Use of Data Mining for Intelligent Evaluation of Imputation Methods

David L. la Red Martínez[1]*, Carlos R. Primorac[2]

[1] National Technological University, Resistencia Regional Faculty, Resistencia (Argentina)
[2] Computer Science Department, National University of the Northeast, Corrientes (Argentina)

* Corresponding author: lrmdavid@ca.frre.utn.edu.ar

Received 6 February 2022 | Accepted 1 March 2023 | Published 23 March 2023

Abstract

In real-world situations, researchers frequently face the difficulty of missing values (MV), i.e., values not observed in a data set. Data imputation techniques allow the estimation of MV using different algorithms, by means of which important data can be imputed for a particular instance. Most of the literature in this field deals with different imputation methods. However, few studies deal with a comparative evaluation of the different methods as to provide more appropriate guidelines for the selection of the method to be applied to impute data for specific situations. The objective of this work is to show a methodology for evaluating the performance of imputation methods by means of new metrics derived from data mining processes, using quality metrics of data mining models. We started from the complete dataset that was amputated with different amputation mechanisms to generate 63 datasets with MV; these were imputed using Median, k-NN, k-Means and Hot-Deck imputation methods. The performance of the imputation methods was evaluated using new metrics derived from quality metrics of the data mining processes, performed with the original full file and with the imputed files. This evaluation is not based on measuring the error when imputing (usual operation), but on considering the similarity of the values of the quality metrics of the data mining processes obtained with the original file and with the imputed files. The results show that –globally considered and according to the new proposed metric, the imputation methods that showed the best performance were k-NN and k-Means. An additional advantage of the proposed methodology is that it provides predictive data mining models that can be used *a posteriori.*

Keywords

Computer Science, Data Imputation, Data Mining, Interdisciplinary Applications, Performance Evaluation of Imputation Methods.

DOI: 10.9781/ijimai.2023.03.002

I. Introduction

MVS (Missing Values) introduce an element of ambiguity in data analysis. They can affect the properties of statistical estimators such as mean, variance or percentages, resulting in a loss of power and false conclusions. Data imputation is an alternative to deal with MV. Most of the published work in this field deals with the development of new imputation methods. However, few studies report a comprehensive evaluation of existing methods to provide guidelines to make the most appropriate methodological choice in practice [1].

The literature proposes two general approaches to dealing with MVs [2]. In the simplest case, they are omitted. A second option is to use imputation techniques and, from the complete data, estimate them using different algorithms, whereby an important feature can be imputed for a particular instance [3].

A classical approach to performance evaluation of imputation methods is described in [4].

Other works have proposed the use of machine learning (ML) algorithms as imputation methods [5]. These techniques are based on building a predictive model to estimate missing data based on the available values in the dataset [6]. In [5], the suitability of supervised (classification) and unsupervised (clustering) learning algorithms for imputation is studied. ML algorithms such as decision trees (DT), k-Nearest Neighbors (k-NN), k-Means Clustering and Bayesian Networks have been used as imputation methods in different domains [5], [6], [7], [8], [9], [10].

In this work, a continuation of [11], we do not propose the use of ML and data mining (DM) algorithms to impute. We rather propose an innovative criterion to evaluate the performance of imputation methods (IM), in this case Medians, k-NN, k-Means and Hot-Deck, using the value of quality indicator metrics of a data mining model (DMM) obtained through data mining processes (DMP). The polynomial regression technique was used to create predictive DMMs.

Please cite this article in press as:
D. L. la Red Martínez, C. R. Primorac. Use of Data Mining for Intelligent Evaluation of Imputation Methods, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 82-95, 2025, http://dx.doi.org/10.9781/ijimai.2023.03.002

We use the criterion of highest similarity between the results of the data mining processes using the original dataset (with complete data) and the imputed datasets after being amputated. New specific metrics were defined from the values of the metrics obtained by the DMPs.

We used the original "Iris" data set and 252 data sets imputed after amputation.

Quality, accuracy (precision) and classification metrics were considered as indicators of DMM quality [12].

The article is organized as follows: the Data Mining (DM) concept review section introduces the main algorithms and model evaluation metrics, the Materials and methods section describes the datasets, the data mining algorithm and the quality indicator metrics used, the Results and discussions section discusses and compares them in detail, and concludes with Conclusions, Future work, Acknowledgements and References.

II. Review of Data Mining Concepts (DM)

Historically, the notion of discovering hidden patterns in data has been given a variety of names including data mining (DM) and knowledge discovery (KDD: Knowledge Discovery in Databases). KDD refers to the general process of discovering useful knowledge from data. KDD is the application of specific algorithms to extract patterns from data. DM is a stage within the general KDD process that refers to the algorithmic means by which patterns are extracted and enumerated from data [13].

The generation of a DMM is part of a larger process that includes from the formulation of questions about the data and the creation of a model to answer them, to the implementation of the model in a working environment. In a broad sense, DMP can be defined by the following basic steps: data acquisition, preprocessing, model generation, evaluation, and exploitation [14].

In addition, DMP is cyclical in nature, meaning that the generation of a DMM is a dynamic and iterative process [15], [16].

A. Generation of DM Models

In practice, the two main objectives of DM, prediction and description, can be achieved by using a variety of methods [17].

Predictive methods include supervised learning techniques such as classification and regression. Descriptive methods include unsupervised learning techniques such as clustering, association rules or sequence discovery [12].

A DM algorithm is a set of calculations and heuristic rules that allows the creation of a DMM from data. To create a model, the algorithm first analyzes the data provided, looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating the DMM. These parameters are then applied across the entire dataset to extract actionable patterns and detailed statistics [14].

The most common classification techniques include tree algorithms and decision rules, Bayesian classifiers, nearest neighbor-based classifiers, logistic regression, support vector machines (SVM) and artificial neural networks (ANN) [12], [15], [18].

The most common regression techniques include linear regression algorithms (simple and multiple), polynomial and weighted local regression, regression trees, SVM for regression and ANN [19], [12], [18].

In general, the main clustering algorithms include partitioning, hierarchical, distance-based and mesh-based methods [15].

The performance evaluation of a DMM is probably the most critical step in the entire DMP [16].

The quality of classification models is often assessed by the classification accuracy and the confusion matrix [18].

In regression problems, measures are based on the difference between the true value and the value predicted by the model [18].

III. Materials and Methods

This section describes the procedure followed to evaluate the performance of four imputation methods (IM): Medians, k-NN, k-Means and Hot-Deck, using the values of quality, accuracy (precision) and classification metrics obtained through data mining processes, using polynomial regression models to classify the "Iris" plant type.

A. Data Mining

IBM InfoSphere Warehouse (ISW) V.9.7 software was used, which includes, among others, tools (Intelligent Miner, Design Studio, etc.), for the creation, interpretation, and evaluation of DMM [20].

The original "Iris" data set and 252 imputed "Iris" data sets, obtained from imputing by Mean, k-NN, k-Means and Hot-Deck IM the amputated data sets in the 63 combinations of mechanisms, patterns and MV percentages, as thoroughly detailed in [11], were used.

In the DM stage, the techniques to be used were selected and the corresponding mining flows were created, in which the respective algorithms were parameterized.

The polynomial regression technique was considered. Its objective is to predict the numerical value of the dependent variable on known values thus creating models that can then be used to predict new or unknown values.

For the analysis of results, the "Iris" data set was considered, corresponding to the plant species of the same name. The type of plant was selected as the objective variable t and the width and length of petals and sepals as independent variables, as presented in Table I.

TABLE I. "Iris" Correlation Matrix [11]

	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

In Design Studio, DMPs are performed by creating and executing DM flows. The design of a flow includes, at a minimum, an input table operator, and a DM operator specific to the DM technique being used. Additionally, most flows include one or more output operators, such as the visualization operator that presents the value of the metrics to evaluate the obtained model [20].

