Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Trabajo Final de Maestría en Tecnologías de la Información

Modelo de Decisión para la Validación de Métodos de Imputación Mediante la Utilización de Algoritmos de Minería de Datos

Autor: Lic. Carlos Roberto Primorac

Director: Mgter. Julio César Acosta

Codirector: Dr. David Luís la Red Martínez

Corrientes, Argentina, 2022

Agradecimientos

A David, a Marcelo y Julio. A mis amigos, especialmente a Viviana y a Sergio. A Fer, Martín, Ani y Claudio.

Dedicatoria

A mis padres, hermanos y sobrinos.

Resumen

Muchos de los conjuntos de datos existentes u obtenidos en investigaciones científicas contienen valores faltantes y otras anomalías originadas en diferentes causas. En minería de datos, estas imperfecciones pueden afectar negativamente la calidad del proceso de aprendizaje supervisado o el rendimiento de algoritmos de agrupamiento de datos. La imputación es una técnica para reemplazar valores faltantes con valores calculados utilizando los datos existentes. Se desarrolló una metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador y un entorno para realizar los experimentos de amputación y posterior imputación. Además se trabajó en encontrar los métodos de imputación más adecuados para completar los valores faltantes en un conjunto de datos mediante la utilización de algoritmos de minería de datos. En todos los escenarios evaluados, los métodos más apropiados resultaron ser k-NN y K-Means.

Palabras Clave: *valores faltantes, amputación de datos, imputación de datos, minería de datos, evaluación de desempeño de métodos de imputación.*

Abstract

Many existing data sets or data obtained in scientific research contain missing values and other anomalies originating from different causes. In data mining, these imperfections can negatively affect the quality of the supervised learning process or the performance of data clustering algorithms. Imputation is a technique to replace missing values with values calculated using existing data. A methodology for evaluating the performance of imputation methods was developed using a traditional metric complemented with a new indicator and an environment for performing amputation and subsequent imputation experiments. In addition, work was done to find the most suitable imputation methods to fill in missing values in a dataset using data mining algorithms. In all scenarios evaluated, the most appropriate methods proved to be k-NN and K-Means.

Key Words: *missing values, data amputation, data imputation, data mining, imputation methods performance evaluation.*

Reconocimientos

El presente trabajo se ha desarrollado en el contexto del PI código SIUTIRE0005231TC, de la Facultad Regional Resistencia de la Universidad Tecnológica Nacional (FRRe-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 las revisiones a los trabajos publicados, 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 y al becario, alumno Alejandro Nadal, por el esfuerzo y dedicación brindados en los múltiples procesos de minería de datos.

Indice

Indice de figuras

Indice de tablas

Aclaración: todas las tablas y figuras donde no se indica referencia, son de elaboración propia.

Capítulo 1

Introducción

1.1. Introducción

Históricamente, la noción de descubrir patrones ocultos en los datos ha recibido una variedad de denominaciones incluidos el de minería de datos y descubrimiento del conocimiento (Fayyad et al., 1996).

La minería de datos se refiere a los medios algorítmicos mediante los cuales se extraen y enumeran patrones a partir de los datos, constituyendo una etapa dentro del proceso general de descubrimiento del conocimiento (La Red Martínez et al., 2016).

Los valores faltantes, valores no disponibles en una o más variables en el registro de interés, constituyen un obstáculo común que enfrentan investigadores en contextos del mundo real (Luengo et al., 2012).

Muchos de los conjuntos de datos existentes u obtenidos en investigaciones científicas, incluyendo medicina y biomedicina, informática, ingeniería, psicología, sociología y educación entre otras, contienen valores faltantes y otras anomalías[1] asociadas a procedimientos de entrada manuales deficientes, mediciones incorrectas o errores en los instrumentos de medición (Twala et al., 2005), (Farhangfar et al., 2007), (Jerez et al., 2010), (Enders, 2010), (Madhu & Rajinikanth, 2012), (Luengo et al., 2012), (Liu & Gopalakrishnan, 2017).

La presencia de estas imperfecciones generalmente requiere de una etapa de preprocesamiento en la cual, con el fin de que resulten útiles y suficientemente claros para el proceso de extracción de conocimiento, los datos se deben preparar y limpiar (M. G. Rahman & Islam, 2010), (Luengo et al., 2012), (Pattanodom et al., 2016).

En minería de datos se pueden encontrar tres problemas principales asociados a valores faltantes y anomalías[2]: i) pérdida de eficiencia de los algoritmos, ii) dificultad en la manipulación y el análisis de los datos y iii) sesgo resultante de las diferencias entre valores faltantes y completos (Luengo et al., 2012), (Pattanodom et al., 2016).

En la literatura se proponen dos enfoques generales para enfrentar a los problemas asociados a los valores faltantes. El caso más simple consiste en ignorarlos. Una segunda alternativa consiste en utilizar técnicas de imputación y reemplazarlos con valores calculados utilizando los datos existentes (Twala et al., 2005), (Farhangfar et al., 2007),

[1] Conocidos en inglés como *outliers*, valores extremos que se desvían de otros valores en el conjunto de datos.

[2] Las anomalías generalmente son eliminadas constituyendo entonces valores faltantes.

(Vergouwe et al., 2010), (Madhu & Rajinikanth, 2012), (Aljuaid & Sasi, 2016), (Liu & Gopalakrishnan, 2017).

Tradicionalmente, el tratamiento de los valores faltantes se realizaba antes del análisis de los datos mediante métodos diseñados "had-hoc" (Peugh & Enders, 2004).

Algunas de estas estrategias consistían en trabajar con información completa, eliminando todos los casos con valores faltantes en una o más variables (listwise-deletion), considerando únicamente los registros con valores completos en la variable de análisis (pairwise-deletion) o sustituyendo los valores faltantes con el promedio de valores de la variable considerada (Van Buuren, 2012).

Sin embargo, el sesgo introducido por estas técnicas ha hecho que sean fuertemente criticadas en la literatura (Peugh & Enders, 2004).

La imputación es una técnica para reemplazar valores faltantes con valores calculados. Una característica importante para una instancia en particular puede imputarse (Aljuaid & Sasi, 2016). Estos métodos utilizan diferentes algoritmos que se pueden dividir en imputación simple e imputación múltiple (tal como se indica en el Capítulo 2); en los últimos años se ha propuesto el uso de algoritmos de aprendizaje automático (Peugh & Enders, 2004), (Twala et al., 2005), (Farhangfar et al., 2007), (Pantanowitz & Marwala, 2008), (Luengo et al., 2012), (Arima et al., 2014), (Tutz & Ramzan, 2015), (Pattanodom et al., 2016), (Liu & Gopalakrishnan, 2017).

El mecanismo de pérdida de datos es un factor clave para decidir el método de imputación a utilizar (Aljuaid & Sasi, 2016).

En (Rubin, 1976) se definieron tres mecanismos por los cuales se genera la pérdida de datos: i) valores faltantes completamente aleatorios (MCAR: Missing Completely At Random), ii) valores faltantes aleatorios (MAR: Missing At Random) y iii) valores faltantes no aleatorios (MNAR: Missing Not At Random).

Asimismo, el tamaño del conjunto de datos y el porcentaje de valores faltantes en el mismo, influye en la elección de un método de imputación (Twala et al., 2005).

Por otro lado, el desempeño de un método de imputación depende, además, de los patrones de valores faltantes (Aljuaid & Sasi, 2016).

El patrón de valores faltantes, se refiere a la disposición de los mismos en el conjunto de datos (Enders, 2010).

Existen diferentes patrones de valores faltantes (como se detalla en el Capítulo 2), algunos asociados con el registro y otros con el atributo (Aljuaid & Sasi, 2016). En el primer caso

pueden ser simples, complejos, medios y mixto. En el segundo univariados, monótonos y arbitrarios (Aljuaid & Sasi, 2016).

Los distintos métodos de imputación funcionan sobre diferentes tipos de datos, algunos trabajan bien con tipos de datos numéricos (valores enteros o reales) y otros únicamente con variables categóricas (valores que representan categorías o grupos mutuamente excluyentes) (M. G. Rahman & Islam, 2010).

Finalmente, algoritmos de agrupamiento (no supervisados) y de clasificación (supervisados) se pueden adaptar también para la imputación (Liu & Gopalakrishnan, 2017).

La mayoría de los artículos publicados en este campo se ocupan del desarrollo de nuevos métodos de imputación, sin embargo, pocos estudios informan una evaluación global de los métodos existentes con el fin de proporcionar directrices para hacer la elección metodológica más apropiada en la práctica (Schmitt et al., 2015).

1.2. Objetivo

El propósito general de este trabajo final de maestría consistió en encontrar y determinar una metodología para seleccionar los métodos de imputación más adecuados para completar valores faltantes en un conjunto de datos mediante la utilización de nuevas métricas que utilizan error de la raíz cuadrada media[3] y otras basadas en algoritmos de minería de datos.

Se utilizó como criterio de validación, el criterio de mayor similitud entre los resultados de procesos de minería de datos antes de la imputación (es decir, considerando solamente registros con valores completos y excluyendo aquellos con valores faltantes) y luego de la aplicación de cada uno de los métodos de imputación utilizados (es decir, incluyendo ahora los archivos con los registros con valores imputados), para lo cual se plantearon los siguientes objetivos específicos:

1. Definir las métricas a utilizar para la validación de los diferentes métodos de imputación aplicados al conjunto de datos.
2. Definir el procedimiento de selección de los métodos de imputación adecuados para ser aplicados al conjunto de datos.
3. Definir el orden de prioridad de las variables para la aplicación de los diferentes métodos de imputación de datos en el conjunto de datos.

[3] RMSE: Root Mean Squared Error.

1.3. Desarrollos realizados

Con el fin de obtener un modelo de decisión que proporcione directrices para hacer una elección metodológica apropiada en la práctica se ha realizado una evaluación de algunos de los métodos de imputación existentes utilizando algoritmos de minería de datos.

La estrategia utilizada consistió en aplicar distintos métodos de imputación de validez reconocida para el tipo de dato a imputar, y comparar luego la efectividad de cada método utilizando para la métrica de comparación los conjuntos de valores representativos de los datos analizados determinados por los procesos de minería de datos, ejecutados antes y después de realizar la imputación con un método en particular y para una variable específica.

Los procesos de minería de datos cuyos resultados se tomaron como referencia para la comparación se realizaron sobre un conjunto de datos completo y correcto, es decir, descartando los valores faltantes y anómalos.

Una vez determinado el método de minería de datos a utilizar para una variable en particular, los métodos de imputación aplicables al tipo de variable en cuestión y las métricas de comparación a utilizar, se obtuvieron los elementos de entrada para el modelo de decisión que se utilizó, el cual permitió seleccionar el método de imputación más adecuado para la variable en cuestión.

Estos pasos se repitieron para las distintas variables para las que se determinó el modelo de decisión, el cual permitió imputar datos de las mismas de la mejor manera posible en el contexto indicado.

El contenido de este trabajo se ha estructurado de la siguiente manera:

En el Capítulo 1, se presenta la problemática de los valores faltantes y se comentan los enfoques para enfrentarse a ellos. Seguidamente, se plantean los objetivos del trabajo final de maestría y se comentan los desarrollos realizados, finalizándose con la mención de las publicaciones relacionadas con el capítulo.

En el Capítulo 2, se comentan los conceptos fundamentales acerca de los valores faltantes y se mencionan las estrategias de imputación generales para enfrentarse a estos. Se describen el conjunto de datos utilizado, los procedimientos para la generación de valores faltantes, los métodos de imputación empleados y la medida de desempeño utilizada, como así también el software utilizado y el diseño de los experimentos efectuados. Seguidamente, se comentan y analizan los resultados obtenidos, considerándose la métrica de evaluación de desempeño propuesta de los métodos de imputación utilizados

concluyendo con unos breves comentarios y mencionando las publicaciones relacionadas con el capítulo.

En el Capítulo 3, se revisan el concepto de minería de datos, se introducen los principales algoritmos y métricas de evaluación de modelos, se describen los conjuntos de datos, el algoritmo de minería de datos y las métricas indicadoras de calidad utilizadas. A continuación, se comentan y comparan los resultados detalladamente, finalizándose con unos breves comentarios y mencionando las publicaciones relacionadas con el capítulo.

Finalmente, en el Capítulo 4, se presentan las conclusiones y se mencionan las futuras líneas de trabajo.

Publicación relacionada

Primorac, Carlos R.; Acosta, Julio César; La Red Martínez, David L. Modelo de decisión para la validación de MI mediante la utilización de algoritmos de minería de datos. *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018)*, Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste, 26 y 27 de abril de 2018.

Capítulo 2

Metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador

2.1. Introducción

El objetivo de este Capítulo es presentar una metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador, basado en el promedio normalizado del RMSE.

El Capítulo se organiza de la siguiente manera: en primer lugar se comentan los conceptos fundamentales acerca de los valores faltantes y se mencionan las estrategias de imputación generales para enfrentarse a estos. A continuación, se describen el conjunto de datos utilizado, los procedimientos para la generación de valores faltantes, los métodos de imputación empleados y la medida de desempeño utilizada, como así también el software utilizado y el diseño de los experimentos efectuados. Seguidamente, se comentan y analizan los resultados obtenidos, considerándose la métrica de evaluación de desempeño propuesta de los métodos de imputación utilizados. Finalmente, se exponen las conclusiones.

2.2. Conocimientos previos

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

1. Recopilación o simulación de conjuntos de datos completos.
2. Generación de conjuntos de datos con valores faltantes a partir de 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.

2.2.1. El modelo de datos

Un conjunto de datos ordenado 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 (Fig. 1) (J. L. Schafer, 1997), (Enders, 2010).

El conjunto de datos completo Y es hipotético, puesto que contiene valores faltantes, denotados por el símbolo "?", que pueden aparecer en diferentes proporciones, patrones y de acuerdo con algún mecanismo. En consecuencia el conjunto de datos Y, se puede considerar como formado por dos componentes (Y_{obs}, Y_{mis}). La componente Y_{obs} contiene

valores completos y la componente Y_{mis} contiene valores faltantes (J. L. Schafer, 1997), (Enders, 2010).

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 $r_{ij} = 0$ (Enders, 2010).

Los valores completos se denotan colectivamente por Y_{obs}, los valores faltantes 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 (Enders, 2010).

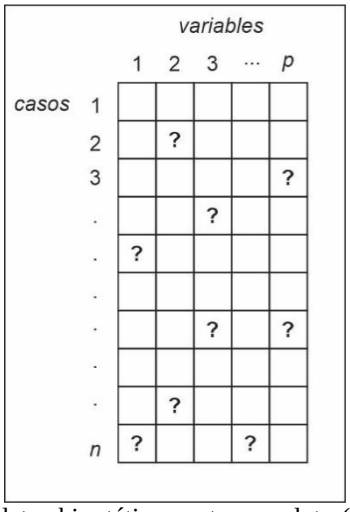

Fig. 1. Conjunto de datos hipotéticamente completo (J. L. Schafer, 1997).

2.2.2. Mecanismos de valores faltantes

En la teoría de (Rubin, 1976) se clasifican a los problemas de valores faltantes en tres categorías: MCAR, MAR y MNAR.

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 matriz indicadora de respuesta R indica los valores faltantes en el conjunto de datos Y. La distribución de probabilidades de la matriz indicadora de respuesta R depende de los valores de $Y = (Y_{obs}, Y_{mis})$, ya sea por diseño o por casualidad, y esta relación se describe mediante el modelo de valores faltantes (Van Buuren, 2012).

