedimientos de amputación y los métodos de imputación utilizados, como así también ha facilitado la gestión de los respectivos archivos originales, amputados e imputados.

Además, el entorno permite la incorporación a futuro de otros procedimientos de amputación y otros métodos de imputación, siendo parte esencial de la metodología propuesta.

La metodología propuesta y el indicar presentado, han permitido llegar a un valor global (ya que tiene en cuenta todas las variables que fueron amputadas y luego imputadas por varios métodos), indicativo del desempeño de cada método de imputación, expresado en valores comparables (normalizados), integrando los resultados de multitud de ensayos representativos de diferentes escenarios, con distintos porcentajes, diversidad de patrones, considerando además los tres mecanismos más frecuentes de aparición de datos faltantes.

Esta metodología permite también la incorporación de otras métricas e indicadores de desempeño de los métodos de imputación considerados en la misma, por lo que resulta flexible en cuanto a su aplicación.

Líneas futuras de trabajo

Con el propósito de ampliar los alcances de la metodología propuesta, se tiene previsto desarrollar nuevas métricas e indicadores utilizando algoritmos de minería de datos aplicados sobre los archivos completos y luego sobre los archivos imputados por diferentes métodos luego de haber sido amputados por diferentes mecanismos.

Agradecimientos

El presente trabajo se ha desarrollado en el contexto del PI código SIUTIRE0005231TC, de la Facultad Regional Resistencia de la UTN, correspondiendo hacer un significativo agradecimiento al Codirector de dicho proyecto, Dr. Marcelo Karanik, por su destacado aporte a la definición y validación de los alcances del software desarrollado, y a los becarios del mencionado proyecto, alumnos Matías R. Jaime y Nicolás F. Mussin, por el esfuerzo y dedicación brindados en el desarrollo del software.

También un agradecimiento especial al Mgter. Julio C. Acosta, por el apoyo brindado durante la formulación matemática utilizada en este trabajo.

References:

1. Ben-Gal, I. (2005). Outlier detection. In Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers.
2. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1–22.
3. Enders, C. K. (2010). Applied Missing Data Analysis. The Guilford Press.

4. Foundation, P. S. (2020). rpy2. https://rpy2.readthedocs.io/en/latest/
5. Gajawada, S., & Toshniwal, D. (2012). Missing Value Imputation Method Based on Clustering and Nearest Neighbours. International Journal of Future Computer and Communication, 1(2), 206–208.
6. García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), 263–282.
7. Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Systems with Applications, 89, 52–65.
8. Iris Data Set. (2020). https://archive.ics.uci.edu/ml/datasets/iris
9. Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), 913–933.
10. Jerez, J. M., Molina, I., García-Laencina, E. A., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. In Artificial Intelligence in Medicine (Vol. 50, pp. 105–115).
11. Liu, Y., & Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. Data, 2(1).
12. Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. In Knowledge and Information Systems (Vol. 32, Issue 1).
13. Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology, 110, 63–73.
14. missingpy 0.2.0. (2020). https://pypi.org/project/missingpy/
15. Multiple Hot-Deck Imputation. (2020). https://cran.r-project.org/web/packages/hot.deck/hot.deck.pdf
16. Multivariate Imputation by Chained Equations. (2020). https://cran.r-project.org/web/packages/mice/index.html
17. Nadzurah, Z. A., Amelia Ritahani, I., & Nurul, A. (2018). Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. International Journal of Advanced Computer Science and Applications, 9.
18. Pigott, T. D. (2001). A review of methods for missing data. International Journal of Phytoremediation, 21(1), 353–383.
19. Python. (2020). https://www.python.org/
20. Rahman, M. M., & Davis, D. N. (n.d.). Machine Learning-Based Missing Value Imputation Method for Clinical Datasets.

21. Rubin, D. B. (1976). Inference and missing data. In Biometrika (Vol. 63, Issue 3, pp. 581–592). https://doi.org/10.1093/biomet/63.3.581
22. Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons.
23. Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J. P., Santos, J., & Abreu, P. H. (2019). Generating synthetic missing data: A review by missing mechanism. IEEE Access, 7, 11651–11667. https://doi.org/10.1109/ACCESS.2019.2891360
24. Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.
25. Schafer, Josepn L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147
26. Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation, 88(15), 2909–2930.
27. scikit-learn. (2007). https://scikit-learn.org/stable/index.html
28. Seo, S. (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Set.
29. sklearn.cluster.KMeans. (2020). https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans
30. sklearn.impute.SimpleImputer. (2020). https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html#sklearn.impute.SimpleImputer
31. The R Project for Statistical Computing. (2020). https://www.r-project.org/about.html
32. Tobias, O. (2017). Performance of Imputation Algorithms on Artificially Produced Missing at Random Data. ProQuest Dissertations and Theses, 129.
33. Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence, 23(5), 373–405.
34. Twala, B., Cartwright, M., & Shepperd, M. (2006). Ensemble of missing data techniques to improve software prediction accuracy. Proceedings - International Conference on Software Engineering, 2006, 909–912.
35. Van Buuren, S. (2012). Flexible Imputation of Missing Data, 2nd ed. Taylor & Francis Group, LLC.
36. van der Meijs, A. (2018). Missing Data Imputation: Predicting Missing Values (Issue July). Tilburg University.