The DM flow used to perform the DMP has the following structure: The <Source Table> operator defines the data set, which in this case consists of one record for each sample of the "Iris" plant file, composed of the four predictor attributes and the target attribute described in Table I. The <Predictor> operator executes the indicated DM algorithm (polynomial regression) and sends the obtained DMM to the <Visualizer> operator, which finally presents the information to evaluate the DMP result.

The model quality metrics, which range from 0 to 1, are presented by the Design Studio viewer operator, and considered to evaluate the quality of the DMM obtained in each of the DMPs: i) model *quality*, ii) *accuracy (precision)* and iii) *classification* [12].

Model quality compares the model's predictive performance with the predictive performance of a trivial model that always returns the mean of the target attribute as the prediction value. A quality value of zero indicates that the model's predictive performance is no better than predicting the standard. In contrast, a value close to one indicates that the model's predictive performance is far superior to predicting the mean [12].

Accuracy (precision) represents the probability that a prediction be correct [12].

Finally, the *classification* is a measure of the model's ability to sort the records correctly. It is calculated according to the order of the test set records when sorted by the predicted values with the order of the same data records when sorted by the actual values of the target variable [12].

DM flows were run for the original "Iris" dataset and the 252 datasets imputed by the Means, k-NN, k-Means and Hot-Deck IM, after being amputated in the different amputation combinations described in [11].

For each DM flow, the values of three metrics indicating the quality of the DMM achieved were obtained.

B. Evaluation of the Performance of Imputation Methods (IM) Using Metrics Obtained From Data Mining Processes (DMP)

It is considered:

- The data set Y shown in Table II, with n cases and p variables, where y_{ij} are observed values, with $1 \le i \le n$ and $1 \le j \le p$.
- The imputed data sets $Y^{a_r m_s}$ depicted in Table III, with $1 \le r \le l$ and $1 \le s \le t$.
- The metrics q_i indicated in Table IV, which are quality indicators of DMM, with $1 \le i \le k$.

Table IV shows the values of the metrics q_i, with $1 \le i \le k$, which are the quality indicators of the DMM obtained by the DMP using the data set Y.

TABLE II. Original Data Set Y [11]

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}

TABLE III. Datasets Y^{ARMS} With Elements $y_{ij}^{a_r m_s}$ Imputed by the M_s Method After Having Been Amputated by the A_R Mechanism [11]

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

TABLE IV. Values of the Metrics $q_i(Y)$ Indicating the Quality of the DMM (Own Elaboration)

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

It is considered $q_i(Y^{a_r m_s})$ the values of the q_i metrics, indicators of quality of the DMM obtained through the DMP using the data sets $Y^{a_r m_s}$, with $1 \le r \le l$ and $1 \le s \le t$, represented in Table V.

The metric Δq_i^{rs}, with $1 \le i \le k$; $1 \le r \le l$ and $1 \le s \le t$, equation (1), was defined. That is, the difference in absolute value, between the values of the metrics $q_i(Y)$ and $q_i(Y^{a_r m_s})$ represented in Tables IV and V respectively.

Thus, with respect to the Δq_i^{rs} metric, the best imputation method m_s, with $1 \le s \le t$, for imputing the amputated Y data set in the combination a_r, with $1 \le r \le l$, is the one that minimizes the value of the Δq_i^{rs} metric, with $1 \le i \le k$.

TABLE V. Values of $q_i(Y^{a_r m_s})$ (Own Elaboration)

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

Table VI summarizes the values as expressed in equation (1).

$$\Delta q_i^{rs} = |q_i(Y) - q_i(Y^{a_r m_s})| \tag{1}$$

TABLE VI. Δq_i^{rs} Values (Own Elaboration)

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}	...

Thus, by ascendingly ordering the imputation methods by the values given by equation (1), we obtain the order of goodness of fit of the m_s, with $1 \le s \le t$, imputation methods used to impute the amputated Y data set in the combination a_r, with $1 \le r \le l$, with respect to the metric Δq_i^{rs}, with $1 \le i \le k$.

The performance of the imputation methods used to impute an amputated data set was evaluated using this newly defined metric, which made it possible to obtain an order of goodness of imputation methods, considering an evaluation criterion.

The order of goodness of imputation methods with respect to the criterion considered was defined as an ordered list or ratio of imputation methods according to their performance in imputing an amputated data set, considering an evaluation criterion. In this list, the best method is ranked first and the worst last.

In this scenario, the best imputation method according to one criterion (and its corresponding metric) may turn out to be the worst according to the remaining criteria. Evaluating an imputation method according to a single metric may not be sufficient, as the best method in terms of two or more metrics simultaneously may be of interest.

An aggregation operator makes it possible to aggregate, merge or synthesize information, that is, to combine a series of data from different sources to reach a certain conclusion or make a certain decision [21], [22].

In order to find the best imputation method m_s to impute an amputated *data set* in the combination a_r in terms of the Δq_i, with $1 \leq i \leq k$, metrics simultaneously, a new metric was defined, based on an aggregation operator, $Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \ldots, \Delta q_k^{rs})$ or simply Q_{rs} for short, with $1 \leq r \leq l$; $1 \leq s \leq t$. In this case, the *arithmetic average* of the values of the metrics used was considered, as shown in equation (2). It is considered convenient to use an aggregate value of the values of the metrics used, to avoid biases that could occur when using a single metric.

$$Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \ldots, \Delta q_k^{rs}) = \frac{1}{k}\sum_{i=1}^{k} \Delta q_i^{rs} \text{ ; with } \begin{matrix} 1 \leq s \leq t \\ 1 \leq r \leq l \end{matrix} \quad (2)$$

Thus, by ascendingly ordering the imputation methods by the values given by equation (2), we obtain the goodness-of-fit order of m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data set in the combination a_r, with $1 \leq s \leq l$, with respect to the Δq_i, with $1 \leq i \leq k$, metrics simultaneously.

To evaluate the performance of m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, i.e., *considering all amputation scenarios (all data sets considered)*, two criteria were used.

Criterion 1. It is considered a new metric $R_{si}(\Delta q_i^{rs}, \Delta q_i^{rs}, \ldots, \Delta q_i^{rs})$ or simply R_{si} for short, with $1 \leq r \leq l$; $1 \leq i \leq k$ and $1 \leq s \leq t$, given by equation (3). This metric thus defined, allows to compute the *arithmetic average* of the values of the metric Δq_i ($Y^{a_r m_s}$), for the imputation method m_s used to impute all amputed data sets in the a_r combinations.

R_{si} is an average indicator of the performance of the s imputation method for all files amputated with different mechanisms and then imputed with the s method, considering one of the metrics Δq_i.

$$R_{si}(\Delta q_i \ , \Delta q_i \ , \ldots, \Delta q_i \ ) = \frac{1}{l}\sum_{r=1}^{l} \Delta q_i(Y^{a_r m_s}); \text{ with} \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \quad (3)$$

Thus, by ascendingly ordering the imputation methods by the values given by equation (3), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, with respect to the metric Δq_i, with $1 \leq i \leq k$.

Similarly, a new metric was defined, $T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \ldots, \Delta q_k^{rs})]$ or simply T_s, as shown in equation (4), which allows to obtain the arithmetic average of the values of the metric Q_{rs} for the imputation method m_s, with $1 \leq s \leq t$, used to impute all amputed data sets in the a_r, with $1 \leq r \leq l$ combinations.

$$T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \ldots, \Delta q_k^{rs})] = \frac{1}{l}\sum_{r=1}^{l} Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \ldots, \Delta q_k^{rs}) \text{ ; with } 1 \leq s \leq t \quad (4)$$

Ascendingly ordering the imputation methods by the values given by the first term of equation (4), we obtain the order of goodness of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, with respect to the Δq_i metrics *simultaneously*, with $1 \leq i \leq k$.