La expresión general del modelo de valores faltantes es:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi)$$

Dónde P es un símbolo genérico para la distribución de probabilidad, R es la matriz indicadora de respuesta, Y_{obs} e Y_{mis} representan respectivamente las componentes con valores completos y los valores faltantes en el conjunto de datos Y, ψ es el parámetro (o conjunto de parámetros) que describe las relaciones entre la matriz indicadora de respuesta R y el conjunto de datos Y (Van Buuren, 2012).

Se dice que los valores faltantes son MCAR si $P(R = 0|\psi)$. Es decir que, la probabilidad de $R = 0$, depende únicamente de algunos parámetros ψ, en este caso del parámetro de la distribución de probablidad de R (Van Buuren, 2012).

Se dice que los valores faltantes son MAR si $P(R = 0|Y_{obs}, \psi)$. Es decir, la probabilidad de $R = 0$, depende de valores observados a través de algunos parámetros ψ que relacionan a Y_{obs} con R (Van Buuren, 2012).

Finalmente, se dice que los valores faltantes son MNAR si $P(R = 0|Y_{obs}, Y_{mis}, \psi)$. En este caso, la probabilidad de $R = 0$ puede depender tanto de valores observados Y_{obs} como de valores no observados Y_{mis} a través de algunos parámetros ψ que los relacionan con R (Van Buuren, 2012).

En la práctica, en general no es posible determinar con certeza los parámetros ψ de la matriz indicadora de respuesta R. Sin embargo, no es importante conocerlos en detalle, solamente es necesario comprender si existe una relación entre R y los componentes de Y (Enders, 2010).

2.2.3. Patrones de valores faltantes

Es importante distinguir entre patrones y mecanismos de valores faltantes. 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 el valor de las variables y la probabilidad de un valor faltante (Van Buuren, 2012).

El patrón simplemente describe la localización de los valores faltantes en el conjunto de datos pero no explica las causas (Van Buuren, 2012).

La Fig. 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 valores faltantes.

El patrón univariado (Fig. 2a) presenta valores faltantes en una única variable. Es relativamente raro en algunas disciplinas, sin embargo, puede surgir en estudios experimentales (Enders, 2010), (Van Buuren, 2012).

Los patrones monótonos o escalonados (Fig. 2b), se asocian con estudios longitudinales en los cuales participantes abandonan la encuesta y no regresan (Enders, 2010), (Van Buuren, 2012).

Finalmente, el patrón multivariado o general (Fig. 2c) es quizás la configuración más común, los valores faltantes 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 preferencia a los datos faltantes en Y_2) (Enders, 2010), (Van Buuren, 2012).

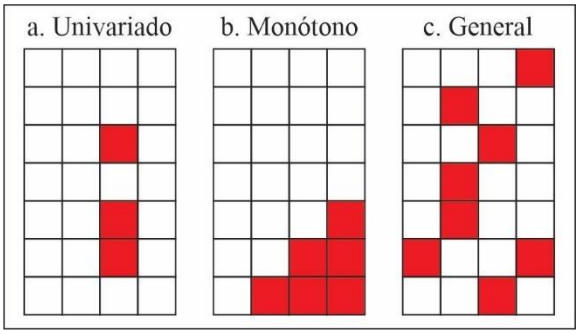

Fig. 2. Patrones típicos de valores faltantes en la literatura (Elaboración propia en base a (Enders, 2010) y (Van Buuren, 2012)).

2.2.4. Porcentaje de valores faltantes

El porcentaje o proporción de valores faltantes, conocidos como tasa de valores faltantes (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 tasas de valores faltantes 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).

2.2.5. Generación de valores faltantes

Un aspecto clave en la evaluación de los métodos de imputación de datos es la generación de valores faltantes (Santos et al., 2019).

El proceso por el cual se generan conjuntos de datos con valores faltantes 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 valores faltantes 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 valores faltantes en múltiples variables para cualquier conjunto de datos.

La amputación multivariada, facilita la definición de patrones y su ocurrencia relativa, permite manipular la tasa y ajustar con precisión los distintos mecanismos de valores faltantes (Schouten et al., 2018). Estas características han sido determinantes para elegir este método en los ensayos realizados que se describirán más adelante.

2.2.6. Enfoques para enfrentarse a los valores faltantes

Los enfoques tradicionales más utilizados que ignoran los valores faltantes son: a) el método de eliminación por listas (listwise deletion) y el método de eliminación por pares (pairwise deletion) (Jadhav et al., 2019).

En el método de eliminación por listas, el análisis se realiza con información completa, simplemente omitiendo todos aquellos casos con valores faltantes. La principal ventaja de este método es su sencillez, no obstante, reduce el tamaño de la muestra sobre todo si la tasa de valores es alta (Enders, 2010).

El método de eliminación por pares 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 los valores faltantes y constituyen los enfoques más populares para sustituirlos 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 media, por regresión, por regresión estocástica, hot-deck y cold-deck (Josepn L. Schafer & Graham, 2002).

La imputación por media, también conocida como sustitución por la media o imputación por media no condicionadas, es el método más simple para sustituir valores faltantes mediante el promedio aritmético de los valores completos para una variable (Josepn L. Schafer & Graham, 2002).

El método de imputación por regresión o imputación por media condicionadas reemplaza los valores faltantes 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 valores faltantes utilizando una ecuación de regresión que incluye un término de error residual normalmente distribuido para mejorar las predicciones (Josepn L. Schafer & Graham, 2002).

Los métodos de imputación hot-deck, consisten en una colección de técnicas que imputan los valores faltantes con valores de observaciones similares. En su forma más simple, reemplaza valores faltantes por valores calculados 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 (Josepn L. Schafer & Graham, 2002).

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 valores faltantes, 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 valores faltantes (Pigott, 2010).

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 maxima verosimilitud de parámetros en conjuntos de datos con valores faltantes. 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, 2010).

Propuestos por (Rubin, 1987), los métodos de imputación múltiple 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, 2010) (Tobias, 2017).

Constituye una estrategia robusta, puesto que requiere supuestos menos estrictos acerca del mecanismo de valores faltantes. Sin embargo, continúan siendo débiles, ya que en el caso de valores MNAR las estimaciones podrían estar sesgadas (Pigott, 2010) (Tobias, 2017).

En los últimos años, se ha propuesto el uso de algoritmos de aprendizaje automático como métodos de imputación (Liu & Gopalakrishnan, 2017).

Las técnicas de aprendizaje automático 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 aprendizaje automático bien conocidos tales como los árboles de decisión (DT: Decision Trees), k-vecinos más cercanos (k-NN: k-Nearest Neighbours), agrupamiento por k-Media (k-Means Clustering) y redes bayesianas (Bayesian Networks), han sido utilizados como métodos de imputaciónn en diferentes dominios (García-Laencina et al., 2010), (Jerez et al., 2010), (Luengo et al., 2012), (M. M. Rahman & Davis, 2013), (Liu & Gopalakrishnan, 2017) (Nadzurah et al., 2018).

2.3. Materiales y métodos

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.

2.3.1. 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 valores faltantes, 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 valores extremos que se desvían de otros valores en el conjunto de datos (Ben-Gal, 2005).

Se utilizó el método de la desviación estándar para cada variable y se eliminaron todos aquellos registros cuyos valores para la variable considerada estuvieron fuera del intervalo $\bar{x} \pm 2\sigma$ (Seo, 2006), obteniendo un conjunto de datos completo con 139 casos.

2.3.2. Generación de los conjuntos de datos con valores faltantes

Se adoptó el método de amputación multivariado descripto en (Schouten et al., 2018) implementado en la función "ampute" del paquete "mice" (Buuren et al., 2021) del software R (*The R Project for Statistical Computing*, n.d.).

Se definieron tres valores de tasa de valores faltantes (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 valores faltantes en una variable, un patrón multivariado simple (multiva2) tiene valores faltantes 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 faltantes según un mecanismo 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 faltantes según un mecanismo MCAR univa se seleccionó "petal width" como variable a ser amputada.

En la generación de valores faltantes según un mecanismo MCAR multiva2, se seleccionaron como variables a ser amputadas "petal width" y "petal length".

Finalmente, para la generación de valores faltantes según un mecanismo MCAR multiva3, se seleccionaron como variables a ser amputadas "petal width", petal length" y "sepal length".

La amputación de valores faltantes según un mecanismo 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 faltantes según un mecanismo MAR univa; {("petal width", "petal length"), ("petal length", "sepal length)} para la generación de valores faltantes según un

mecanismo MAR multiva2 y {("petal width", "petal length"), ("petal length", "sepal length"), ("sepal width", ("sepal length")} para la generación de valores faltantes según un mecanismo MAR multiva3.

La generación de valores faltantes según un mecanismo MNAR univa, MNAR multiva2 y MNAR multiva3 sigue un enfoque similar al propuesto para la generación de valores faltantes según un mecanismo 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 faltantes según un mecanismo MAR y MNAR los casos candidatos a contener valores faltantes reciben una probabilidad basada en un puntaje de suma ponderada (Schouten et al., 2018).

La asignación de estas probabilidades se realiza en base a una función con distribución logística. Los puntajes de suma ponderada se obtienen multiplicando cada variable por un peso elegido arbitrariamente. Para cada patrón definido, es posible ponderar cada variable de manera que gobierne el impacto de éstas en la formación de los puntajes de suma ponderada. Variables con pesos altos tendrán mayor influencia que variables con pesos bajos.

De esta manera, es posible generar valores faltantes según un mecanismo 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 faltantes según un mecanismo MNAR.

Para la generación valores faltantes según un mecanismo 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).

2.3.3. Imputación de los conjuntos de datos amputados

Se seleccionaron los métodos de imputación por Media, k-NN, k-Means y Hot-Deck.

Imputación por Media

Es una de las técnicas más simples para reemplazar valores faltantes en una variable. Consiste en calcular la media aritmética de los valores completos para la variable en cuestión y utilizarlo para imputar el valor faltante de esa variable (Jerez et al., 2010), (Aljuaid & Sasi, 2016).

Imputación mediante k-NN

Es un método de imputación basado en métodos de aprendizaje automático supervisados (Jerez et al., 2010), (Liu & Gopalakrishnan, 2017), (Nadzurah et al., 2018).

El método de imputación k-NN clasifica los datos en grupos y entonces reemplaza los valores faltantes en una variable con el valor correspondiente al vecino más cercano, es decir, el valor más próximo a en base a la distancia euclidea. Los valores faltantes se imputan teniendo en cuenta a un determinado número de registros que son, en su mayoría, similares al registro de interés. Esta similitud se determina utilizando la distancia euclidea (Aljuaid & Sasi, 2016).

Imputación Hot-Deck

La imputación Hot-Deck consiste en sustituir los valores faltantes (receptores) de una o varias variables con valores completos (donantes) son similares a los receptores con respecto a las características completas (Aljuaid & Sasi, 2016).

En algunas versiones, conocidas como *métodos hot deck aleatorios*, el donante se selecciona de manera aleatoria de un conjunto de potenciales donantes (Aljuaid & Sasi, 2016).

En otras versiones, conocidas como *métodos hot deck deterministas*, se identifica un donante único y los valores se imputan desde ese registro (por ejemplo, el vecino más cercando basado en alguna métrica). Otros métodos, utilizan como donantes valores como el promedio de un conjunto de donantes (Aljuaid & Sasi, 2016).

Imputación k-Means

Es un método de imputación basado en métodos de aprendizaje automático no supervisados (Jerez et al., 2010), (Liu & Gopalakrishnan, 2017).

El método de imputación k-Means el algoritmo de clasificación no supervisada que agrupa registros en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada registro y el centroide de su grupo o clúster. Los valores faltantes se completan en base a la información del clúster, aplicando el criterio del vecino más cercano (Gajawada & Toshniwal, 2012)

En este trabajo se utilizaron los lenguajes de programación R (*The R Project for Statistical Computing*, n.d.) y Python (*Python*, n.d.).

Se utilizó la librería "rpy2" (*rpy2 - R in Python*, n.d.) como interfaz entre ambos lenguajes para la ejecución de experimentos de amputación e imputación.

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

La clase "k-NNImputer", de la librería "missingpy" (*missingpy 0.2.0*, n.d.) de Python, soportó la imputación utilizando el enfoque k-NN. La función "hot.deck" del paquete "hot.deck-package" (Gill et al., 2017) del software R (*The R Project for Statistical Computing*, n.d.) 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*, n.d.) de la librería "scikit-learn" (*scikit-learn*, n.d.) de Python (*Python*, n.d.). En el primer paso, se calculó la cantidad óptima de clústeres. Seguidamente, la información de los clústeres obtenidos se utilizó para imputar los valores faltantes.

2.4. 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.

Tabla 1. Conjunto de datos completo Y.

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}

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.

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

$Y_1^{a_1 m_1}$	$Y_2^{a_r m_s}$	…	$Y_j^{a_r m_s}$	…	$Y_p^{a_r m_s}$
$y_{11}^{a_r m_s}$	$y_{12}^{a_r m_s}$	…	$y_{1j}^{a_r m_s}$	…	$y_{1p}^{a_r m_s}$
$y_{21}^{a_r m_s}$	$y_{22}^{a_r m_s}$	…	$y_{2j}^{a_r m_s}$	…	$y_{2p}^{a_r m_s}$
⋮	⋮	…	…	…	⋮
$y_{i1}^{a_r m_s}$	$y_{i2}^{a_r m_s}$	…	$y_{ij}^{a_r m_s}$	…	$y_{ip}^{a_r m_s}$
⋮	⋮	…	⋮	…	⋮
$y_{n1}^{a_r m_s}$	$y_{n2}^{a_r m_s}$	…	$y_{nj}^{a_r m_s}$	…	$y_{np}^{a_r m_s}$

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, según la ecuación (1). Estos resultados se pueden sintetizar en la Tabla 3.

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

Tabla 3. RMSE para la variable $Y_j^{a_r m_s}$.

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

Paso 2: Utilizando la ecuación (2), se normaliza el RMSE para cada variable $Y_j^{a_r m_s}$. Esto facilita comparar los valores de la RMSE para cada variable, dado que éstas pueden estar en diferentes escalas. Estos resultados se pueden sintetizar en la Tabla 4.

$$RMSEN\left(Y_j^{a_r m_s}\right) = \frac{RMSE\left(Y_j^{a_r m_s}\right) - min\{RMSE\left(Y_j^{a_t m_u}\right)\}}{max\left\{RMSE\left(Y_j^{a_t m_u}\right)\right\} - min\left\{RMSE\left(Y_j^{a_t m_u}\right)\right\}}; con \begin{matrix} 1 \le j \le p \\ 1 \le r \le l \\ 1 \le s \le q \\ 1 \le t \le l \\ 1 \le u \le q \end{matrix} \quad (2)$$

Tabla 4. RMSEN para cada variable $Y_j^{a_r m_s}$.

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

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, según la ecuación (3). Estos valores se pueden sintetizar en la Tabla 5.

$$PRMSEN\left(Y_j^{m_s}\right) = \frac{\sum_{r=1}^{l} RMSEN\left(Y_j^{a_r m_s}\right)}{l}; con \begin{matrix} 1 \le j \le p \\ 1 \le s \le q \end{matrix} \quad (3)$$

Tabla 5. PRMSEN 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.

	m_1	m_2	⋯	m_s	⋯	m_q
Y_1	$PRMSEN\left(Y_1^{m_1}\right)$	$PRMSEN\left(Y_1^{m_2}\right)$	⋮	$PRMSEN\left(Y_1^{m_s}\right)$	⋮	$PRMSEN\left(Y_1^{m_q}\right)$
Y_2	$PRMSEN\left(Y_2^{m_1}\right)$	$PRMSEN\left(Y_2^{m_2}\right)$	⋮	$PRMSEN\left(Y_2^{m_s}\right)$	⋮	$PRMSEN\left(Y_2^{m_q}\right)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Y_j	$PRMSEN\left(Y_j^{m_1}\right)$	$PRMSEN\left(Y_j^{m_2}\right)$	⋮	$PRMSEN\left(Y_j^{m_s}\right)$	⋮	$PRMSEN\left(Y_j^{m_q}\right)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Y_p	$PRMSEN\left(Y_p^{m_1}\right)$	$PRMSEN\left(Y_p^{m_2}\right)$	⋮	$PRMSEN\left(Y_p^{m_s}\right)$	⋮	$PRMSEN\left(Y_p^{m_q}\right)$

Paso 4: Finalmente, con la ecuación (4), 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. Se puede sintetizar en la Tabla 6. Errores bajos, indican mejor desempeño del método de imputación.

$$E(m_s) = \frac{\sum_{j=1}^{p} PRMSEN\left(Y_j^{m_s}\right)}{p};\ con\ s = \{1, \dots, q\} \quad (4)$$

Tabla 6. Promedio de errores producidos por el método m_s sobre todas las variables Y_j

(Elaboración propia).

m_1	m_2	$\cdots$	m_s	$\cdots$	m_q
$E(m_1)$	$E(m_2)$	$\cdots$	$E(m_s)$	$\cdots$	$E(m_q)$

2.5. Resultados y discusiones

A partir del conjunto de datos original sin valores extremos, se ejecutaron los procedimientos de amputación y se obtuvieron en total 63 conjuntos de datos con valores faltantes. De estos, 9 (3 patrones x 3 tasas de valores faltantes) corresponden a conjuntos de datos amputados bajo el supuesto MCAR, 27 (3 patrones x 3 tasas de valores faltantes x 3 tipos) a conjuntos de datos amputados bajo el supuesto MAR y 27 (3 patrones x 3 tasas de valores faltantes 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 Media, 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 Tabla 7 a Tabla 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 de imputación Media, k-NN, k-Means y Hot-Deck, luego de haber sido amputadas de acuerdo a los 63 procedimientos definidos.

Tabla 7. RMSE para la variable “petal width” imputada por Media, 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.

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 8. RMSE para las variables “petal width” y “petal length” imputadas por Media, 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.

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 9. RMSE para las variables “petal width”, “petal length” y “sepal length” imputadas por Media, 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.

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
		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 10. RMSE para la variable “petal width” imputada por Media, 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.

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 11. RMSE para la variables “petal width” y “petal length” imputadas por Media, 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

.

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 12. RMSE para las variables “petal width”, “petal length” y “sepal length” imputadas por Media, 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.

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
	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
	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
	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
	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 13. RMSE para la variable “petal width” imputada por Media, 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.

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 14. RMSE para las variables “petal width” y “petal length” imputadas por Media, k-NN, k-Means y Hot-Deckluego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 20% de los casos.

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 15. RMSE para las variables “petal width”, “petal length” y “sepal length” imputadas por Media, k-NN, k-Means y Hot-Deck luego de haber sido amputadas bajo los supuestos MCAR, MAR y MNAR en un 20% de los casos.

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

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

En las Tabla 16 a Tabla 24, se presentan los valores de la 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.

Tabla 16. RMSEN para la variable “petal width” amputada en un 10% de los casos.

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

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 patrón 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.

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

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

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.

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

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

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.

Tabla 19. RMSE Normalizado para la variable “petal width” amputada en un 15% de los casos.

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

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.

Tabla 20. RMSEN para la variable "petal length" amputada en un 15% de los casos.

Mecanismo	Patrón	Tipo	Método de Imputación			
			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

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 uno k-NN resultó ser el mejor método de imputación, mientras que en los restantes 6, el mejor fue k-Means.

Tabla 21. RMSE Normalizado para la variable “sepal length” amputada en un 15% de los casos.

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

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 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 mejor 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.

Tabla 22. RMSEN para la variable "petal width" amputada en el 20% de los casos.

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

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.

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

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

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.

El gráfico de barras de la Fig. 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 Fig. 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.

Tabla 24. RMSE Normalizado para la variable "sepal length" amputada en el 20% de los casos.

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

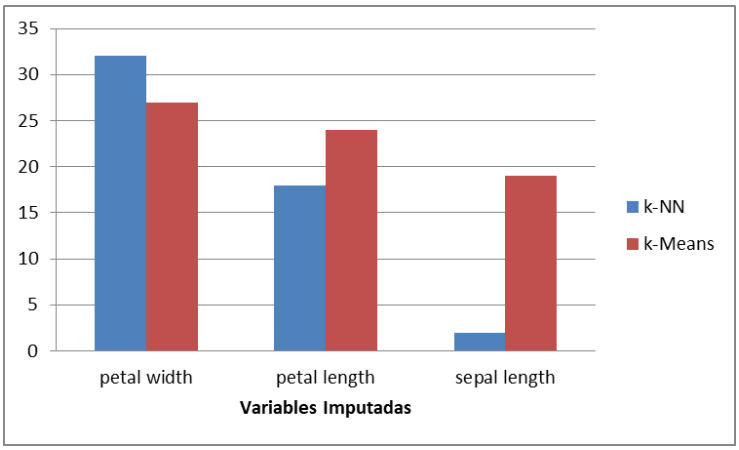

Fig. 3. Cantidad de veces que los métodos de imputación k-NN y k-Means resultaron los mejores para imputar las variables "petal width", "petal length" y "sepal length" en la totalidad de imputaciones realizadas.

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 Media, 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.

Tabla 25. Promedio de los RMSEN.

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

Finalmente, en la Tabla 26, se presenta el Promedio de los Errores producidos por los métodos de imputación Media, k-NN, k-Means y Hot-Deck, 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.

Tabla 26. Error promedio de cada método de imputación para la totalidad de conjuntos de datos imputados.

	Media	k-NN	k-Means	Hot-Deck
Error del Método de Imputación	0,80196084	0,32979837	**0,16661107**	0,44904304

2.6. Comentarios

En este Capítulo se presentó 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. A partir de un conjunto de datos completo, se generaron 63 conjuntos de datos con valores faltantes. Estos fueron imputados mediante los métodos de imputación por Media, 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.

Publicación relacionada

Primorac, C. R., La Red Martínez, D. L., Giovannini, M. E. (2020). "Metodología de Evaluación del Desempeño de MI Mediante una Métrica Tradicional Complementada con un Nuevo Indicador"; *European Scientific Journal (ESJ)*; Volume 16 – N° 18; pp. 61-92; ISSN N° 1857-7881; University Ss "Cyril and Methodius" Skopje, Macedonia.

Capítulo 3

Utilización de minería de datos para evaluar métodos de imputación

3.1. Introducción

Los valores faltantes introducen un elemento de ambigüedad en el análisis de datos. Pueden afectar las propiedades de estimadores estadísticos tales como la media, varianza o porcentajes, dando como resultado una pérdida de potencia y conclusiones falsas. (Schmitt et al., 2015).

En (Santos et al., 2019) se describe un enfoque clásico para la evaluación del desempeño de los métodos de imputación.

En otros trabajos se ha propuesto el uso de algoritmos de aprendizaje automático como métodos de imputación (Liu & Gopalakrishnan, 2017). Estas técnicas se basan en la construcción de un modelo predictivo para estimar datos faltantes en función de los valores disponibles en el conjunto de datos (García-Laencina et al., 2010).

En (Liu & Gopalakrishnan, 2017) se estudia la adecuación de los algoritmos de aprendizaje supervisados (clasificación) y los no supervisados (clústering) para la imputación. Algoritmos de aprendizaje automático tales como los árboles de decisión (DT: Decision Trees), k-vecinos más cercanos (k-NN: k-Nearest Neighbors), agrupamiento por k-Media (k-Means Clustering) y redes bayesianas (Bayesian Networks), han sido utilizados como métodos de imputación en diferentes dominios (García-Laencina et al., 2010), (Jerez et al., 2010), (Luengo et al., 2012), (M. M. Rahman & Davis, 2013), (Liu & Gopalakrishnan, 2017), (Nadzurah et al., 2018),.

En este Capítulo, se propone un criterio innovador para evaluar el desempeño de los métodos de imputación, en este caso Media, k-NN, k-Means y Hot-Deck, utilizando el valor de métricas indicadoras de calidad de un modelo de modelo de minería de datos obtenidas mediante procesos de minería de datos.

Se utilizó la técnica de regresión polinómica para crear modelos de minería de datos predictivos.

Se utilizó el criterio de mayor similitud entre los resultados de los procesos de minería de datos utilizando el conjunto de datos original (con valores completos) y los conjuntos de datos imputados luego de haber sido amputados, para lo cual se definieron nuevas métricas específicas a partir de los valores de las métricas obtenidas por los procesos de minería de datos.

Se utilizaron el conjunto de datos “Iris” original y 252 conjuntos de datos imputados luego de haber sido amputados descriptos en el Capítulo 2.

Las métricas indicadoras de calidad del modelo de minería de datos consideradas fueron calidad, precisión y clasificación (Ballard et al., 2010).

El Capítulo se organiza de la siguiente manera: en la sección revisión de conceptos de minería de datos se introducen los principales algoritmos y métricas de evaluación de modelos, en la sección materiales y métodos se describen los conjuntos de datos, el algoritmo de minería de datos y las métricas indicadora de calidad utilizadas, en la sección resultados y discusiones se comentan y comparan los mismos detalladamente, finalizándose con unos breves comentarios.

3.2. Revisión de conceptos de minería de datos

La minería de datos se refiere a los medios algorítmicos mediante los cuales se extraen y enumeran patrones a partir de los datos (La Red Martínez et al., 2016).

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. En un sentido amplio, el proceso de minería de datos se puede definir mediante los siguientes pasos básicos: adquisición de datos, preprocesamiento, generación del modelo, evaluación y explotación (La Red Martínez et al., 2016).

Además, el proceso de minería de datos tiene naturaleza cíclica, lo que significa que la generación de un modelo de minería de datos es un proceso dinámico e iterativo (Roiger, 2017).

3.2.1. Generación de modelos de minería de datos

En la práctica, los dos objetivos principales de la minería de datos, la predicción y la descripción, se pueden alcanzar utilizando una variedad de métodos particulares (Fayyad et al., 1996).

Los métodos predictivos, incluyen técnicas de aprendizaje supervisado como la clasificación y la regresión. Los métodos descriptivos, incluyen técnicas de aprendizaje no supervisado como el agrupamiento o clústering, las reglas de asociación o el descubrimiento de secuencias (Ballard et al., 2010).

Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos. El algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la

creación del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas (La Red Martínez et al., 2016).

Las técnicas de clasificación más comunes incluyen algoritmos de árbol y reglas de decisión, clasificadores Bayesianos, clasificadores basados en los vecinos más cercanos, regresión logística, máquinas de vectores de soporte (SVM: Support Vector Machines) y redes neuronales artificiales (ANN: Artificial Neural Networks) (Kononenko & Kukar, 2007), (Ballard et al., 2010).

Las técnicas de regresión más comunes incluyen algoritmos de regresión lineal (simple y múltiple), polinómica y regresión local ponderada, arboles de regresión, SVM para regresión y ANN (Kononenko & Kukar, 2007), (Chakrabarti et al., 2008), (Ballard et al., 2010).

En general, entre los principales algoritmos de agrupamiento se incluyen los métodos de particionado, jerárquicos, basados en distancia y basados en malla (Han et al., 2012).

La evaluación del desempeño de un modelo de minería de datos es probablemente el paso más crítico en todo el proceso de minería de datos (Roiger, 2017).

La calidad de los modelos de clasificación frecuentemente se evalúa mediante la precisión de la clasificación y la matriz de confusión (Kononenko & Kukar, 2007).

En los problemas de regresión las medidas se basan en la diferencia entre el valor verdadero y el predicho por el modelo (Kononenko & Kukar, 2007).

3.3. Materiales y métodos

En esta sección se describe el procedimiento seguido para evaluar el desempeño de cuatro método s de imputación: Media, k-NN, k-Means y Hot-Deck, utilizando los valores de las métricas de calidad, precisión y clasificación obtenidas mediante procesos de minería de datos utilizando modelos de regresión polinómica para clasificar el tipo de planta “Iris”.

3.3.1. Minería de datos

Se utilizó el software de IBM InfoSphere Warehouse (ISW) V.9.7 que incluye entre otras, herramientas (Intelligent Miner, Design Studio, etc.), para la creación, interpretación y evaluación de los modelos de minería de datos (Ballard et al., 2010).

Se utilizaron el conjunto de datos “Iris” original y 252 conjuntos de datos “Iris” imputados, obtenidos de imputar por los métodos de imputación por Media, k-NN, K-Means y Hot-

Deck los conjuntos de datos amputados en las 63 combinaciones de mecanismos, patrones y porcentajes de valores faltantes, según se detalla minuciosamente en el Capítulo 2.

En la etapa de minería de datos, se seleccionaron las técnicas a utilizar, creándose los flujos de minería correspondiente, en los cuales se parametrizaron los respectivos algoritmos.

Se consideró la técnica de regresión polinómica, cuyo objetivo es predecir el valor numérico de la variable dependiente sobre valores conocidos y así crear modelos que luego puedan ser utilizados para predecir valores nuevos o desconocidos.

Un modelo de regresión predice el valor numérico de la variable dependiente (objetivo) y en un registro, a partir los valores de las variables independientes (predictores) $x_1, x_2, \dots, x_n$ en el mismo registro. El modelo de regresión se crea y entrena utilizando conjunto de datos con registros para los cuales se conoce el valor de la variable objetivo (Ballard et al., 2010).

Se puede comprobar la calidad de predicción del modelo creado, utilizando un nuevo conjunto de datos con registros para los cuales se conoce el valor de la variable objetivo y comparando el valor predicho de la variable objetivo con su valor real (Ballard et al., 2010).

Matemáticamente, un modelo de regresión es una función $y = f(x_1, x_2, \dots, x_n)$ (Ballard et al., 2010).

El algoritmo de regresión polinómica asume una relación polinómica entre las variables independientes y la variable dependiente y extiende el modelo lineal al agregar predictores adicionales, obtenidos al elevar cada uno de los predictores originales a una potencia (Ballard et al., 2010).

Matemáticamente $y = a_0 + a_{11}x_1 + \cdots + a_{1m}x_1^m + \cdots + a_{n1}x_n + \cdots + a_{nm}x_n^m$ (Ballard et al., 2010).

El análisis de los resultados se basó en considerar como parámetro de minería la variable relacionada a la especie de planta “Iris” utilizando el conjunto de datos “Iris” original. Se seleccionó como variable objetivo el tipo de planta y como variables independientes, el ancho y largo del pétalo y del sépalo (Tabla 27).

Tabla 27. Matriz de correlaciones de “Iris”.

Variables	sepal length	sepal width	petal length	petal width	class
sepal length	1,0000	-0,1777	0,8774	0,8288	0,7885
sepal width	-0,1777	1,0000	-0,4434	-0,3549	-0,4320
petal length	0,8774	-0,4434	1,0000	0,9619	0,9462
petal width	0,8288	-0,3549	0,9619	1,0000	0,9526
class	0,7885	-0,4320	0,9462	0,9526	1,0000

En Design Studio los procesos de minería de datos se realizan creando y ejecutando flujos de minería de datos. El diseño de un flujo incluye, como mínimo, un operador tabla de entrada y un operador de minería de datos específico para la técnica de minería de datos que se utiliza. Adicionalmente, la mayor parte de los flujos incluyen uno o más operadores de salida, como el operador de visualización que presenta el valor de las métricas para evaluar el modelo obtenido (Ballard et al., 2010).