T_s is an average indicator of the performance of the s imputation method for all files amputated with different mechanisms and then imputed with the s method, considering simultaneously all metrics Δq_i.

Criterion 2. It is considered the order of goodness of the imputation methods m_s, with $1 \leq s \leq t$, used to impute the amputed data set in the combination a_r, with $1 \leq r \leq l$, with respect to the metrics Δq_i^{rs}, with $1 \leq i \leq k$, and with respect to the metric Q_{rs}.

A score p_i^{rs} was assigned to the imputation method m_s, used to impute the amputated Y data set in the combination a_r, which comes *first in the order of goodness* of fit with respect to the values of the metric Δq_i^{rs} obtained using equation (1). Similarly, a score P_{rs} is assigned to the imputation method m_s, used to impute the amputated data set in the combination a_r, which comes first in the order of goodness of fit with respect to the values of the metric Q_{rs}.

The score was assigned according to the following criteria. If an imputation method m_s results first in the goodness-of-fit order, 1 (one) point is assigned to the method. If two imputation methods m_s and $m_{s'}$ tie for first place in the order of goodness of fit, ½ (half) point is assigned to each of them. If three imputation methods m_s, $m_{s'}$ and $m_{s''}$ tie for first place in the order of goodness of fit, each of them is assigned 1/3 (one third) of a point and, in general, if all t imputation methods tie for first place in the order of goodness of fit, each of them is assigned $1/t$ points.

Applying the above mentioned procedure, a new metric w_{si} was defined as shown in equation (5), as the score obtained by the imputation method m_s, considering the metric Δq_i. The value of w_{si} indicates the score obtained by the imputation method s for the metric Δq_i.

$$w_{si} = \sum_{r=1}^{l} p_i^{rs} \text{ ; with } \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \quad (5)$$

Sorting the imputation methods descendingly by the values given by equation (5), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the w_{si} metric.

Similarly, a new metric W_s was defined as the score obtained by the imputation method m_s, considering the values of the metric P_{rs}, average score of all metrics Δq_i.

$$W_s = \sum_{r=1}^{l} P_{rs} \text{ ; with } 1 \leq s \leq t \quad (6)$$

Sorting the imputation methods descendingly by the values given by equation (6), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the W_s metric.

Finally, a new metric G_s given by equation (7) was defined as the overall score obtained by each imputation method m_s considering all metrics, w_{si} and W_s.

$$G_s = \left(\sum_{i=1}^{k} w_{si}\right) + W_s; \text{ with } 1 \leq s \leq t \quad (7)$$

Sorting the imputation methods descendingly by the values given by equation (7), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the G_s metric.

This metric is considered the global indicator of this proposal, although each of the summands of equation (7) separately could also be considered as proxy indicators.

IV. Results and Discussions

Table VII presents the values of the quality indicator metrics of the DMM obtained through the DMP using the original "Iris" dataset. These are *quality (Cal)*, *precision (accuracy) (Prec)* and *classification (Clas)*.

TABLE VII. Values of the Metrics for the Original Data Set (Own Elaboration)

	Quality (Cal)	Precision (Prec)	Classification (Clas)
Iris	0.884	0.972	0.796

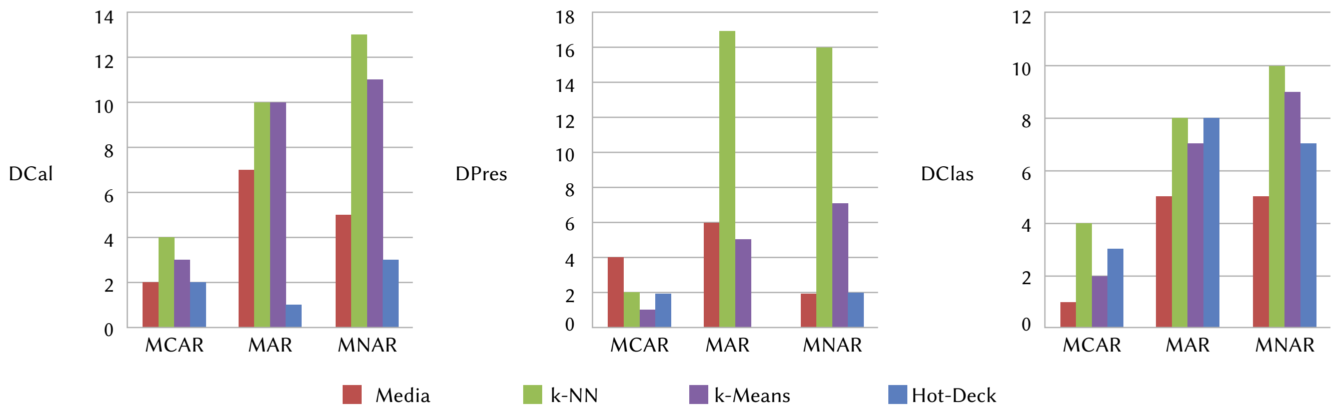

Fig. 1. First place according to MV mechanisms (Own elaboration).

Table VIII presents the values of the DMM quality indicator metrics obtained by DMP using the “Iris” datasets imputed by the Means, k-NN, k-Means and Hot-Deck imputation methods, after being amputated in each of the 63 combinations of mechanisms, patterns and MV percentages described in [11]. In total, for 63 amputated datasets, 252 imputed datasets were obtained (63 x 4). MR indicates the percentage of missing records.

Each row of Table VIII represents the characteristics of the amputated datasets and the value of each of the DMM goodness-of-fit indicator metrics obtained by DMP using the dataset imputed by Mean, k-NN, k-Means and Hot-Deck imputation methods after amputation.

Thus, for example, the *accuracy* value of the DMM obtained with the “Iris” data set imputed by the k-NN imputation method after having been amputated according to the MCAR assumption, in univariate pattern in 10% of the records is 0.967.

Table IX shows the values of the metrics *differences in absolute value* between the values of the DMM quality indicator metrics mentioned in Table VII and Table VIII, obtained using equation (1).

Thus, for example, the values of the *differences in absolute value* between the quality metrics (Δ*Cal*) for the “Iris” data sets imputed by Mean, k-NN, k-Means and Hot-Deck after amputation in the MCAR assumption, in univariate pattern on 10% of the records, are 0.007; 0.001; 0.004 and 0.008 respectively.

Sorting the preceding values in ascending order gives the k-NN, k-Means, Medians and Hot-Deck methods, ranked according to their order of goodness of fit for the relevant imputation method.

The results presented in Table IX for each of the metrics and the number of times each imputation method came first, second, third and fourth in the order of goodness of fit to impute each of the 63 amputated data sets are described below.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *Mean* imputation method came first, second, third and fourth 14, 1, 17 and 31 out of 63 times respectively. Also, of the 14 times it came first, it shared position with the k-NN method and in one with the k-NN and k-Means methods. In terms of the *absolute value difference* between the *precision* metrics (Δ*Prec*), the *Mean* imputation method came first, second, third and fourth 12, 5, 20 and 26 out of 63 times respectively. Finally, for the *absolute value differences* between the *classification* metrics (Δ*Clas*), the *Mean* imputation method came first, second, third and fourth 11, 19, 11 and 22 out of 63 times, respectively. Of the 12 times it came first in the order of goodness of fit, it was accompanied by the k-NN method once and the k-Means method once.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *k-NN* imputation method came first, second, third, and fourth in 27, 25, 8, and 3 of 63 times, respectively. Also, of the 27 times it came first in the goodness-of-fit order, in one it shared position with the Hot-Deck method and in three with the k-Means method. In terms of the *difference in absolute value* between the *precision* metrics (Δ*Prec*), the *k-NN* imputation method came first, second, third and fourth 35, 20, 5 and 3 times out of 63, respectively. Of the 35 times it came first, once it did so jointly with k-Means. Finally, for the *absolute value difference* between metric *classification* (Δ*Clas*), the *k-NN* imputation method came first, second, third and fourth 22, 13, 24 and 4 times out of 63, respectively. Of the 22 times it came first, four times it did so jointly with k-Means.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *k-Means* imputation method came first, second, third and fourth 24, 23, 14 and 2 out of 63 times, respectively. Likewise, of the 24 times it came first in the goodness-of-fit order, once it did so jointly with Mean and k-NN, 3 times with k-Means and once with Hot-Deck. In terms of the *difference in absolute value* between the *precision* metrics (Δ*Prec*), the *k-Means* imputation method came first, second, third and fourth 13, 31, 17 and 2 times out of 63, respectively. Of the 13 times it came first, once it did so jointly with k-NN. Finally, for the *absolute value difference* between the *classification* metrics Δ*Clas*), the *k-Means* imputation method came first, second, third and fourth 18, 14, 19 and 12 times out of 63, respectively. Of the 18 times it came first, once it did so jointly with Mean, twice with k-NN and once with Hot-Deck.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *Hot-Deck* imputation method came first, second, third and fourth in 6, 10, 20 and 27 out of 63 times, respectively. Also, of the 6 times it came first in the order of goodness of fit, it did so jointly with k-NN once and once with k-Means. In terms of the *absolute value difference* between the *precision* metrics (Δ*Prec*), the *Hot-Deck* imputation method came first, second, third and fourth 13, 31, 17 and 2 times out of 63, respectively. Finally, for the *absolute value difference* between *classification* metrics (Δ*Clas*), the *Hot-Deck* imputation method came first, second, third and fourth in 18, 18, 8 and 19 times out of 63 respectively. Of the 18 times it came first, once it did so jointly with k-NN and once with k-Means.

Fig. 1 presents the number of times that the Mean, k-NN, k-Means, and Hot-Deck imputation methods came first, with respect to each metric and under each of the three assumed *MV mechanisms.* It is clearly observed that the k-NN imputation method results first overall, except with respect to the Δ*Cal* and Δ*Clas* metrics under the MAR assumption where it ranks first with k-Means and with respect to the Δ*Prec* metric under the MCAR assumption where the first place is for Mean.

The number of times that the Mean, k-NN, k-Means and Hot-Deck imputation methods came first for each metric considering the three

TABLE VIII. Values of the Metrics for the Imputed Data Sets (Own Elaboration)

Amputation data set in the amputation combination				Imputation method used to impute the amputated dataset											
				Media			k-NN			k-Means			Hot-Deck		
Mechanism	**Type**	**Pattern**	**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

TABLE IX. Value of the Metrics Differences in Absolute Value (Own Elaboration).

Amputation data set in the amputation combination				Imputation method											
				Medias			k-NN			k-Means			Hot-Deck		
Mechanism	**Type**	**Pattern**	**MR**	**Δ*Cal***	**Δ*Prec***	**Δ*Clas***	**Δ*Cal***	**Δ*Prec***	**Δ*Clas***	**Δ*Cal***	**Δ*Prec***	**Δ*Clas***	**Δ*Cal***	**Δ*Prec***	**Δ*Clas***
			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
		univa	**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
			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
MCAR	**-**	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
	LEFT	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
MAR	**MID**	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
	RIGHT	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
	LEFT	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
MNAR	**MID**	**multiva2**	**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
			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
		multiva3	**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
			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
		univa	**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
			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
	RIGHT	**multiva2**	**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
			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
		multiva3	**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

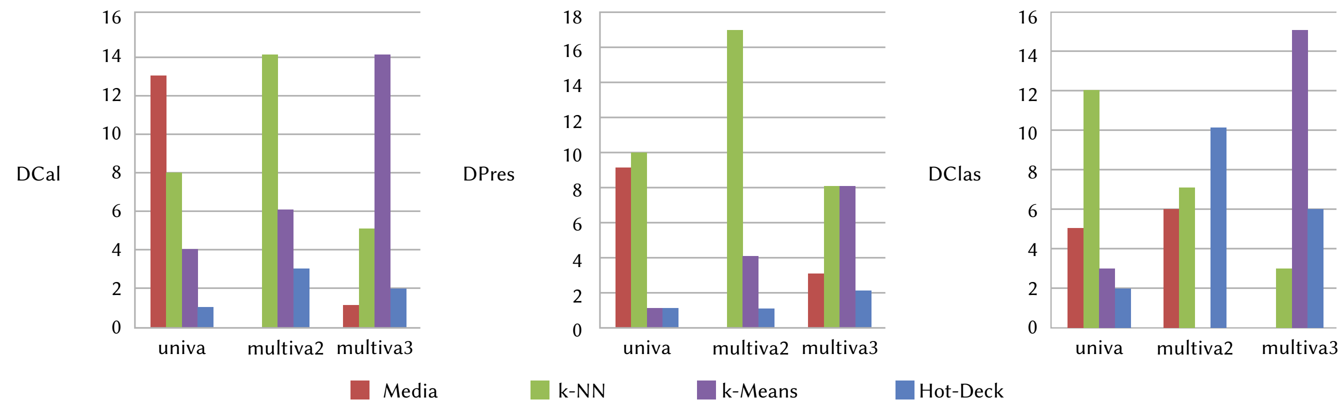

Fig. 2. First place according to MV patterns (Own elaboration).

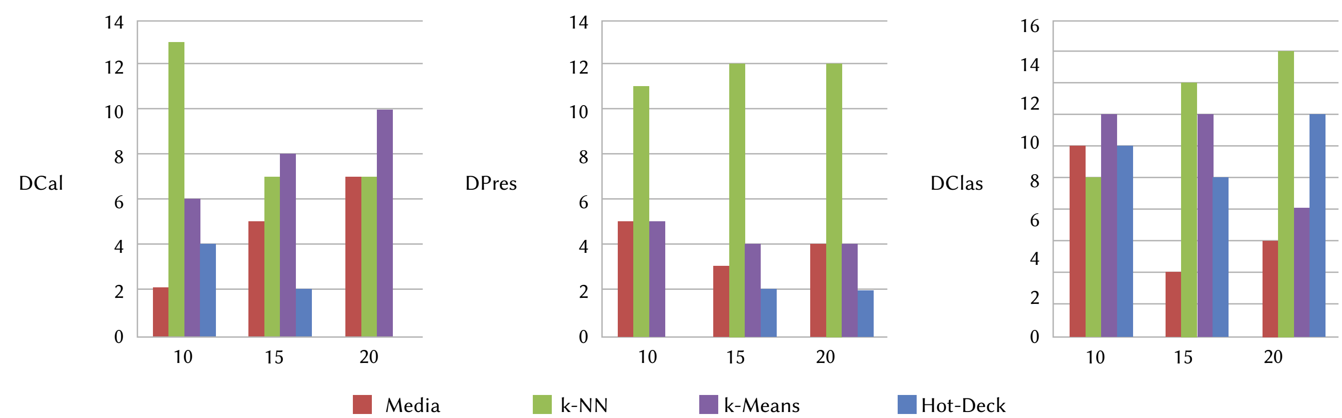

Fig. 3. First place according to percentage of MV (Own elaboration).