En la Fig. 4, se presenta el flujo de minería de datos utilizado para realizar los procesos de minería de datos. El operador <Tabla Fuente> define el conjunto de datos, que en este caso consiste en un registro para cada muestra del archivo de la planta "Iris", compuesto por los cuatro atributos predictores y el atributo objetivo descriptos en la Tabla 27. El operador <Predictor (Pronosticador)> ejecuta el algoritmo de minería de datos indicado (regresión polinómica) y envía el modelo de minería de datos obtenido al operador <Visualizador>, que presenta finalmente la información para evaluar el resultado del proceso de minería de datos.

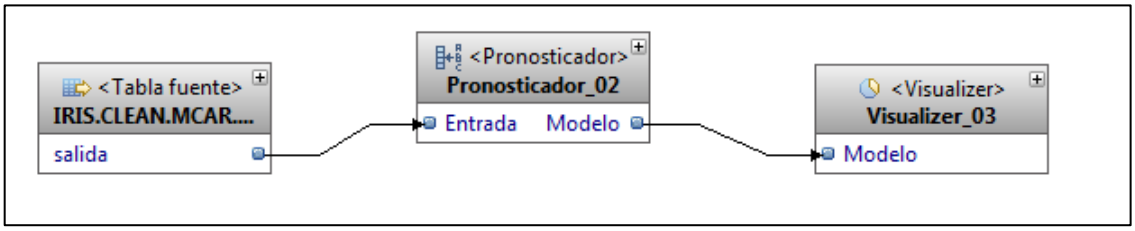

Fig. 4. Flujo de minería de datos en Design Studio (ISW V.9.7).

En la Fig. 5, se muestran las métricas de calidad del modelo, que varían entre 0 y 1, presentadas por el operador visualizador de Design Studio, consideradas para evaluar la calidad de los modelos de minería de datos obtenidos en cada uno de los procesos de minería de datos: i) calidad del modelo, ii) precisión y iii) clasificación (Ballard et al., 2010).

La calidad del modelo compara el rendimiento predictivo del modelo con el rendimiento predictivo de un modelo trivial que siempre devuelve como valor de predicción la media del atributo objetivo. Un valor de calidad de cero, indica que el desempeño predictivo del modelo no es mejor que predecir la media, mientras que un valor cercano a uno indica que el desempeño predictivo del modelo es muy superior a la predicción de la media (Ballard et al., 2010).

La precisión, representa la probabilidad de que una predicción sea correcta (Ballard et al., 2010).

Finalmente, la clasificación es una medida de la capacidad del modelo para ordenar los registros correctamente, calculada según el orden de los registros del conjunto de prueba cuando se ordenan por los valores de predicción con el orden de los mismos registros de datos cuando se ordenan por los valores reales de la variable objetivo (Ballard et al., 2010).

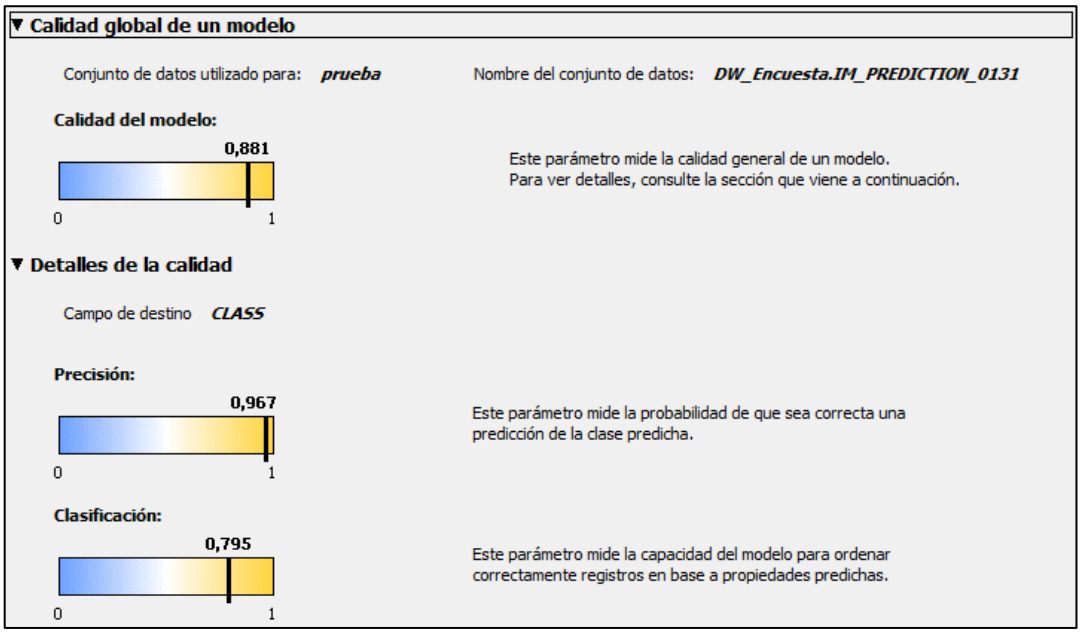

Fig. 5. Resultados mostrados por el Visualizador (ISW V.9.7).

Se ejecutaron los flujos de minería de datos para el conjunto de datos “Iris” original y los 252 conjuntos de datos imputados por los métodos de imputación por Media, k-NN, k-Means y Hot-Deck luego de haber sido amputados en las diferentes combinaciones de amputación descriptas en el Capítulo 2.

Para cada flujo de minería de datos, se obtuvieron los valores de tres métricas indicadoras de la calidad de los modelos de minería de datos obtenidos.

3.3.2. Evaluación del desempeño de los métodos de imputación utilizando métricas obtenidas por los procesos de minería de datos

Sean el conjunto de datos Y de la Tabla 28, con n casos y p variables donde y_{ij}, con $1 \leq i \leq n$ y $1 \leq j \leq p$, son valores observados; los conjuntos de datos imputados $Y^{a_r m_s}$, con $1 \leq r \leq l$ y $1 \leq s \leq t$, representados en la Tabla 29 y las métricas q_i, con $1 \leq i \leq k$, indicadoras de calidad del modelo de minería de datos, indicadas en la Tabla 30.

Tabla 28. Conjunto de datos original Y.

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 29. Conjuntos de datos $Y^{a_r m_s}$ con elementos $y_{ij}^{a_r m_s}$ imputados por el método m_s luego de haber sido amputados por el mecanismo a_r.

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

Sean $q_i(Y)$ los valores de las métricas q_i, indicadoras de calidad del modelo de minería de datos obtenido mediante el proceso de minería de datos utilizando el conjunto de datos Y, representados en la Tabla 30.

Tabla 30. Valores de las métricas $q_i(Y)$ indicadoras de calidad del modelo de minería de datos.

	q_1	q_2	...	q_i	...	q_k
Y	$q_1(Y)$	$q_2(Y)$	...	$q_i(Y)$	...	$q_k(Y)$

Sean $q_i(Y^{a_r m_s})$ los valores de las métricas q_i, indicadoras de calidad de los modelos de minería de datos obtenidos mediante los procesos de minería de datos utilizando los conjuntos de datos $Y^{a_r m_s}$, con $1 \leq r \leq l$ y $1 \leq s \leq t$, representados en la Tabla 31.

Tabla 31. Valores de $q_i(Y^{a_r m_s})$.

Y	m_1	...	m_1	...	m_s	...	m_s	...
a_1	$q_1(Y^{a_1 m_1})$	...	$q_k(Y^{a_1 m_1})$	...	$q_1(Y^{a_1 m_s})$	...	$q_k(Y^{a_1 m_s})$	...
a_2	$q_1(Y^{a_2 m_1})$	...	$q_k(Y^{a_2 m_1})$	...	$q_1(Y^{a_2 m_s})$	...	$q_k(Y^{a_2 m_s})$	...
...	...	...	...	...	...	...	...	...
a_r	$q_1(Y^{a_r m_1})$	...	$q_k(Y^{a_r m_1})$	...	$q_1(Y^{a_r m_s})$	...	$q_k(Y^{a_r m_s})$	...
...	...	...	...	...	...	...	...	...
a_l	$q_1(Y^{a_l m_1})$	...	$q_k(Y^{a_l m_1})$	...	$q_1(Y^{a_l m_s})$	...	$q_k(Y^{a_l m_s})$	...

Se definió la métrica Δq_i^{rs}, con $1 \leq i \leq k$; $1 \leq r \leq l$ y $1 \leq s \leq t$, ecuación (5). Es decir, la diferencia en valor absoluto, entre los valores de las métricas $q_i(Y)$ y $q_i(Y^{a_r m_s})$ representados en las Tablas 30 y Tabla 31 respectivamente.

De esta manera, respecto de la métrica Δq_i^{rs}, el mejor método de imputación m_s, con $1 \leq s \leq t$, para imputar el conjunto de datos Y amputado en la combinación a_r, con $1 \leq r \leq l$, es aquel que minimice el valor de la métrica Δq_i^{rs}, con $1 \leq i \leq k$.

En la Tabla 32, se resumen los valores del primer término de la ecuación (5).

$$\Delta q_i^{rs} = |q_i(Y) - q_i(Y^{a_r m_s})| \tag{5}$$

Así, ordenando ascendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (5), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar el conjunto de datos Y amputado en la combinación a_r, con $1 \leq r \leq l$, respecto de la métrica Δq_i^{rs}, con $1 \leq i \leq k$.

El desempeño de los métodos de imputación utilizados para imputar un conjunto de datos amputado se evaluó utilizando esta nueva métrica definida, que permitió obtener un orden de bondad de los métodos de imputación, considerando un criterio de evaluación.

Se definió el orden de bondad de los métodos de imputación respecto del criterio considerado, como una lista o relación ordenada de los métodos de imputación según su desempeño para imputar un conjunto de datos amputado, considerando un criterio de evaluación. En esta lista, el mejor método ocupa el primer lugar y el peor el último.

Tabla 32. Valores de Δq_i^{rs}.

Y	m_1	...	m_1	...	m_s	...	m_s	...
a_1	Δq_1^{11}	...	Δq_k^{11}	...	Δq_1^{1s}	...	Δq_k^{1s}	...
a_2	Δq_1^{21}	...	Δq_k^{21}	...	Δq_1^{2s}	...	Δq_k^{2s}	...
...	...	...	...	...	...	...	...	...
a_r	Δq_1^{r1}	...	Δq_k^{r1}	...	Δq_1^{rs}	...	Δq_k^{rs}	...
...	...	...	...	...	...	...	...	...
a_l	Δq_1^{l1}	...	Δq_k^{l1}	...	Δq_1^{ls}	...	Δq_k^{ls}	...

En este escenario, el mejor método de imputación según un criterio (y su correspondiente métrica) puede resultar el peor de acuerdo con los restantes criterios. Evaluar un método de imputación según una única métrica puede no ser suficiente, ya que es posible que interese el mejor método en términos de dos o más métricas simultáneamente.

Un operador de agregación permite agregar, fundir o sintetizar información, esto es, combinar entre sí una serie de datos, procedentes de fuentes diversas, para llegar a una cierta conclusión o tomar una determinada decisión (La Red Martinez & Acosta, 2014).

Con el objeto de encontrar el mejor método de imputación m_s para imputar un conjunto de datos amputado en la combinación a_r en términos de las Δq_i, con $1 \leq i \leq k$, métricas simultáneamente, se definió una nueva métrica, basada en un operador de agregación, $Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})$ o simplemente Q_{rs} para abreviar, con $1 \leq r \leq l$; $1 \leq s \leq t$.

En este caso se consideró el promedio aritmético de los valores de las métricas utilizadas relacionadas con calidad, precisión y clasificación, como se muestra en la ecuación (6).

$$Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs}) = \frac{1}{k}\sum_{i=1}^{k} \Delta q_i^{rs}\,; con\ \begin{matrix} 1 \leq s \leq t \\ 1 \leq r \leq l \end{matrix} \tag{6}$$

De esta forma, ordenando ascendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (6), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar el conjunto de datos Y amputado en la combinación a_r, con $1 \leq s \leq l$, respecto de las Δq_i, con $1 \leq i \leq k$, métricas simultáneamente.

Para evaluar el desempeño de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones, es decir teniendo en cuenta todos los escenarios de amputación (todos los conjuntos de datos considerados), se utilizaron dos criterios.

Criterio 1. Sea una nueva métrica $R_{si}(\Delta q_i^{rs}, \Delta q_i^{rs}, \dots, \Delta q_i^{rs})$ o simplemente R_{si} para abreviar, con $1 \leq r \leq l$; $1 \leq i \leq k$ y $1 \leq s \leq t$, dada por la ecuación (7). Esta métrica así definida, permite calcular el *promedio aritmético* de los valores de la métrica $\Delta q_i(Y^{a_r m_s})$, para el método de imputación m_s utilizado para imputar todos los conjuntos de datos amputados en las a_r combinaciones.

$$R_{si}(\Delta q_i^{rs}, \Delta q_i^{rs}, \dots, \Delta q_i^{rs}) = \frac{1}{l}\sum_{r=1}^{l} \Delta q_i(Y^{a_r m_s});\ con\ \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \tag{7}$$

Así, ordenando ascendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (7), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$,

métodos de imputación utilizados para imputar todos los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones, respecto de la métrica Δq_i, con $1 \leq i \leq k$.

De manera similar, se definió una nueva métrica $T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})]$ o simplemente T_s, como se muestra en la ecuación (8), que permite obtener el promedio aritmético de los valores de la métrica Q_{rs} para el método de imputación m_s, con $1 \leq s \leq t$, utilizado para imputar todos los conjuntos de datos amputados en las a_r, con $1 \leq r \leq l$ combinaciones.

$$T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})] = \frac{1}{l}\sum_{r=1}^{l} Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})\ ; con\ 1 \leq s \leq t \quad (8)$$

Ordenando ascendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (8), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar todos los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones, respecto de las Δq_i métricas simultáneamente, con $1 \leq i \leq k$.

Criterio 2. Sea el orden de bondad de los métodos de imputación m_s, con $1 \leq s \leq t$, utilizados para imputar el conjunto de datos amputado en la combinación a_r, con $1 \leq r \leq l$, respecto de las métricas Δq_i^{rs}, con $1 \leq i \leq k$, y respecto de la métrica Q_{rs}.

Se asignó un puntaje p_i^{rs}, al método de imputación m_s, utilizado para imputar el conjunto de datos Y amputado en la combinación a_r, que resulte *primero en el orden de bondad* respecto de los valores de la métrica Δq_i^{rs} obtenidos utilizando la ecuación (5). De manera similar, se asigna un puntaje P_{rs} al método de imputación m_s, utilizado para imputar el conjunto de datos amputado en la combinación a_r, que resulte primero en el orden de bondad respecto de los valores de la métrica Q_{rs}.

El puntaje se asignó de acuerdo con los siguientes criterios. Si un método de imputación m_s resulta el primero en el orden de bondad, se le asigna 1 (un) punto al método. Si dos métodos de imputación m_s y $m_{s'}$ empatan el primer lugar en el orden de bondad, se le asigna ½ (medio) punto a cada uno de los ellos. Si tres métodos de imputación m_s, $m_{s'}$ y $m_{s''}$ empatan el primer lugar en el orden en el orden de bondad, a cada uno de ellos se le asigna 1/3 (un tercio) del punto y, en general, si los t métodos de imputación empatan el primer lugar en el orden de bondad, a cada uno le corresponde $1/t$ puntos.

Se definió una nueva métrica w_{si} como se muestra en la ecuación (9), como el puntaje obtenido por el método de imputación m_s, considerando la métrica Δq_i.

$$w_{si} = \sum_{r=1}^{l} p_i^{rs} ; con \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \quad (9)$$

Ordenando descendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (9), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones respecto de la métrica w_{si}.

De manera similar, se definió una nueva métrica W_s, como el puntaje obtenido por el método de imputación m_s, considerando los valores de la métrica P_{rs}, como se muestra en la ecuación (10).

$$W_s = \sum_{r=1}^{l} P_{rs} ; con\ 1 \leq s \leq t \quad (10)$$

Ordenando descendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (10), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones respecto de la métrica W_s.

Finalmente, se definió una nueva métrica G_s dada por la ecuación (11), como el puntaje general obtenido por cada método de imputación m_s considerando todas las métricas.

$$G_s = \left(\sum_{i=1}^{k} w_{si} \right) + W_s ; con\ 1 \leq s \leq t \quad (11)$$

Ordenando descendentemente los métodos de imputación por los valores dados por el primer término de la ecuación (11), se obtiene el orden de bondad de los m_s, con $1 \leq s \leq t$, métodos de imputación utilizados para imputar todos los conjuntos de datos Y amputados en las a_r, con $1 \leq r \leq l$, combinaciones respecto de la métrica G_s.

3.4. Resultados y discusiones

En la Tabla 33, se presentan los valores de las métricas indicadoras de calidad del modelo de minería de datos obtenido mediante el proceso de minería de datos utilizando el

conjunto de datos “Iris” original. Estas son calidad (Cal), precisión (Prec) y clasificación (Clas).

Tabla 33. Valores de las métricas para el conjunto de datos original.

	Cal	*Prec*	*Clas*
Iris	0,884	0,972	0,796

En la Tabla 34, se presentan los valores de las métricas indicadoras de calidad del modelo de minería de datos obtenido mediante los procesos de minería de datos utilizando los conjuntos de datos “Iris” imputados por los métodos de imputación por Media, k-NN, k-Means y Hot-Deck, luego de haber sido amputados en cada una las 63 combinaciones de mecanismos, patrones y porcentajes de valores faltantes descriptos en el Capítulo 2. En total, para 63 conjuntos de datos amputados, se obtuvieron 252 conjuntos de datos imputados (63 x 4).

Cada fila de la Tabla 34, representa las características de los conjuntos de datos amputados y el valor de cada una de las métricas indicadoras de bondad del modelo de minería de datos obtenido mediante el proceso de proceso de minería de datos utilizando el conjunto de datos imputado por los métodos de imputación por Media, k-NN, k-Means y Hot-Deck, luego de haber sido amputado.

Así, por ejemplo, el valor de precisión del modelo de minería de datos obtenido con el conjunto de datos “Iris” imputado por el método de imputación k-NN luego de haber sido amputado según el supuesto MCAR, en patrón univariado en un 10% de los registros es 0,967.

En la Tabla 35, se presentan los valores de las métricas diferencias en valor absoluto entre los valores de las métricas indicadoras de calidad del modelo de minería de datos mencionados en las Tabla 33 y Tabla 34, obtenidas mediante la ecuación (5).

Así, por ejemplo, los valores de las diferencias en valor absoluto entre las métricas de calidad (ΔCal) para los conjuntos de datos “Iris” imputados por Media, k-NN, k-Means y Hot-Deck luego de haber sido amputado en el supuesto MCAR, en patrón univariado en un 10% de los registros, son 0,007; 0,001; 0,004 y 0,008 respectivamente.

Ordenando los valores precedentes ascendentemente resultan los métodos k-NN, k-Means, Media y Hot-Deck, ordenados según su orden de bondad para el método de imputación pertinente.

A continuación, se describen los resultados presentados en la Tabla 35, respecto de cada una de las métricas y la cantidad de veces que cada método de imputación resultó primero,

segundo, tercero y cuarto en el orden de bondad para imputar cada uno de los 63 conjuntos de datos amputados.

Respecto de la diferencia en valor absoluto entre las métricas calidad (ΔCal), el método de imputación Media resultó primero, segundo, tercero y cuarto en 14, 1, 17 y 31 de 63 veces respectivamente. Asimismo, de las 14 veces que resultó en primer lugar, compartió posición con el método k-NN y en una con los métodos k-NN y k-Means. En términos de la diferencia en valor absoluto entre las métricas precisión (ΔPrec), el método de imputación por Media, resultó primero, segundo, tercero y cuarto en 12, 5, 20 y 26 de 63 veces respectivamente. Finalmente, para las diferencias en valor absoluto entre las métricas clasificación (ΔClas), el método de imputación Media resultó primero, segundo, tercero y cuarto en 11, 19, 11 y 22 de 63 veces respectivamente. De las 12 veces que resultó primero en el orden de bondad, en una oportunidad fue acompañado por el método k-NN y en una por el método k-Means.

Respecto de la diferencia en valor absoluto entre las métricas calidad (ΔCal), el método de imputación k-NN resultó primero, segundo, tercero y cuarto en 27, 25, 8, y 3 de 63 veces respectivamente. Asimismo, de las 27 veces que resultó primero en el orden de bondad, en una compartió posición con el método Hot-Deck y en tres con el método k-Means. En términos de la diferencia en valor absoluto entre las métricas precisión (ΔPrec), el método de imputación k-NN, resultó primero, segundo, tercero y cuarto en 35, 20, 5 y 3 veces de las 63 respectivamente. De las 35 veces que resultó primero, una vez lo hizo conjuntamente con k-Means. Finalmente, para la diferencia en valor absoluto entre las métricas clasificación (ΔClas), el método de imputación k-NN resultó primero, segundo, tercero y cuarto en 22, 13, 24 y 4 veces de las 63 respectivamente. De las 22 veces que resultó primero, en cuatro oportunidades lo hizo conjuntamente con k-Means.

Respecto de la diferencia en valor absoluto entre las métricas calidad (ΔCal), el método de imputación k-Means resultó primero, segundo, tercero y cuarto en 24, 23, 14 y 2 de 63 veces respectivamente. Asimismo, de las 24 veces que resulto primero en el orden de bondad, en una oportunidad lo hizo conjuntamente con Media y k-NN, en 3 con k-Means y en una con Hot-Deck. En términos de la diferencia en valor absoluto entre las métricas precisión (ΔPrec), el método de imputación k-Means resulto primero, segundo, tercero y cuarto en 13, 31, 17 y 2 veces de las 63 respectivamente. De las 13 veces que resulto primero, una vez lo hizo conjuntamente con k-NN. Finalmente para la diferencia en valor absoluto entre las métricas clasificación (ΔClas), el método de imputación k-Means resulto

primero, segundo, tercero y cuarto en 18, 14, 19 y 12 veces de las 63 respectivamente. De las 18 veces que resulto primero, en una oportunidad lo hizo conjuntamente con Media, en dos con k-NN y en una con Hot-Deck.

Respecto de la diferencia en valor absoluto entre las métricas calidad (ΔCal), el método de imputación Hot-Deck resultó primero, segundo, tercero y cuarto en 6, 10, 20 y 27 de 63 veces respectivamente. Asimismo, de las 6 veces que resulto primero en el orden de bondad, en una oportunidad lo hizo conjuntamente con k-NN y en una con k-Means. En términos de la diferencia en valor absoluto entre las métricas precisión (ΔPrec), el método de imputación Hot-Deck resultó primero, segundo, tercero y cuarto en 13, 31, 17 y 2 veces de las 63 respectivamente. Finalmente, para la diferencia en valor absoluto entre las métricas clasificación (ΔClas), el método de imputación Hot-Deck resulto primero, segundo, tercero y cuarto en 18, 18, 8 y 19 veces de las 63 respectivamente. De las 18 veces que resulto primero, en una oportunidad lo hizo conjuntamente con k-NN y en una con k-Means.

Tabla 34. Valores de las métricas para los conjuntos de datos imputados.

Conjunto de datos amputado en la combinación de amputación				Método de imputación utilizado para imputar el conjunto de datos amputado											
				Media			k-NN			k-Means			Hot-Deck		
Mecanismo	Tipo	Patrón	MR	*Cal*	*Prec*	*Clas*	*Cal*	*Prec*	*Clas*	*Cal*	*Prec*	*Clas*	*Cal*	*Prec*	*Clas*
MCAR	-	univa	0,1	0,877	0,97	0,784	0,883	**0,967**	0,798	0,88	0,967	0,793	0,876	0,964	0,788
			0,15	0,877	0,971	0,783	0,889	0,972	0,806	0,888	0,979	0,796	0,868	0,949	0,786
			0,2	0,881	0,974	0,788	0,881	0,967	0,795	0,881	0,963	0,8	0,872	0,955	0,79
		multiva2	0,1	0,897	0,997	0,797	0,891	0,973	0,809	0,88	0,968	0,792	0,878	0,96	0,796
			0,15	0,818	0,872	0,763	0,896	0,985	0,806	0,898	0,99	0,806	0,88	0,965	0,796
			0,2	0,773	0,753	0,793	0,872	0,942	0,801	0,901	0,99	0,812	0,835	0,895	0,775
		multiva3	0,1	0,848	0,92	0,775	0,859	0,908	0,808	0,858	0,907	0,808	0,852	0,893	0,811
			0,15	0,753	0,718	0,788	0,86	0,916	0,803	0,855	0,902	0,808	0,879	0,978	0,781
			0,2	0,782	0,806	0,758	0,88	0,973	0,786	0,811	0,824	0,798	0,803	0,811	0,796
MAR	LEFT	univa	0,1	0,893	0,988	0,798	0,867	0,924	0,809	0,866	0,923	0,809	0,675	0,549	0,802
			0,15	0,892	0,993	0,79	0,874	0,943	0,805	0,799	0,792	0,806	0,79	0,782	0,798
			0,2	0,892	0,994	0,79	0,79	0,777	0,803	0,786	0,768	0,804	0,812	0,824	0,801
		multiva2	0,1	0,784	0,764	0,805	0,861	0,91	0,813	0,86	0,907	0,813	0,852	0,893	0,811
			0,15	0,811	0,828	0,793	0,862	0,912	0,813	0,861	0,909	0,813	0,618	0,464	0,772
			0,2	0,713	0,642	0,784	0,863	0,919	0,808	0,865	0,919	0,812	0,863	0,908	0,818
		multiva3	0,1	0,742	0,716	0,768	0,88	0,991	0,768	0,859	0,912	0,806	0,817	0,826	0,808
			0,15	0,787	0,798	0,777	0,876	0,988	0,764	0,89	0,973	0,806	0,842	0,876	0,808
			0,2	0,815	0,877	0,753	0,812	0,86	0,764	0,893	0,973	0,813	0,842	0,93	0,755
	MID	univa	0,1	0,826	0,866	0,786	0,896	0,989	0,803	0,861	0,912	0,81	0,794	0,814	0,775
			0,15	0,881	0,972	0,79	0,89	0,981	0,799	0,863	0,917	0,81	0,593	0,435	0,752
			0,2	0,879	0,955	0,804	0,89	0,981	0,799	0,895	0,986	0,805	0,835	0,906	0,764
		multiva2	0,1	0,795	0,784	0,806	0,905	1	0,81	0,86	0,909	0,812	0,853	0,908	0,798
			0,15	0,757	0,707	0,808	0,882	0,96	0,803	0,9	0,991	0,808	0,594	0,453	0,734
			0,2	0,753	0,697	0,808	0,884	0,965	0,803	0,883	0,956	0,81	0,799	0,797	0,802
		multiva3	0,1	0,812	0,873	0,752	0,873	0,979	0,768	0,893	0,976	0,81	0,722	0,649	0,796
			0,15	0,839	0,943	0,736	0,872	0,976	0,768	0,894	0,988	0,8	0,841	0,869	0,813
			0,2	0,802	0,873	0,731	0,856	0,965	0,746	0,894	0,983	0,805	0,731	0,709	0,753
	RIGHT	univa	0,1	0,793	0,794	0,791	0,885	0,983	0,787	0,853	0,889	0,818	0,863	0,936	0,79
			0,15	0,878	0,956	0,8	0,89	0,981	0,799	0,861	0,922	0,8	0,576	0,411	0,74
			0,2	0,884	0,965	0,803	0,89	0,981	0,799	0,859	0,904	0,815	0,744	0,727	0,76
		multiva2	0,1	0,787	0,765	0,81	0,9	0,992	0,808	0,896	0,982	0,81	0,893	0,998	0,788
			0,15	0,773	0,742	0,804	0,887	0,97	0,803	0,895	0,982	0,808	0,822	0,854	0,791
			0,2	0,744	0,695	0,793	0,887	0,978	0,797	0,882	0,955	0,81	0,644	0,553	0,734
		multiva3	0,1	0,855	0,976	0,734	0,818	0,891	0,745	0,898	0,998	0,798	0,858	0,912	0,803
			0,15	0,816	0,898	0,734	0,84	0,936	0,745	0,861	0,917	0,805	0,819	0,839	0,8
			0,2	0,785	0,865	0,705	0,858	0,976	0,741	0,894	0,999	0,789	0,858	0,928	0,787
MNAR	LEFT	univa	0,1	0,893	0,988	0,798	0,867	0,924	0,809	0,867	0,923	0,811	0,675	0,549	0,802
			0,15	0,889	0,988	0,79	0,89	0,981	0,799	0,87	0,934	0,806	0,704	0,62	0,789
			0,2	0,889	0,988	0,79	0,89	0,981	0,799	0,873	0,94	0,806	0,846	0,886	0,805
		multiva2	0,1	0,789	0,771	0,806	0,861	0,91	0,813	0,861	0,908	0,813	0,856	0,901	0,811
			0,15	0,823	0,852	0,793	0,862	0,91	0,813	0,86	0,902	0,818	0,767	0,757	0,776
			0,2	0,726	0,673	0,779	0,861	0,91	0,811	0,861	0,909	0,813	0,859	0,907	0,811
		multiva3	0,1	0,714	0,662	0,767	0,871	0,956	0,786	0,859	0,911	0,806	0,852	0,897	0,808
			0,15	0,834	0,9	0,769	0,872	0,977	0,766	0,857	0,902	0,812	0,849	0,886	0,812
			0,2	0,766	0,786	0,747	0,825	0,892	0,757	0,896	0,908	0,812	0,855	0,929	0,781
	MID	univa	0,1	0,753	0,721	0,784	0,89	0,981	0,799	0,863	0,92	0,806	0,878	0,985	0,771
			0,15	0,883	0,975	0,79	0,89	0,981	0,799	0,863	0,92	0,806	0,55	0,351	0,75
			0,2	0,888	0,987	0,788	0,89	0,981	0,799	0,872	0,938	0,806	0,796	0,845	0,747
		multiva2	0,1	0,814	0,825	0,803	0,856	0,899	0,813	0,852	0,887	0,816	0,732	0,667	0,797
			0,15	0,757	0,707	0,808	0,882	0,96	0,803	0,9	0,991	0,808	0,594	0,453	0,734
			0,2	0,753	0,697	0,808	0,884	0,965	0,803	0,883	0,956	0,81	0,799	0,797	0,802
		multiva3	0,1	0,852	0,952	0,752	0,875	0,982	0,768	0,894	0,978	0,81	0,585	0,402	0,768
			0,15	0,829	0,917	0,741	0,848	0,95	0,746	0,884	0,971	0,796	0,847	0,906	0,788
			0,2	0,802	0,872	0,731	0,846	0,946	0,746	0,895	0,982	0,808	0,748	0,734	0,763
	RIGHT	univa	0,1	0,774	0,755	0,792	0,874	0,967	0,782	0,894	0,988	0,8	0,69	0,607	0,774
			0,15	0,688	0,612	0,764	0,739	0,719	0,759	0,892	0,985	0,8	0,838	0,907	0,769
			0,2	0,893	0,991	0,795	0,89	0,981	0,799	0,854	0,887	0,82	0,865	0,969	0,76
		multiva2	0,1	0,787	0,765	0,81	0,9	0,992	0,808	0,899	0,988	0,81	0,893	0,998	0,788
			0,15	0,773	0,742	0,804	0,887	0,97	0,803	0,883	0,952	0,815	0,801	0,828	0,775
			0,2	0,744	0,695	0,763	0,887	0,978	0,797	0,88	0,95	0,81	0,644	0,533	0,734
		multiva3	0,1	0,511	0,294	0,728	0,567	0,368	0,766	0,897	0,996	0,798	0,871	0,942	0,8
			0,15	0,658	0,591	0,724	0,82	0,893	0,746	0,889	0,984	0,795	0,894	0,99	0,798
			0,2	0,671	0,63	0,713	0,856	0,96	0,751	0,857	0,91	0,803	0,809	0,817	0,801

Tabla 35. Valor de las métricas diferencias en valor absoluto.