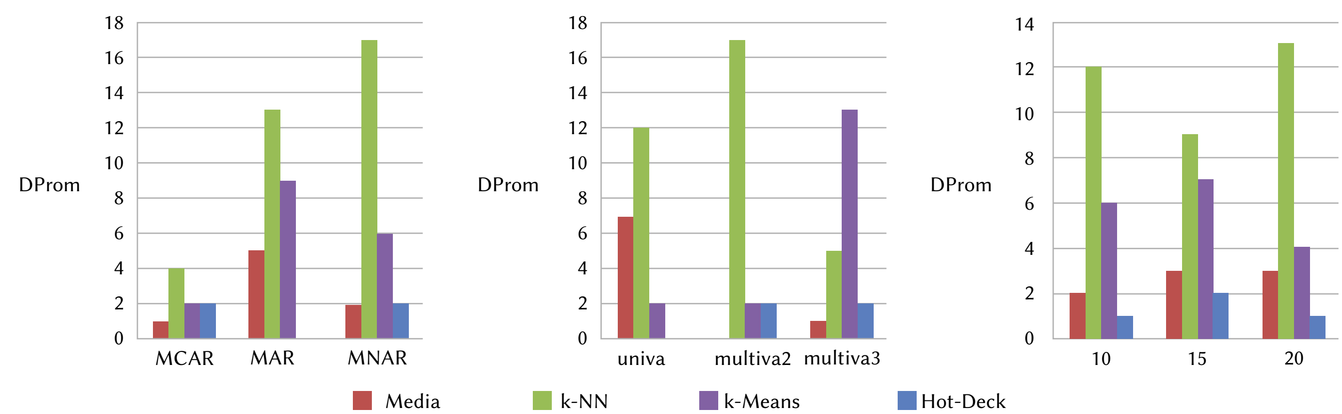

Fig. 4. First place with respect to the arithmetic average metric (Own elaboration).

MV patterns is presented in Fig. 2. The graphs show a clear dispute for first place between the k-NN and k-Means methods. Regarding the Δ*Cal* metric, the Mean imputation method clearly results in the first place when dealing with a univariate pattern. However, in the case of a simple multivariate pattern k-NN comes first; something similar happens with the complex multivariate pattern where k-Means comes first. Concerning Δ*Prec*, the first place is for k-NN for both the univariate and simple multivariate pattern, however, it shares the first place with k-Means in the case of a complex multivariate pattern. Finally, regarding *Clas*, the results are mixed, k-NN came first in the case of a univariate pattern, Hot-Deck in the case of a simple multivariate one and k-Means in the case of complex multivariate pattern.

Finally, the number of times that the Mean, k-NN, k-Means and Hot-Deck imputation methods came first, with respect to each metric and considering the different *MV percentages* are shown in Fig. 3. k-NN comes first with respect to Δ*Cal* for an MV percentage of 10% while for 15% and 20% k-Means comes first. With respect to Δ*Prec*, it is clearly observed that in all cases k-NN comes out first. Finally, with respect to Δ*Clas*, k-Means came first for 10% while k-NN came first for 15% and 20%.

In Table X, the values of the metrics obtained using equation (2), i.e., the *arithmetic average* of the metric values Δ*Cal*, Δ*Prec* and Δ*Clas*, indicated in Table IX, for each imputation method m_s used to impute the amputed data set in the combination a_r, are presented.

By sorting the imputation methods in ascending order by the values of this metric, we obtain the order of goodness of fit of the Medias, k-NN, k-Means and Hot-Deck imputation methods used to impute the "Iris" data set in each of the 63 amputation combinations.

TABLE X. Arithmetic Average Metric Values of ΔCal, $\Delta Prec$ and $\Delta Clas$ (Own Elaboration)

Amputation combination				Imputation method			
				Medias	k-NN	k-Means	Hot-Deck
Mechanism	**Type**	**Pattern**	**MR**	**Average (ΔQ)**	**Average (ΔQ)**	**Average (ΔQ)**	**Average (ΔQ)**
MCAR	**-**		**0.1**	0.007	0.003	0.004	0.008
MCAR	**-**	**univa**	**0.15**	0.007	0.005	0.004	0.016
MCAR	**-**		**0.2**	0.004	0.003	0.005	0.012
MCAR	**-**		**0.1**	0.013	0.007	0.004	0.006
MCAR	**-**	**multiva2**	**0.15**	0.066	0.012	0.014	0.004
MCAR	**-**		**0.2**	0.111	0.016	0.017	0.049
MCAR	**-**		**0.1**	0.036	0.034	0.034	0.042
MCAR	**-**	**multiva3**	**0.15**	0.131	0.029	0.037	0.009
MCAR	**-**		**0.2**	0.021	0.090	0.074	0.081
MAR	**LEFT**		**0.1**	0.009	0.026	0.027	0.213
MAR	**LEFT**	**univa**	**0.15**	0.012	0.016	0.092	0.095
MAR	**LEFT**		**0.2**	0.012	0.099	0.103	0.075
MAR	**LEFT**		**0.1**	0.106	0.034	0.035	0.042
MAR	**LEFT**	**multiva2**	**0.15**	0.073	0.033	0.034	0.266
MAR	**LEFT**		**0.2**	0.171	0.029	0.029	0.036
MAR	**LEFT**		**0.1**	0.142	0.017	0.032	0.075
MAR	**LEFT**	**multiva3**	**0.15**	0.097	0.019	0.006	0.050
MAR	**LEFT**		**0.2**	0.069	0.072	0.009	0.042
MAR	**MID**		**0.1**	0.058	0.012	0.032	0.090
MAR	**MID**	**univa**	**0.15**	0.003	0.006	0.030	0.291
MAR	**MID**		**0.2**	0.010	0.006	0.011	0.049
MAR	**MID**		**0.1**	0.096	0.021	0.034	0.032
MAR	**MID**	**multiva2**	**0.15**	0.135	0.007	0.016	0.290
MAR	**MID**		**0.2**	0.139	0.005	0.010	0.089
MAR	**MID**		**0.1**	0.072	0.015	0.009	0.162
MAR	**MID**	**multiva3**	**0.15**	0.045	0.015	0.010	0.054
MAR	**MID**		**0.2**	0.082	0.028	0.010	0.153
MAR	**RIGHT**		**0.1**	0.091	0.007	0.045	0.021
MAR	**RIGHT**	**univa**	**0.15**	0.009	0.006	0.026	0.308
MAR	**RIGHT**		**0.2**	0.005	0.006	0.037	0.140
MAR	**RIGHT**		**0.1**	0.106	0.016	0.012	0.014
MAR	**RIGHT**	**multiva2**	**0.15**	0.116	0.004	0.011	0.062
MAR	**RIGHT**		**0.2**	0.140	0.003	0.011	0.240
MAR	**RIGHT**		**0.1**	0.032	0.066	0.014	0.031
MAR	**RIGHT**	**multiva3**	**0.15**	0.068	0.044	0.029	0.067
MAR	**RIGHT**		**0.2**	0.099	0.028	0.015	0.026
MNAR	**LEFT**		**0.1**	0.009	0.026	0.027	0.213
MNAR	**LEFT**	**univa**	**0.15**	0.009	0.006	0.021	0.180
MNAR	**LEFT**		**0.2**	0.009	0.006	0.018	0.044
MNAR	**LEFT**		**0.1**	0.102	0.034	0.035	0.038
MNAR	**LEFT**	**multiva2**	**0.15**	0.061	0.034	0.039	0.117
MNAR	**LEFT**		**0.2**	0.158	0.033	0.034	0.035
MNAR	**LEFT**		**0.1**	0.170	0.013	0.032	0.040
MNAR	**LEFT**	**multiva3**	**0.15**	0.050	0.016	0.038	0.046
MNAR	**LEFT**		**0.2**	0.118	0.059	0.031	0.029
MNAR	**MID**		**0.1**	0.131	0.006	0.028	0.015
MNAR	**MID**	**univa**	**0.15**	0.003	0.006	0.028	0.334
MNAR	**MID**		**0.2**	0.009	0.006	0.019	0.088
MNAR	**MID**		**0.1**	0.075	0.039	0.046	0.153
MNAR	**MID**	**multiva2**	**0.15**	0.135	0.007	0.016	0.290
MNAR	**MID**		**0.2**	0.139	0.005	0.010	0.089
MNAR	**MID**		**0.1**	0.032	0.016	0.010	0.299
MNAR	**MID**	**multiva3**	**0.15**	0.055	0.036	0.000	0.037
MNAR	**MID**		**0.2**	0.082	0.038	0.011	0.136
MNAR	**RIGHT**		**0.1**	0.110	0.010	0.010	0.194
MNAR	**RIGHT**	**univa**	**0.15**	0.196	0.145	0.008	0.046
MNAR	**RIGHT**		**0.2**	0.010	0.006	0.046	0.019
MNAR	**RIGHT**		**0.1**	0.106	0.016	0.015	0.014
MNAR	**RIGHT**	**multiva2**	**0.15**	0.116	0.004	0.013	0.083
MNAR	**RIGHT**		**0.2**	0.150	0.003	0.013	0.247
MNAR	**RIGHT**		**0.1**	0.373	0.317	0.013	0.016
MNAR	**RIGHT**	**multiva3**	**0.15**	0.226	0.064	0.006	0.010
MNAR	**RIGHT**		**0.2**	0.213	0.028	0.032	0.078