Conjunto de datos amputado en la combinación de amputación				Método de imputación											
				Media			k-NN			k-Means			Hot-Deck		
Mecanismo	Tipo	Patrón	MR	ΔCal	$\Delta Prec$	$\Delta Clas$	ΔCal	$\Delta Prec$	$\Delta Clas$	ΔCal	$\Delta Prec$	$\Delta Clas$	ΔCal	$\Delta Prec$	$\Delta Clas$
MCAR	-	univa	0,1	**0,007**	0,002	0,012	**0,001**	0,005	0,002	**0,004**	0,005	0,003	**0,008**	0,008	0,008
			0,15	0,007	0,001	0,013	0,005	0,000	0,010	0,004	0,007	0,000	0,016	0,023	0,010
			0,2	0,003	0,002	0,008	0,003	0,005	0,001	0,003	0,009	0,004	0,012	0,017	0,006
		multiva2	0,1	0,013	0,025	0,001	0,007	0,001	0,013	0,004	0,004	0,004	0,006	0,012	0,000
			0,15	0,066	0,100	0,033	0,012	0,013	0,010	0,014	0,018	0,010	0,004	0,007	0,000
			0,2	0,111	0,219	0,003	0,012	0,030	0,005	0,017	0,018	0,016	0,049	0,077	0,021
		multiva3	0,1	0,036	0,052	0,021	0,025	0,064	0,012	0,026	0,065	0,012	0,032	0,079	0,015
			0,15	0,131	0,254	0,008	0,024	0,056	0,007	0,029	0,070	0,012	0,005	0,006	0,015
			0,2	0,014	0,010	0,038	0,084	0,177	0,010	0,073	0,148	0,002	0,081	0,161	0,000
MAR	LEFT	univa	0,1	0,009	0,016	0,002	0,017	0,048	0,013	0,018	0,049	0,013	0,209	0,423	0,006
			0,15	0,008	0,021	0,006	0,010	0,029	0,009	0,085	0,180	0,010	0,094	0,190	0,002
			0,2	0,008	0,022	0,006	0,094	0,195	0,007	0,098	0,204	0,008	0,072	0,148	0,005
		multiva2	0,1	0,100	0,208	0,009	0,023	0,062	0,017	0,024	0,065	0,017	0,032	0,079	0,015
			0,15	0,073	0,144	0,003	0,022	0,060	0,017	0,023	0,063	0,017	0,266	0,508	0,024
			0,2	0,171	0,330	0,012	0,021	0,053	0,012	0,019	0,053	0,016	0,021	0,064	0,022
		multiva3	0,1	0,142	0,256	0,028	0,004	0,019	0,028	0,025	0,060	0,010	0,067	0,146	0,012
			0,15	0,097	0,174	0,019	0,008	0,016	0,032	0,006	0,001	0,010	0,042	0,096	0,012
			0,2	0,069	0,095	0,043	0,072	0,112	0,032	0,009	0,001	0,017	0,042	0,042	0,041
	MID	univa	0,1	0,058	0,106	0,010	0,012	0,017	0,007	0,023	0,060	0,014	0,090	0,158	0,021
			0,15	0,003	0,000	0,006	0,006	0,009	0,003	0,021	0,055	0,014	0,291	0,537	0,044
			0,2	0,005	0,017	0,008	0,006	0,009	0,003	0,011	0,014	0,009	0,049	0,066	0,032
		multiva2	0,1	0,089	0,188	0,010	0,021	0,028	0,014	0,024	0,063	0,016	0,031	0,064	0,002
			0,15	0,127	0,265	0,012	0,002	0,012	0,007	0,016	0,019	0,012	0,290	0,519	0,062
			0,2	0,131	0,275	0,012	0,000	0,007	0,007	0,001	0,016	0,014	0,085	0,175	0,006
		multiva3	0,1	0,072	0,099	0,044	0,011	0,007	0,028	0,009	0,004	0,014	0,162	0,323	0,000
			0,15	0,045	0,029	0,060	0,012	0,004	0,028	0,010	0,016	0,004	0,043	0,103	0,017
			0,2	0,082	0,099	0,065	0,028	0,007	0,050	0,010	0,011	0,009	0,153	0,263	0,043
	RIGHT	univa	0,1	0,091	0,178	0,005	0,001	0,011	0,009	0,031	0,083	0,022	0,021	0,036	0,006
			0,15	0,006	0,016	0,004	0,006	0,009	0,003	0,023	0,050	0,004	0,308	0,561	0,056
			0,2	0,000	0,007	0,007	0,006	0,009	0,003	0,025	0,068	0,019	0,140	0,245	0,036
		multiva2	0,1	0,097	0,207	0,014	0,016	0,020	0,012	0,012	0,010	0,014	0,009	0,026	0,008
			0,15	0,111	0,230	0,008	0,003	0,002	0,007	0,011	0,010	0,012	0,062	0,118	0,005
			0,2	0,140	0,277	0,003	0,003	0,006	0,001	0,002	0,017	0,014	0,240	0,419	0,062
		multiva3	0,1	0,029	0,004	0,062	0,066	0,081	0,051	0,014	0,026	0,002	0,026	0,060	0,007
			0,15	0,068	0,074	0,062	0,044	0,036	0,051	0,023	0,055	0,009	0,065	0,133	0,004
			0,2	0,099	0,107	0,091	0,026	0,004	0,055	0,010	0,027	0,007	0,026	0,044	0,009
MNAR	LEFT	univa	0,1	0,009	0,016	0,002	0,017	0,048	0,013	0,017	0,049	0,015	0,209	0,423	0,006
			0,15	0,005	0,016	0,006	0,006	0,009	0,003	0,014	0,038	0,010	0,180	0,352	0,007
			0,2	0,005	0,016	0,006	0,006	0,009	0,003	0,011	0,032	0,010	0,038	0,086	0,009
		multiva2	0,1	0,095	0,201	0,010	0,023	0,062	0,017	0,023	0,064	0,017	0,028	0,071	0,015
			0,15	0,061	0,120	0,003	0,022	0,062	0,017	0,024	0,070	0,022	0,117	0,215	0,020
			0,2	0,158	0,299	0,017	0,023	0,062	0,015	0,023	0,063	0,017	0,025	0,065	0,015
		multiva3	0,1	0,170	0,310	0,029	0,013	0,016	0,010	0,025	0,061	0,010	0,032	0,075	0,012
			0,15	0,050	0,072	0,027	0,012	0,005	0,030	0,027	0,070	0,016	0,035	0,086	0,016
			0,2	0,118	0,186	0,049	0,059	0,080	0,039	0,012	0,064	0,016	0,029	0,043	0,015
	MID	univa	0,1	0,131	0,251	0,012	0,006	0,009	0,003	0,021	0,052	0,010	0,006	0,013	0,025
			0,15	0,001	0,003	0,006	0,006	0,009	0,003	0,021	0,052	0,010	0,334	0,621	0,046
			0,2	0,004	0,015	0,008	0,006	0,009	0,003	0,012	0,034	0,010	0,088	0,127	0,049
		multiva2	0,1	0,070	0,147	0,007	0,028	0,073	0,017	0,032	0,085	0,020	0,152	0,305	0,001
			0,15	0,127	0,265	0,012	0,002	0,012	0,007	0,016	0,019	0,012	0,290	0,519	0,062
			0,2	0,131	0,275	0,012	0,000	0,007	0,007	0,001	0,016	0,014	0,085	0,175	0,006
		multiva3	0,1	0,032	0,020	0,044	0,009	0,010	0,028	0,010	0,006	0,014	0,299	0,570	0,028
			0,15	0,055	0,055	0,055	0,036	0,022	0,050	0,000	0,001	0,000	0,037	0,066	0,008
			0,2	0,082	0,100	0,065	0,038	0,026	0,050	0,011	0,010	0,012	0,136	0,238	0,033
	RIGHT	univa	0,1	0,110	0,217	0,004	0,010	0,005	0,014	0,010	0,016	0,004	0,194	0,365	0,022
			0,15	0,196	0,360	0,032	0,145	0,253	0,037	0,008	0,013	0,004	0,046	0,065	0,027
			0,2	0,009	0,019	0,001	0,006	0,009	0,003	0,030	0,085	0,024	0,019	0,003	0,036
		multiva2	0,1	0,097	0,207	0,014	0,016	0,020	0,012	0,015	0,016	0,014	0,009	0,026	0,008
			0,15	0,111	0,230	0,008	0,003	0,002	0,007	0,001	0,020	0,019	0,083	0,144	0,021
			0,2	0,140	0,277	0,033	0,003	0,006	0,001	0,004	0,022	0,014	0,240	0,439	0,062
		multiva3	0,1	0,373	0,678	0,068	0,317	0,604	0,030	0,013	0,024	0,002	0,013	0,030	0,004
			0,15	0,226	0,381	0,072	0,064	0,079	0,050	0,005	0,012	0,001	0,010	0,018	0,002
			0,2	0,213	0,342	0,083	0,028	0,012	0,045	0,027	0,062	0,007	0,075	0,155	0,005

En la Fig. 6, se presenta el número de veces que los método de imputación por Media, k-NN, k-Means y Hot-Deck resultaron primeros, respecto de cada métrica y según cada uno de los tres mecanismos de valores faltantes supuestos. Claramente se observa que el método de imputación k-NN resulta primero en general, excepto respecto de las métricas ΔCal y $\Delta Clas$ bajo el supuesto MAR donde ocupa el primer lugar con k-Means y respecto de la métrica $\Delta Prec$ en el supuesto MCAR donde el primer lugar es para Media.

El número de veces que los métodos de imputación por Media, k-NN, k-Means y Hot-Deck resultaron primero, respecto de cada métrica considerando los tres patrones de valores faltantes se presenta en la Fig. 7. Los gráficos muestran una clara disputa del primer lugar entre los métodos k-NN y k-Means. Respecto de la métrica ΔCal, el método de imputación por Media resulta claramente en el primer lugar cuándo se trata de un patrón univariado. Sin embargo, en el caso de un patrón multivariado simple el primero resulta k-NN; algo similar ocurre con el patrón multivariado complejo donde k-Means resulta primero. Respecto de $\Delta Prec$, el primer lugar es para k-NN tanto para el patrón univariado como multivariado simple, sin embargo, comparte el primer lugar con k-Means en el caso de un patrón multivariado complejo. Finalmente, respecto de $\Delta Clas$, los resultados son dispares, k-NN resultó primero en el caso de un patrón univariado, Hot-Deck en el caso de uno multivariado simple y k-Means en el caso de patrón multivariado complejo.

Por último, el número de veces que los métodos de imputación por Media, k-NN, k-Means y Hot-Deck resultaron primero, respecto de cada métrica y teniendo en cuenta los distintos porcentajes de valores faltantes se muestran en la Fig. 8. K-NN resulta primero respecto de ΔCal para un porcentaje de valores faltantes del 10% mientras que para el 15% y 20% el primero resulta k-Means. Respecto de $\Delta Prec$, claramente se observa que en todos los casos el primero resulto ser k-NN. Finalmente, respecto de $\Delta Clas$, k-Means resultó primero el 10% mientras que k-NN lo hizo para el 15% y 20%.

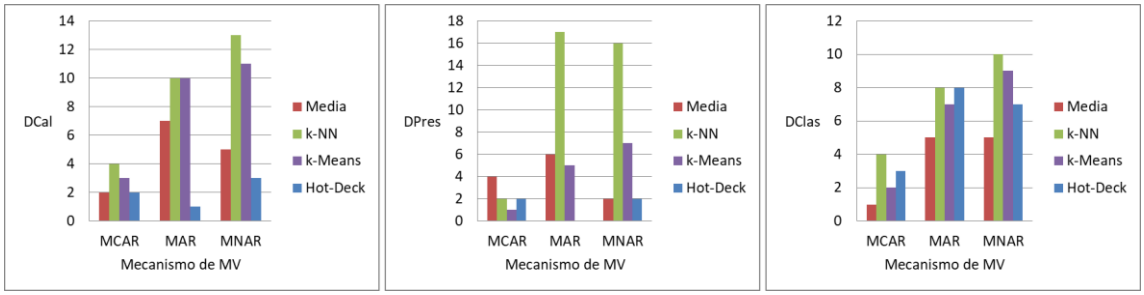

Fig. 6. Primer lugar según mecanismos de valores faltantes.

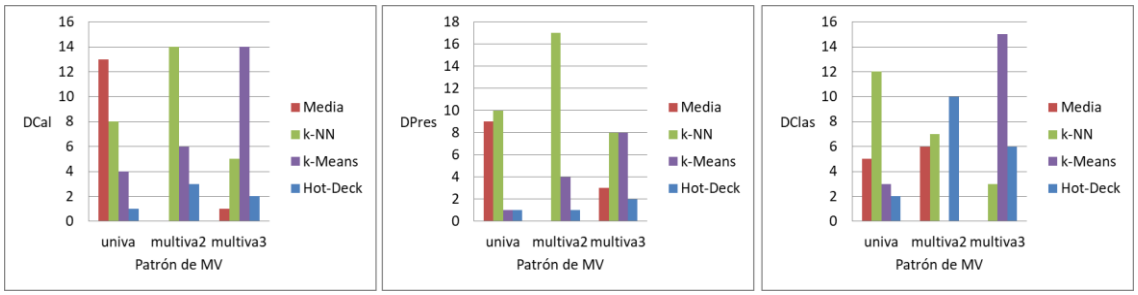

Fig. 7. Primer lugar según patrones de valores faltantes.