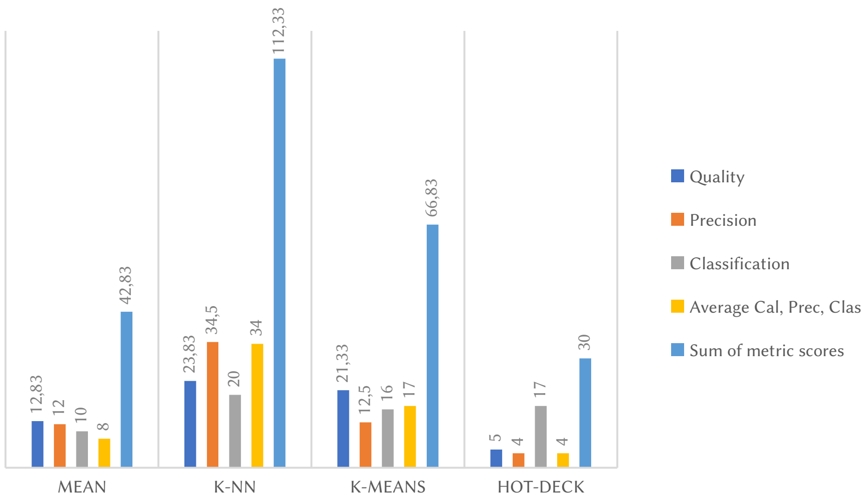

Fig. 5. Overall scores obtained by the imputation methods according to the metrics used (Own elaboration).

For example, by sorting the imputation methods in ascending order by the values indicated in the first row, we obtain the order of goodness of the imputation methods Medias, k-NN, k-Means and Hot-Deck used to impute the "Iris" data set after the original "Iris" data set was amputated according to the MCAR mechanism/assumption, in a univariate pattern on 10% of the records.

The results presented in Table X for this metric and the number of times each imputation method came first, second, third and fourth in the order of goodness of fit to impute each of the 63 amputated data sets are summarized below.

The *Mean* imputation method came first, second, third and fourth in 8, 7, 19 and 29 times out of 63, respectively. Likewise, *k-NN* ranked first, second, third and fourth 34, 17, 9 and 3 out of 63 times. The *k-Means* method came first, second, third and fourth 18, 14, 19 and 12 times out of 63, and finally, *Hot-Deck* came first, second, third and fourth 4, 12, 18 and 29 times out of 63, respectively.

Fig. 4 shows the number of times that the Mean, k-NN, k-Means and Hot-Deck methods came *first* in order of goodness of fit with respect to the *arithmetic average* aggregation operator metric and considering MV mechanisms, patterns and percentages. Clearly, the k-NN method came out first in all cases, except in the case of a complex multivariate MV pattern, where the k-Means method came out first.

Table XI shows the results obtained by applying equations (3) and (4), defined in *Criterion 1*, to the values obtained in Tables IX and X, i.e., the *arithmetic average* of the values of the quality, precision, classification, and aggregate metrics obtained by each imputation method.

By ascending the values in Table XI, the imputation methods were obtained for each metric, according to their order of goodness of fit.

TABLE XI. Values of the Arithmetic Average Metrics (Own Elaboration)

Imputation Method	Metrics			
	Pro. ΔCal	**Pro. ΔPre**	**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

Regarding the *arithmetic average* of the values of the ΔCal metric (Pro. ΔCal), the k-Means, k-NN, Mean and Hot-Deck methods resulted according to their order of goodness of fit. Similarly, considering the *arithmetic average* of the values of the $\Delta Prec$ metric (Pro. $\Delta Prec$), the k-Means, k-NN, Mean and Hot-Deck methods were obtained, according to their order of goodness. However, considering *arithmetic average* of the values of the $\Delta Clas$ metric (Pro. $\Delta Clas$), the k-Means, k-NN, Hot-Deck and Mean methods resulted. Finally, with respect to the *arithmetic average* of the aggregate metric values (Pro. *Met. Agr.*), the k-Means, k-NN, Mean and Hot-Deck methods resulted according to their order of goodness of fit.

Table XII presents the scores obtained by the imputation methods that came first in the order of goodness of fit with respect to the metrics ΔCal, $\Delta Prec$ and $\Delta Clas$ considering the values obtained using equation (1) and presented in Table IX, i.e., considering *Criterion 2*.

Thus, for example, considering the order of goodness of IM given by the value of the ΔCal, $\Delta Prec$ and $\Delta Clas$ metrics in Table IX, with respect to the ΔCal metric, one point was assigned to the k-NN IM used to impute the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 10% of the records; similarly, with respect to the $\Delta Prec$ metric, the Mean imputation method scored one point when imputing the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 10% of the records.

Similarly, with respect to the ΔCal metric, 0.33 points were assigned to the IM by Mean, k-NN and k-Means used to impute the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 20% of the records.

Similarly, with respect to the $\Delta Clas$ metric, 0.5 points were assigned to the MI k-Means and Hot-Deck used to impute the amputated "Iris" dataset according to the MCAR mechanism, in complex multivariate pattern, in 10% of the records.

Table XIII presents the scores obtained, considering *Criterion 2*, by the imputation methods that resulted first in the order of goodness of fit with respect to the aggregate metric considering the values obtained by equation (2) (metric ΔQ average of the metrics ΔCal, $\Delta Prec$ and $\Delta Clas$) and systematized in Table X.

Thus, for example, considering the order of goodness of IM given by the value of the ΔQ metric in Table X, a point was assigned to the k-NN IM used to impute the amputated "Iris" data set according to the MCAR mechanism, in univariate pattern, in 10% of the records.

Finally, Table XIV summarizes the score obtained by each IM for each metric, resulting from applying equations (5) and (6) to the data in Tables XII and XIII, and the overall score obtained by each imputation method, resulting from applying equation (7) to Table XIV.