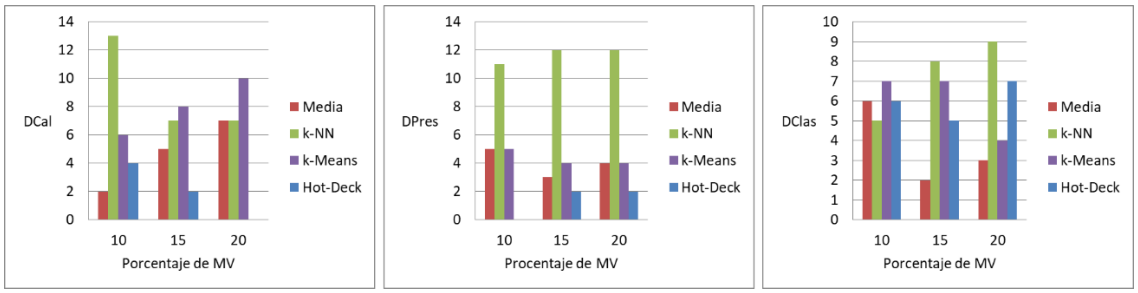

Fig. 8. Primer lugar según porcentaje de valores faltantes.

En la Tabla 36, se presentan los valores de las métricas obtenidas utilizando la ecuación (6), es decir, el promedio aritmético de los valores de métricas ΔCal, $\Delta Prec$ y $\Delta Clas$, indicadas en la Tabla 35, para cada método de imputación m_s utilizado para imputar el conjunto de datos amputado en la combinación a_r.

Ordenando ascendentemente los métodos de imputación por los valores de esta métrica, se obtiene el orden de bondad de los métodos de imputación Media, k-NN, k-Means y Hot-Deck utilizados para imputar el conjunto de datos "Iris" en cada una de las 63 combinaciones de amputación.

Por ejemplo, ordenando ascendentemente los métodos de imputación por los valores indicados en la primera fila, se obtiene el orden de bondad de los métodos de imputación Media, k-NN, k-Means y Hot-Deck utilizados para imputar el conjunto de datos "Iris"

luego de haber sido el conjunto de datos “Iris” original amputado según el mecanismo/en el supuesto MCAR, en patrón univariado en un 10% de los registros.

A continuación, se resumen los resultados presentados en la Tabla 36, respecto de esta métrica y la cantidad de veces que cada método de imputación resultó primero, segundo, tercero y cuarto en el orden de bondad para imputar cada uno de los 63 conjuntos de datos amputados.

El método de imputación por Media resultó primero, segundo, tercero y cuarto en 8, 7, 19 y 29 veces de 63 respectivamente. Asimismo, k-NN ocupó el primero, segundo, tercero y cuarto lugar en 34, 17, 9 y 3 de 63 veces. El método k-Means resultó primero, segundo, tercero y cuarto en 18, 14, 19 y 12 veces de 63 y, finalmente, Hot-Deck ocupó el primero, segundo, tercero y cuarto lugar 4, 12, 18 y 29 de 63 veces respectivamente.

Tabla 36. Valores de la métrica promedio aritmético de ΔCal, $\Delta Prec$ y $\Delta Clas$.

Combinación de amputación				Método de imputación			
				Media	k-NN	k-Means	Hot-Deck
Mecanismo	Tipo	Patrón	MR	Promedio (ΔQ)	Promedio (ΔQ)	Promedio (ΔQ)	Promedio (ΔQ)
MCAR	-	univa	0,1	0,007	0,003	0,004	0,008
			0,15	0,007	0,005	0,004	0,016
			0,2	0,004	0,003	0,005	0,012
		multiva2	0,1	0,013	0,007	0,004	0,006
			0,15	0,066	0,012	0,014	0,004
			0,2	0,111	0,016	0,017	0,049
		multiva3	0,1	0,036	0,034	0,034	0,042
			0,15	0,131	0,029	0,037	0,009
			0,2	0,021	0,090	0,074	0,081
MAR	LEFT	univa	0,1	0,009	0,026	0,027	0,213
			0,15	0,012	0,016	0,092	0,095
			0,2	0,012	0,099	0,103	0,075
		multiva2	0,1	0,106	0,034	0,035	0,042
			0,15	0,073	0,033	0,034	0,266
			0,2	0,171	0,029	0,029	0,036
		multiva3	0,1	0,142	0,017	0,032	0,075
			0,15	0,097	0,019	0,006	0,050
			0,2	0,069	0,072	0,009	0,042
	MID	univa	0,1	0,058	0,012	0,032	0,090
			0,15	0,003	0,006	0,030	0,291
			0,2	0,010	0,006	0,011	0,049
		multiva2	0,1	0,096	0,021	0,034	0,032
			0,15	0,135	0,007	0,016	0,290
			0,2	0,139	0,005	0,010	0,089
		multiva3	0,1	0,072	0,015	0,009	0,162
			0,15	0,045	0,015	0,010	0,054
			0,2	0,082	0,028	0,010	0,153
	RIGHT	univa	0,1	0,091	0,007	0,045	0,021
			0,15	0,009	0,006	0,026	0,308
			0,2	0,005	0,006	0,037	0,140
		multiva2	0,1	0,106	0,016	0,012	0,014
			0,15	0,116	0,004	0,011	0,062
			0,2	0,140	0,003	0,011	0,240
		multiva3	0,1	0,032	0,066	0,014	0,031
			0,15	0,068	0,044	0,029	0,067
			0,2	0,099	0,028	0,015	0,026
MNAR	LEFT	univa	0,1	0,009	0,026	0,027	0,213
			0,15	0,009	0,006	0,021	0,180
			0,2	0,009	0,006	0,018	0,044
		multiva2	0,1	0,102	0,034	0,035	0,038
			0,15	0,061	0,034	0,039	0,117
			0,2	0,158	0,033	0,034	0,035
		multiva3	0,1	0,170	0,013	0,032	0,040
			0,15	0,050	0,016	0,038	0,046
			0,2	0,118	0,059	0,031	0,029
	MID	univa	0,1	0,131	0,006	0,028	0,015
			0,15	0,003	0,006	0,028	0,334
			0,2	0,009	0,006	0,019	0,088
		multiva2	0,1	0,075	0,039	0,046	0,153
			0,15	0,135	0,007	0,016	0,290
			0,2	0,139	0,005	0,010	0,089
		multiva3	0,1	0,032	0,016	0,010	0,299
			0,15	0,055	0,036	0,000	0,037
			0,2	0,082	0,038	0,011	0,136
	RIGHT	univa	0,1	0,110	0,010	0,010	0,194
			0,15	0,196	0,145	0,008	0,046
			0,2	0,010	0,006	0,046	0,019
		multiva2	0,1	0,106	0,016	0,015	0,014
			0,15	0,116	0,004	0,013	0,083
			0,2	0,150	0,003	0,013	0,247
		multiva3	0,1	0,373	0,317	0,013	0,016
			0,15	0,226	0,064	0,006	0,010
			0,2	0,213	0,028	0,032	0,078

En la Fig. 9, se ilustra el número veces que los métodos Media, k-NN, k-Means y Hot-Deck resultaron primeros en orden de bondad respecto de la métrica operador de agregación promedio aritmético y considerando los mecanismos, patrones y porcentajes de valores faltantes. Claramente, se observa que el método k-NN resultó primero en todos los casos, excepto en el caso de un patrón de valores faltantes multivariado complejo, donde resultó primero el método k-Means.

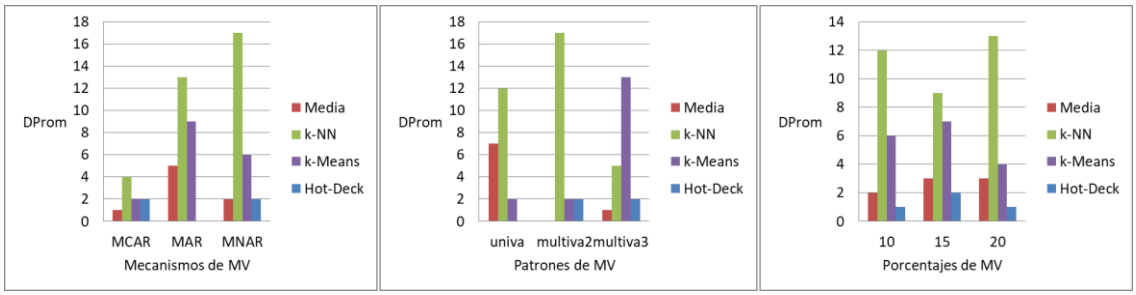

Fig. 9. Primer lugar respecto de la métrica promedio aritmético.

En la Tabla 37 se presentan los resultados obtenidos de aplicar las ecuaciones (7) y (8), definidas en el *Criterio 1*, a los valores obtenidos en las Tablas 35 y Tabla 36. Es decir, el promedio aritmético de los valores de las métricas de calidad, precisión, clasificación y la métrica agregada obtenido por cada método de imputación.

Ordenando ascendentemente los valores de la Tabla 37 se obtuvieron los métodos de imputación respecto de cada métrica, según su orden de bondad.

Respecto del promedio aritmético de los valores de la métrica ΔCal (Pro. ΔCal), resultaron los métodos k-Means, k-NN, Media y Hot-Deck según su orden de bondad. De manera similar, considerando el promedio aritmético de los valores de la métrica $\Delta Prec$ (Pro. $\Delta Prec$), se obtuvieron los métodos k-Means, k-NN, Media y Hot-Deck, según su orden de bondad. Sin embargo, teniendo en cuenta promedio aritmético de los valores de la métrica $\Delta Clas$ (Pro. $\Delta Clas$), resultaron los métodos k-Means, k-NN, Hot-Deck y Media. Finalmente, respecto del promedio aritmético de los valores de la métrica agregada (Pro. *Met. Agr.*), resultaron los métodos k-Means, k-NN, Media y Hot-Deck según su orden de bondad.

Tabla 37. Valores de las métricas promedio aritmético.

Método de Imputación	Métricas			
	Pro. ΔCal	Pro. $\Delta Prec$	Pro. $\Delta Clas$	Pro. Met. Agr.
Media	0,081	0,146	0,023	0,083
k-NN	0,026	0,044	0,017	0,029
k-Means	0,019	0,043	0,011	0,024
Hot-Deck	0,095	0,178	0,019	0,097

En la Tabla 38 se presentan los puntajes obtenidos por los métodos de imputación que resultaron primeros en el orden de bondad respecto de las métricas ΔCal, $\Delta Prec$ y $\Delta Clas$ considerando los valores obtenidos mediante la ecuación (5) y presentados en la Tabla 35, es decir considerando el *Criterio 2*.

Así, por ejemplo, considerando el orden de bondad de los métodos de imputación dado por el valor de las métricas ΔCal, $\Delta Prec$ y $\Delta Clas$ de la Tabla 35, respecto de la métrica ΔCal, se asignó un punto al método de imputación k-NN utilizado para imputar el conjunto de datos "Iris" amputado de acuerdo con el mecanismo MCAR, en patrón univariado, en un 10% de los registros; de manera similar, respecto de la métrica $\Delta Prec$, el método de imputación por Media obtuvo un punto al imputar el conjunto de datos "Iris" amputado de acuerdo al mecanismo MCAR, en patrón univariado, en un 10% de los registros.

De manera similar, respecto de la métrica ΔCal, se asignó 0,33 puntos a los métodos de imputación por Media, k-NN y k-Means utilizados para imputar el conjunto de datos "Iris" amputado de acuerdo con el mecanismo MCAR, en patrón univariado, en un 20% de los registros.

Igualmente, respecto de la métrica $\Delta Clas$, se asignó 0,5 puntos a los métodos de imputación k-Means y Hot-Deck utilizados para imputar el conjunto de datos "Iris" amputado de acuerdo con el mecanismo MCAR, en patrón multivariado complejo, en un 10% de los registros.

En la Tabla 39 se presentan los puntajes obtenidos, considerando el *Criterio 2*, por los métodos de imputación que resultaron primeros en el orden de bondad respecto de la métrica agregada considerando los valores obtenidos mediante la ecuación (6) (métrica ΔQ promedio de las métricas ΔCal, $\Delta Prec$ y $\Delta Clas$) y sistematizados en la Tabla 35. Así, por ejemplo, considerando el orden de bondad de los métodos de imputación dado por el valor de la métrica ΔQ de la Tabla 35, se asignó un punto al método de imputación k-NN utilizado para imputar el conjunto de datos "Iris" amputado de acuerdo con el mecanismo MCAR, en patrón univariado, en un 10% de los registros.

Finalmente, en la Tabla 40, se sintetiza el puntaje obtenido por cada método de imputación respecto de cada métrica, resultado de aplicar las ecuaciones (9) y (10) a los datos de las Tabla 38 y Tabla 39 el puntaje general obtenido por cada método de imputación, resultado de aplicar sobre la misma Tabla 40 la ecuación (11).

Tabla 38. Puntajes obtenidos respecto de cada métrica.

Características de los Conjuntos de Datos Amputados				Puntaje obtenido por cada Método de Imputación respecto de cada métrica												
				Media			k-NN			k-Means			Hot-Deck		
Mecanismo	Tipo	Patrón	MR	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$
MCAR	-	univa	0,1		1,00		**1,00**		1,00						
			0,15					1,00		1,00		1,00			
			0,2	**0,33**	1,00		**0,33**		1,00	**0,33**					
		multiva2	0,1					1,00		1,00					1,00
			0,15										1,00	1,00	1,00
			0,2			1,00	1,00				1,00				
		multiva3	0,1		1,00		1,00		**0,50**			**0,50**			
			0,15						1,00				1,00	1,00	
			0,2	1,00	1,00										1,00
MAR	LEFT	univa	0,1	1,00	1,00	1,00									
			0,15	1,00	1,00										1,00
			0,2	1,00	1,00										1,00
		multiva2	0,1			1,00	1,00	1,00							
			0,15			1,00	1,00	1,00							
			0,2			0,50		0,50	0,50	1,00	0,50				
		multiva3	0,1				1,00	1,00				1,00			
			0,15							1,00	1,00	1,00			
			0,2							1,00	1,00	1,00			
	MID	univa	0,1				1,00	1,00	1,00						
			0,15	1,00	1,00				1,00						
			0,2	1,00				1,00	1,00						
		multiva2	0,1				1,00	1,00							1,00
			0,15				1,00	1,00	1,00						
			0,2				1,00	1,00							1,00
		multiva3	0,1							1,00	1,00				1,00
			0,15					1,00		1,00		1,00			
			0,2					1,00		1,00		1,00			
	RIGHT	univa	0,1			1,00	1,00	1,00							
			0,15	0,50			0,50	1,00	1,00						
			0,2	1,00	1,00				1,00						
		multiva2	0,1								1,00		1,00		1,00
			0,15				1,00	1,00							1,00
			0,2					1,00	1,00	1,00					
		multiva3	0,1		1,00					1,00		1,00			
			0,15					1,00		1,00					1,00
			0,2					1,00		1,00		1,00			
MNAR	LEFT	univa	0,1	1,00	1,00	1,00									
			0,15	1,00				1,00	1,00						
			0,2	1,00				1,00	1,00						
		multiva2	0,1			1,00	0,50	1,00		0,50					
			0,15			1,00	1,00	1,00							
			0,2				0,50	1,00	0,50	0,50					0,50
		multiva3	0,1				1,00	1,00	0,50			0,50			
			0,15				1,00	1,00				0,50			0,50
			0,2							1,00				1,00	1,00
	MID	univa	0,1				0,50	1,00	1,00				0,50		
			0,15	1,00	1,00				1,00						
			0,2	1,00				1,00	1,00						
		multiva2	0,1				1,00	1,00							1,00
			0,15				1,00	1,00	1,00						
			0,2				1,00	1,00							1,00
		multiva3	0,1				1,00				1,00	1,00			
			0,15							1,00	1,00	1,00			
			0,2							1,00	1,00	1,00			
	RIGHT	univa	0,1			0,50	0,50	1,00		0,50		0,50			
			0,15							1,00	1,00	1,00			
			0,2			1,00	1,00							1,00	
		multiva2	0,1								1,00		1,00		1,00
			0,15					1,00	1,00	1,00					
			0,2				1,00	1,00	1,00						
		multiva3	0,1							0,50	1,00	1,00	0,50		
			0,15							1,00	1,00	1,00			
			0,2					1,00		1,00					1,00

Tabla 39. Puntaje obtenido respecto de la métrica promedio aritmético.