TABLE XII. Scores Obtained for Each Metric (Own Elaboration)

Characteristics of Amputated Datasets				Scores obtained by each Imputation Method for each metric											
				Media			k-NN			k-Means			Hot-Deck		
Mechanism	**Type**	**Pattern**	**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			**0.1**		1.00		**1.00**		1.00						
MCAR		**univa**	**0.15**					1.00		1.00		1.00			
MCAR			**0.2**	**0.33**	1.00		**0.33**		1.00	**0.33**					
MCAR			**0.1**					1.00		1.00					1.00
MCAR		**multiva2**	**0.15**										1.00	1.00	1.00
MCAR			**0.2**			1.00	1.00				1.00				
MCAR			**0.1**		1.00		1.00		**0.50**			**0.50**			
MCAR		**multiva3**	**0.15**						1.00				1.00	1.00	
MCAR			**0.2**	1.00	1.00										1.00
MAR	**LEFT**		**0.1**	1.00	1.00	1.00									
MAR	**LEFT**	**univa**	**0.15**	1.00	1.00										1.00
MAR	**LEFT**		**0.2**	1.00	1.00										1.00
MAR	**LEFT**		**0.1**			1.00	1.00	1.00							
MAR	**LEFT**	**multiva2**	**0.15**			1.00	1.00	1.00							
MAR	**LEFT**		**0.2**			0.50		0.50	0.50	1.00	0.50				
MAR	**LEFT**		**0.1**				1.00	1.00				1.00			
MAR	**LEFT**	**multiva3**	**0.15**							1.00	1.00	1.00			
MAR	**LEFT**		**0.2**							1.00	1.00	1.00			
MAR	**MID**		**0.1**				1.00	1.00	1.00						
MAR	**MID**	**univa**	**0.15**	1.00	1.00				1.00						
MAR	**MID**		**0.2**	1.00				1.00	1.00						
MAR	**MID**		**0.1**				1.00	1.00							1.00
MAR	**MID**	**multiva2**	**0.15**				1.00	1.00	1.00						
MAR	**MID**		**0.2**				1.00	1.00							1.00
MAR	**MID**		**0.1**							1.00	1.00				1.00
MAR	**MID**	**multiva3**	**0.15**					1.00		1.00		1.00			
MAR	**MID**		**0.2**					1.00		1.00		1.00			
MAR	**RIGHT**		**0.1**			1.00	1.00	1.00							
MAR	**RIGHT**	**univa**	**0.15**	0.50			0.50	1.00	1.00						
MAR	**RIGHT**		**0.2**	1.00	1.00				1.00						
MAR	**RIGHT**		**0.1**								1.00		1.00		1.00
MAR	**RIGHT**	**multiva2**	**0.15**				1.00	1.00							1.00
MAR	**RIGHT**		**0.2**					1.00	1.00	1.00					
MAR	**RIGHT**		**0.1**		1.00					1.00		1.00			
MAR	**RIGHT**	**multiva3**	**0.15**					1.00		1.00					1.00
MAR	**RIGHT**		**0.2**					1.00		1.00		1.00			
MNAR	**LEFT**		**0.1**	1.00	1.00	1.00									
MNAR	**LEFT**	**univa**	**0.15**	1.00				1.00	1.00						
MNAR	**LEFT**		**0.2**	1.00				1.00	1.00						
MNAR	**LEFT**		**0.1**			1.00	0.50	1.00		0.50					
MNAR	**LEFT**	**multiva2**	**0.15**			1.00	1.00	1.00							
MNAR	**LEFT**		**0.2**				0.50	1.00	0.50	0.50					0.50
MNAR	**LEFT**		**0.1**				1.00	1.00	0.50			0.50			
MNAR	**LEFT**	**multiva3**	**0.15**				1.00	1.00				0.50			0.50
MNAR	**LEFT**		**0.2**							1.00				1.00	1.00
MNAR	**MID**		**0.1**				0.50	1.00	1.00				0.50		
MNAR	**MID**	**univa**	**0.15**	1.00	1.00				1.00						
MNAR	**MID**		**0.2**	1.00				1.00	1.00						
MNAR	**MID**		**0.1**				1.00	1.00							1.00
MNAR	**MID**	**multiva2**	**0.15**				1.00	1.00	1.00						
MNAR	**MID**		**0.2**				1.00	1.00							1.00
MNAR	**MID**		**0.1**				1.00				1.00	1.00			
MNAR	**MID**	**multiva3**	**0.15**							1.00	1.00	1.00			
MNAR	**MID**		**0.2**							1.00	1.00	1.00			
MNAR	**RIGHT**		**0.1**			0.50	0.50	1.00		0.50		0.50			
MNAR	**RIGHT**	**univa**	**0.15**							1.00	1.00	1.00			
MNAR	**RIGHT**		**0.2**			1.00	1.00							1.00	
MNAR	**RIGHT**		**0.1**								1.00		1.00		1.00
MNAR	**RIGHT**	**multiva2**	**0.15**					1.00	1.00	1.00					
MNAR	**RIGHT**		**0.2**				1.00	1.00	1.00						
MNAR	**RIGHT**		**0.1**							0.50	1.00	1.00	0.50		
MNAR	**RIGHT**	**multiva3**	**0.15**							1.00	1.00	1.00			
MNAR	**RIGHT**		**0.2**					1.00		1.00					1.00

TABLE XIII. Score Obtained With Respect to the Arithmetic Average Metric (Own Elaboration)

Characteristics of Amputated Datasets				Imputation Method			
				Media	k-NN	k-Means	Hot-Deck
Mechanism	Type	Pattern	MR	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$
MCAR			0.1		**1.00**		
MCAR		univa	0.15			1.00	
MCAR			0.2		1.00		
MCAR			0.1			1.00	
MCAR		multiva2	0.15				1.00
MCAR			0.2		1.00		
MCAR			0.1		1.00		
MCAR		multiva3	0.15				1.00
MCAR			0.2	1.00			
MAR	LEFT		0.1	1.00			
MAR	LEFT	univa	0.15	1.00			
MAR	LEFT		0.2	1.00			
MAR	LEFT		0.1		1.00		
MAR	LEFT	multiva2	0.15		1.00		
MAR	LEFT		0.2		1.00		
MAR	LEFT		0.1		1.00		
MAR	LEFT	multiva3	0.15			1.00	
MAR	LEFT		0.2			1.00	
MAR	MID		0.1		1.00		
MAR	MID	univa	0.15	1.00			
MAR	MID		0.2		1.00		
MAR	MID		0.1		1.00		
MAR	MID	multiva2	0.15		1.00		
MAR	MID		0.2		1.00		
MAR	MID		0.1			1.00	
MAR	MID	multiva3	0.15			1.00	
MAR	MID		0.2			1.00	
MAR	RIGHT		0.1		1.00		
MAR	RIGHT	univa	0.15		1.00		
MAR	RIGHT		0.2	1.00			
MAR	RIGHT		0.1			1.00	
MAR	RIGHT	multiva2	0.15		1.00		
MAR	RIGHT		0.2		1.00		
MAR	RIGHT		0.1			1.00	
MAR	RIGHT	multiva3	0.15			1.00	
MAR	RIGHT		0.2			1.00	
MNAR	LEFT		0.1	1.00			
MNAR	LEFT	univa	0.15		1.00		
MNAR	LEFT		0.2		1.00		
MNAR	LEFT		0.1		1.00		
MNAR	LEFT	multiva2	0.15		1.00		
MNAR	LEFT		0.2		1.00		
MNAR	LEFT		0.1		1.00		
MNAR	LEFT	multiva3	0.15		1.00		
MNAR	LEFT		0.2				1.00
MNAR	MID		0.1		1.00		
MNAR	MID	univa	0.15	1.00			
MNAR	MID		0.2		1.00		
MNAR	MID		0.1		1.00		
MNAR	MID	multiva2	0.15		1.00		
MNAR	MID		0.2		1.00		
MNAR	MID		0.1			1.00	
MNAR	MID	multiva3	0.15			1.00	
MNAR	MID		0.2			1.00	
MNAR	RIGHT		0.1		1.00		
MNAR	RIGHT	univa	0.15			1.00	
MNAR	RIGHT		0.2		1.00		
MNAR	RIGHT		0.1				1.00
MNAR	RIGHT	multiva2	0.15		1.00		
MNAR	RIGHT		0.2		1.00		
MNAR	RIGHT		0.1			1.00	
MNAR	RIGHT	multiva3	0.15			1.00	
MNAR	RIGHT		0.2		1.00		

TABLE XIV. Scores Obtained by IM for Each Metric (Own Elaboration)

Imputation Method	Score obtainedfor each metric				
	$o_1(\Delta Cal)$	$o_2(\Delta Prec)$	$o_3(\Delta Clas)$	$O(\Delta Q)$	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

The values in Table XIV are plotted in Fig. 5.

By sorting the values in Table XIV in descending order, the imputation methods for each metric were obtained, according to their order of goodness of fit to impute the set/group of data sets (files).

Regarding the values of the ΔCal metric, the k-NN, k-Means, Mean and Hot-Deck methods, according to their order of goodness of fit, were better. Similarly, considering the values of the $\Delta Prec$ metric, the k-NN, k-Means, Mean and Hot-Deck methods, according to their order of goodness of fit, were obtained. However, considering the values of the $\Delta Clas$ metric, the k-NN, Hot-Deck, k-Means and Mean methods resulted. Finally, as for the values of the arithmetic average metric ΔQ, the k-NN, k-Means, Mean and Hot-Deck methods resulted according to their order of goodness.

Finally, considering the values of the overall score metric G, the k-NN, k-Means, Mean and Hot-Deck methods were ranked according to their order of goodness of fit.

Summarizing, the best imputation methods globally considered turned out to be k-Means and k-NN according to criterion 1, k-NN and k-Means according to criterion 2 of this proposal, and k-Means and k-NN according to the calculation methodology based on the square root of the mean square error shown in [11].

V. Conclusions

This paper has presented an innovative methodology to evaluate the performance of imputation methods, based on metrics derived from data mining processes, instead of the generally used methods based on the root mean square error and its derivatives.

The proposed methodology is applicable to data sets to which data mining processes (e.g. regressions) can be applied, which will provide the information with which the different metrics will be calculated.

The working environment implemented to perform the amputation and subsequent imputation experiments described in [11] was appropriate. It has facilitated the management of the respective original, amputated and imputed files, to which the data mining processes performed with ISW V.9.7 software was applied.

The proposed methodology and the metrics presented have made it possible to arrive at an overall value (since it takes into account all the variables that were amputated and then imputed by various methods), indicative of the performance of each imputation method, expressed in comparable values (since it is based on normalized values of data mining metrics), integrating the results of a multitude of tests representative of different scenarios, with different percentages, diversity of patterns, considering also the three most frequent mechanisms of occurrence of missing data.

The results obtained with the proposed methodology in its different variants of metrics (differences in absolute values and scores) are slightly different. However, they concur that the best imputation methods globally considered are k-NN and K-Means, which also coincides with the global results obtained by the metrics indicated in [11].

The proposed methodology, by contemplating several metrics derived from the DMPs, allows working with only one of them or with all of them simultaneously, to determine the best imputation methods for a given scenario. Moreover, it can be applied to the evaluation of any imputation method, since it works with the imputed files and not with the methods themselves.

This methodology makes it possible to use the DMM generated to evaluate the imputation methods, to perform *a posteriori* predictive data mining process, which constitutes an added value of this proposal.

A. Future Lines of Work

To extend the scope of the proposed methodology, we plan to develop new metrics and indicators. We will use combined algorithms based on mean square error and data mining algorithms applied on the complete files, and then on the files imputed by different methods after having been amputated by different mechanisms.

Acknowledgment

This work has been developed in the context of the Research Project code SIUTIRE0005231TC, of the Resistencia Regional Faculty of the National Technological University, Argentine. We would like to thank the Co-Director of this project, Dr. Marcelo Karanik, for reviewing this work, Dr. Jorge Emilio Monzón, for reviewing the English version, and the scholarship holder, student Alejandro Nadal, for his effort and dedication to the multiple data mining processes.

References

[1] P. Schmitt, J. Mandel, and M. Guedj, "A comparison of six methods for missing data imputation," *Journal of Biometrics & Biostatistics*, vol. 6, no. 1, pp. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[2] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 5, pp. 692–709, 2007.

[3] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," in *Proc. 2016 International Conference on Data Science and Engineering (ICDSE)*, 2016.

[4] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, 2019, doi: 10.1109/access.2019.2891360.

[5] Y. Liu and V. Gopalakrishnan, "An overview and evaluation of recent machine learning imputation methods using cardiac imaging data," *Data*, vol. 2, no. 8, pp. 1–15, 2017, doi: 10.3390/data2010008.

[6] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[7] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *IAENG Transactions on Engineering Technologies, Lecture Notes in Electrical Engineering*, vol. 229, pp. 245–257, 2013.

[8] J. M. Jerez, I. Molina, E. A. García-Laencina, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, pp. 105–115, 2010.

[9] N. Z. Abidin, A. R. Ismail, and N. A. Emran, "Performance analysis of machine learning algorithms for missing value imputation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 6, pp. 442–447, 2018.

[10] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and Information Systems*, vol. 32, no. 1, pp. 77–108, 2012.

[11] C. R. Primorac, D. L. La Red Martínez, and M. E. Giovannini, "Metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador," *European*

Scientific Journal (ESJ)*, vol. 16, no. 18, pp. 61–92, 2020.

[12] C. Ballard, J. Rollins, J. Ramos, A. Perkins, R. Hale, A. Dorneich, E. C. Milner, and J. Chodagam, *Dynamic warehousing: Data mining made easy*, IBM Corporation, 2007.

[13] G. Madhu and T. V. Rajinikanth, "A novel index measure imputation algorithm for missing data values: A machine learning approach," in *2012 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2012, doi: 10.1109/ICCIC.2012.6510198.

[14] D. L. La Red Martínez, M. Karanik, M. Giovannini, M. E. Báez, and J. Torre, "Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos," *Revista Científica Iberoamericana de Tecnología Educativa - Scientific Journal of Educational Technology*, vol. 5, no. 2, pp. 70–83, 2016.

[15] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed., Amsterdam, Netherlands: Elsevier, 2012.

[16] R. J. Roiger, *Data mining: A tutorial-based primer*, 2nd ed., Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2016.

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[18] I. Kononenko and M. Kukar, *Machine learning and data mining: Introduction to principles and algorithms*, Amsterdam, Netherlands: Elsevier, 2007.

[19] S. Chakrabarti, E. Cox, E. Frank, R. H. Güting, J. Han, X. Jiang, M. Kamber, S. S. Lightstone, T. P. Nadeau, R. E. Neapolitan, D. Pyle, M. Refaat, M. Schneider, T. J. Teorey, and I. H. Witten, *Data mining: Know it all*, Amsterdam, Netherlands: Elsevier, 2009.

[20] C. Ballard, N. Harris, A. Lawrence, M. Lowry, A. Perkins, and S. Voruganti, *InfoSphere Warehouse: A robust infrastructure for business intelligence*, IBM Corporation, 2010.

[21] D. L. La Red Martínez and J. C. Acosta, "Aggregation operators review - mathematical properties and behavioral measures," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 7, no. 10, pp. 63–76, 2015.

[22] P. Chan Chiu, A. Selamat, O. Krejcar, K. Kuok Kuok, E. Herrera-Viedma, and G. Fenza, "Imputation of rainfall data using the sine cosine function fitting neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 39–48, 2021.

David L. la Red Martínez

David L. la Red Martínez obtained a master's degree in computer science and informatics at the National University of the Northeast - UNNE (Argentina) in 2001 and a PhD in systems engineering and computer science at the University of Malaga - UMA (Spain) in 2011. He is currently a full professor at the National University of the Northeast and at the National Technological University - UTN and director of the "Operating Systems and ICT" Research Group at UNNE. For more than 20 years, he has worked in research projects both at national and international level. His research has focused on distributed systems, decision support systems, data communications, ICT in education and educational and health data mining.

Carlos R. Primorac

Carlos R. Primorac obtained a degree in computer science at the National University of the Northeast - UNNE (Argentina) in 2015. He is currently a professor at the National University of the Northeast and a member of the "Operating Systems and ICT" Research Group at UNNE. For more than 6 years, he has worked in research projects at national level. His research has focused on distributed systems, data communications and data imputation.