Características de los Conjuntos de Datos Amputados				Método de Imputación			
				Media	k-NN	k-Means	Hot-Deck
Mecanismo	Tipo	Patrón	MR	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$
MCAR	-	univa	0,1		**1,00**		
MCAR			0,15			1,00	
MCAR			0,2		1,00		
MCAR		multiva2	0,1			1,00	
MCAR			0,15				1,00
MCAR			0,2		1,00		
MCAR		multiva3	0,1		1,00		
MCAR			0,15				1,00
MCAR			0,2	1,00			
MAR	LEFT	univa	0,1	1,00			
MAR			0,15	1,00			
MAR			0,2	1,00			
MAR		multiva2	0,1		1,00		
MAR			0,15		1,00		
MAR			0,2		1,00		
MAR		multiva3	0,1		1,00		
MAR			0,15			1,00	
MAR			0,2			1,00	
MAR	MID	univa	0,1		1,00		
MAR			0,15	1,00			
MAR			0,2		1,00		
MAR		multiva2	0,1		1,00		
MAR			0,15		1,00		
MAR			0,2		1,00		
MAR		multiva3	0,1			1,00	
MAR			0,15			1,00	
MAR			0,2			1,00	
MAR	RIGHT	univa	0,1		1,00		
MAR			0,15		1,00		
MAR			0,2	1,00			
MAR		multiva2	0,1			1,00	
MAR			0,15		1,00		
MAR			0,2		1,00		
MAR		multiva3	0,1			1,00	
MAR			0,15			1,00	
MAR			0,2			1,00	
MNAR	LEFT	univa	0,1	1,00			
MNAR			0,15		1,00		
MNAR			0,2		1,00		
MNAR		multiva2	0,1		1,00		
MNAR			0,15		1,00		
MNAR			0,2		1,00		
MNAR		multiva3	0,1		1,00		
MNAR			0,15		1,00		
MNAR			0,2				1,00
MNAR	MID	univa	0,1		1,00		
MNAR			0,15	1,00			
MNAR			0,2		1,00		
MNAR		multiva2	0,1		1,00		
MNAR			0,15		1,00		
MNAR			0,2		1,00		
MNAR		multiva3	0,1			1,00	
MNAR			0,15			1,00	
MNAR			0,2			1,00	
MNAR	RIGHT	univa	0,1		1,00		
MNAR			0,15			1,00	
MNAR			0,2		1,00		
MNAR		multiva2	0,1				1,00
MNAR			0,15		1,00		
MNAR			0,2		1,00		
MNAR		multiva3	0,1			1,00	
MNAR			0,15			1,00	
MNAR			0,2		1,00		

Tabla 40. Puntaje obtenido por método de imputación respecto de cada métrica.

Método de Imputación	Puntaje obtenido respecto de cada métrica				
	$o_1(\Delta Cal)$	$o_2(\Delta Prec)$	$o_3(\Delta Clas)$	$O(\Delta Q)$	G
Media	12,83	12,00	10,00	8,00	42,83
k-NN	23,83	34,50	20,00	34,00	112,33
k-Means	21,33	12,50	16,00	17,00	66,83
Hot-Deck	5,00	4,00	17,00	4,00	30,00

En la Fig. 10 se grafican los valores de la Tabla 40.

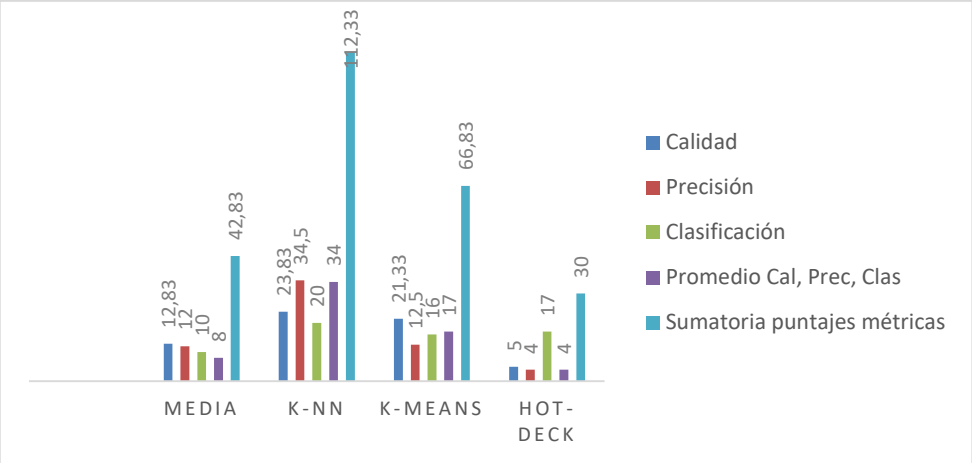

Fig. 10. Puntajes globales obtenidos por los métodos de imputación según las métricas utilizadas.

Ordenando descendentemente los valores de la Tabla 40 se obtuvieron los métodos de imputación respecto de cada métrica, según su orden de bondad para imputar el conjunto/o grupo de conjuntos de datos (archivos).

Respecto de los valores de la métrica ΔCal, resultaron mejores los métodos k-NN, k-Means, Media y Hot-Deck, según su orden de bondad. De manera similar, considerando los valores de la métrica ΔPrec, se obtuvieron los métodos k-NN, k-Means, Media y Hot-Deck, según su orden de bondad. Sin embargo, teniendo en cuenta los valores de la métrica $\Delta Clas$, resultaron los métodos k-NN, Hot-Deck, k-Means y Media. Finalmente, en cuanto a los valores de la métrica promedio aritmético ΔQ , resultaron los métodos k-NN, k-Means, Media y Hot-Deck según su orden de bondad.

Finalmente, considerando los valores de la métrica puntaje general obtenido G, resultaron los métodos k-NN, k-Means, Media y Hot-Deck según su orden de bondad.

Resumiendo, los mejores métodosd de imputación globalmente considerados resultaron ser el k-Means y el k-NN según el *Criterio 1*, el k-NN y el k-Means según el *Criterio 2* de esta

propuesta, y el k-Means y el k-NN según la metodología de cálculo basada en la raíz cuadrada del error cuadrático medio mostrados en Capítulo 2.

3.5. Comentarios

En este Capítulo se presentó una metodología de evaluación del desempeño de métodos de imputación mediante nuevas métricas derivadas de procesos de minería de datos, utilizando métricas de calidad de los modelos de minería de datos.

Se partió del conjunto de datos completos que fue amputado con diferentes mecanismos de amputación para generar 63 conjuntos de datos con valores faltantes; éstos fueron imputados mediante los métodos de imputación por Media, k-NN, k-Means y Hot-Deck.

El desempeño de los métodos de imputación fue evaluado utilizando nuevas métricas derivadas de métricas de calidad de los proceso de minería de datos realizados con el archivo original completo y con los archivos imputados luego haber sido amputados.

Esta evaluación no se basó en medir el error cometido al imputar (operación habitual), sino en considerar la similitud de los valores de las métricas de calidad de los procesos de minería de datos obtenidas con el archivo original y con los archivos imputados luego de haber sido amputados.

Los resultados muestran que, globalmente considerados y según las nuevas métricas propuestas, los métodos de imputación que mostraron mejor desempeño fueron el k-NN y el k-Means. Una ventaja adicional de la metodología propuesta es que permite disponer de modelos de minería de datos predictiva que podrán ser utilizados a posteriori.

Esta metodología contempla la posibilidad de usar modelos de minería de datos con diferentes algoritmos (en este capítulo se ha utilizado regresión), con lo cual se podrían evaluar los resultados arrojados considerando esos diferentes métodos de minería de datos, pudiendo agregar los resultados finales obtenidos con algún operador agregación, obteniéndose así el modelo de decisión a partir de varios modelos de minería de datos.

Publicaciones relacionadas

David L. la Red Martinez, Carlos R. Primorac. Use of Data Mining for Intelligent Evaluation of Imputation Methods, *International Journal of Interactive Multimedia and Artificial Intelligence.* Aceptado para revisión.

Capítulo 4

Conclusiones y futuras líneas de trabajo

4.1. Conclusiones

En el Capítulo 2 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 raíz cuadrada del error cuadrático medio.

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 permitirá 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 indicador 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.

En Capitulo 3 se ha presentado una innovadora metodología para evaluar el desempeño de métodos de imputación, basada en métricas derivadas de procesos de minería de datos, en vez de las generalmente utilizadas basadas en la raíz cuadrada del error cuadrático medio y sus derivaciones.

El entorno de trabajo implementado para realizar los experimentos de amputación y posterior imputación descripto en el Capítulo 2 resultó apropiado, ya que facilitó la gestión de los respectivos archivos originales, amputados e imputados, a los que se aplicaron los procesos de minería de datos realizados con el software ISW V.9.7.

La metodología propuesta y las métricas presentadas, 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 (ya que se parte de valores normalizados de métricas de minería de

datos), 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 valores faltantes.

Los resultados obtenidos mediante la metodología propuesta en sus distintas variantes de métricas (diferencias en valores absolutos y puntajes), son ligeramente diferentes pero coincidentes en que los mejores métodos de imputación globalmente considerados son el k-NN y el K-Means, lo cual también coincide con los resultados globales obtenidos por las métricas indicadas en el Capítulo 2.

La metodología propuesta, al contemplar varias métricas derivadas de los procesos de minería de datos, permite trabajar con sólo una de ellas o con todas simultáneamente, para determinar los mejores métodos de imputación ante un escenario determinado. Además, se puede aplicar a la evaluación de cualquier método de imputación, ya que se trabaja con los archivos imputados y no con los métodos en sí.

Esta metodología permite disponer de los modelos de minería de datos generados para evaluar los métodos de imputación, para realizar a posteriori minería de datos predictiva, constituyendo ello un valor agregado de esta propuesta.

Se ha planteado el uso de un modelo de decisión que contempla el uso de múltiples modelos de minería de datos.

4.2. 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 combinados basados en error cuadrático medio y 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.

Referencias

Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*, 1–5.

Arima, K., Okada, N., Tsuji, Y., & Kiguchi, K. (2014). Evaluations of a multiple SOMs method for estimating missing values. *2014 IEEE/SICE International Symposium on System Integration*, 796–801.

Ballard, C., Harris, N., Lawrence, A., Lowry, M., Perkins, A., & Voruganti, S. (2010). *InfoSphere Warehouse: A Robust Infrastructure for Business Intelligence*. An IBM Redbooks publication.

Ben-Gal, I. (2005). Outlier detection. In *Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook*. Springer.

Buuren, S. van, Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P., Doove, L., Jolani, S., Moreno-Betancur, M., White, I., Gaffert, P., Meinfelder, F., Gray, B., Arel-Bundock, V., Cai, M., Volker, T., Costantini, E., & Lissa, C. van. (2021). *Multivariate Imputation by Chained Equations*. https://cran.r-project.org/web/packages/mice/index.html

Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., Kamber, M., & Lightstone, S. S. (2008). *Data Mining. Know It All* (1st ed.). Elsevier.

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.

Enders, C. K. (2010). *Applied Missing Data Analysis* (2nd ed.). The Guilford Press.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *37*(5), 692–709.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Gajawada, S., & Toshniwal, D. (2012). Missing Value Imputation Method Based on Clustering and Nearest Neighbours. *International Journal of Future Computer and Communication*, 206–208.

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.

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.

Gill, J., Cranmer, S., Jackson, N., Murr, A., Armstrong, D., & Heuberger, S. (2017). *Multiple Hot-Deck Imputation*. https://cran.r-project.org/web/packages/hot.deck/hot.deck.pdf

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Iris Data Set. (2020). https://archive.ics.uci.edu/ml/datasets/iris

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, *33*(10), 913–933.

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. *Artificial Intelligence in Medicine*, *50*(2), 105–115.

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining: introduction to principles and algorithms*. Horwood Publishing Limited.

La Red Martinez, D. L., & Acosta, J. C. (2014). Revisión de Operadores de Agregación. *Campus Virtuales*, *3*, 24–44.

La Red Martínez, D. L., Karanik, M., Giovannini, M., Báez, M. E., & Torre, J. (2016). Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. *Campus Virtuales*, *5*, 70–83.

Liu, Y., & Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data*, *2*(1).

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, *32*(1), 77–108.

Madhu, G., & Rajinikanth, T. V. (2012). A novel index measure imputation algorithm for missing data values: A machine learning approach. *2012 IEEE International Conference on Computational Intelligence and Computing Research*, 1–7.

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.

missingpy 0.2.0. (n.d.). Retrieved October 3, 2020, from https://pypi.org/project/missingpy/

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*(6).

Pantanowitz, A., & Marwala, T. (2008). Evaluating the Impact of Missing Data Imputation through the use of the Random Forest Algorithm. *ArXiv*.

Pattanodom, M., Iam-On, N., & Boongoen, T. (2016). Clustering data with the presence of missing values by ensemble approach. *2016 2nd Asian Conference on Defence Technology, ACDT 2016*, 151–156.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556.

Pigott, T. D. (2010). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383.

Python. (n.d.). https://www.python.org/

Rahman, M. G., & Islam, M. Z. (2010). A decision tree-based missing value imputation technique for data pre-processing. *Conferences in Research and Practice in Information Technology Series*, *121*, 41–50.

Rahman, M. M., & Davis, D. N. (2013). Machine Learning-Based Missing Value Imputation Method for Clinical Datasets. *IAENG Transactions on Engineering Technologies*, 245–257.

Roiger, J. R. (2017). *Data Minig. A Tutorial-Based Primer* (2nd ed.). CRC Press.

rpy2 - R in Python. (n.d.). https://rpy2.github.io/doc/v2.9.x/html/index.html

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

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.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* (1st ed.). Chapman & Hall/CRC.

Schafer, Josepn L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, *06*(01), 1–6.

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.

scikit-learn. (n.d.). https://scikit-learn.org/stable/index.html

Seo, S. (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Set*. University of Pittsburgh.

sklearn.cluster.KMeans. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

sklearn.impute.SimpleImputer. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html#sklearn.impute.SimpleImputer

The R Project for Statistical Computing. (n.d.). https://www.r-project.org/about.html

Tobias, O. (2017). Performance of Imputation Algorithms on Artificially Produced Missing at Random Data. In *ProQuest Dissertations and Theses*. East Tennessee State University.

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, *90*, 84–99.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373–405.

Twala, B., Cartwright, M., & Shepperd, M. (2005). Comparison of various methods for handling incomplete data in software engineering databases. *2005 International Symposium on Empirical Software Engineering*.

Twala, B., Cartwright, M., & Shepperd, M. (2006). Ensemble of missing data techniques to improve software prediction accuracy. *28th International Conference on Software Engineering (ICSE 2006)*, 909–912.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data* (2nd ed.). Taylor & Francis Group, LLC.

van der Meijs, A. (2018). *Missing Data Imputation: Predicting Missing Values*. Tilburg University.

Vergouwe, Y., Royston, P., Moons, K. G. M., & Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology*, *63*(2), 205–214.