

Data-Driven Machine Learning Approach for Predicting Missing Values in Large Data Sets: A Comparison Study

Ogerta Elezaj¹, Sule Yildirim¹, and Edlira Kalem²(✉)

¹ Faculty of Computer Science and Media Technology,
Norwegian University of Science and Technology, 2815 Gjøvik, Norway
ogertae@stud.ntnu.no, sule.yildirim@ntnu.no

² Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK
e.kalemi@surrey.ac.uk

Abstract. Pre-processing of large scale datasets in order to ensure data quality is a very important task in data mining. One of the serious threats to data quality is the lack of data collected during field experiments, which negatively affects the data quality. The missing data usually have significant effects in many real-life pattern classification scenarios, especially when it leads to biased parameter estimates but also disqualify for analysis purposes. The process of filling in the missing data based on other valid values of rest of the variables of a data set is known as the imputation process. In this paper, we present a new data-driven machine learning approach for imputing the missing data. Even though Machine Learning methods are used in order to impute missing data in the literature, it is difficult to decide on a single method to apply on a given data set for imputation. This is because imputation process is not considered as science but as art that focuses on choosing the best method with the least biased value. For this reason, we compare different machine learning methods, such as decision tree (C4.5), Bayesian network, clustering algorithm and artificial neural networks in this work. The comparison of the algorithms indicates that, for predicting categorical and numerical missing information in large survey data sets, clustering method is the most efficient out of the others methods found in literature. A hybrid method is introduced which combines unsupervised learning methods with supervised ones based on the missing ratio, for achieving a data imputation with higher accuracy. Additionally, some statistical imputation methods such as MeanMode, Hot-Deck have been applied emphasizing their limitations in large scale datasets in comparison to the machine learning methods. A comparison of all above mentioned methods, traditional statistical methods and machine learning methods has been made and conclusions are drawn for achieving data imputation with higher accuracy in data sets of large scale survey. Also, another objective of these experiments is to discover the effect of balancing the training data set in the performance of classifiers. All methods are tested to a real world data set, population and housing census.

Keywords: Imputation · Data pre-processing · Clustering · Decision tree
Neural networks · Hot-Deck

1 Introduction

Information plays an important role in today's modern society and it is considered as a valuable asset for companies and organizations as it can be used for future prediction or decision-making processes. The rapid growth of information coming from different sources and in different formats (such as text, numeric, image, video audio and so on) requires interdisciplinary knowledge to be processed and used for decision making. Nowadays, machine learning techniques are playing an important role in knowledge discovery in databases which is a process that aims to identify valid data models. Identifying valid data models is complex and cannot be managed only based on experience or intuition. The successful implementation of this process requires the application of technologies and techniques that are based on analysis and research. Predictive methods such as classification, regression and time series analyses are used to learn models that are used for classification, i.e. predicting a certain class. Descriptive methods such as clustering, association or link analysis deal with extracting interesting, understandable and interpretable knowledge in order to discover regularities in the data and to uncover patterns. In the discipline of knowledge discovery, the term "process" is always used to indicate that knowledge discovery consists of several steps. There are 5 main steps in this process: (1) understanding the domain and selection of relevant data sets; (2) data pre-processing; (3) data transformation (4) data mining; (5) interpretation and evaluation [1]. Data pre-processing is a complex process that faces a number of challenging aspects. Most of these challenges are common such as scalability, dimensionality and heterogeneous data, data quality, data ownership and distribution and data privacy preservation. In this paper, we will be focus on applying machine learning techniques to pre-processing process for dealing with the data quality challenge. The real-world data sets are considered as dirty data sets due to inconsistency among values, missing values or being noisy data sets. The data can be incomplete when the data are not collected, when there are differences between the time when the data are collected and when the data are analyzed or because of human error or hardware/software malfunction during data collection and data capturing. A data set is considered as noisy when it contains errors for different reasons such as faulty data collection instruments, data entry errors or data transmission errors. In these conditions, when working with real world data set, lacking quality data means lacking quality in mining results and consequently misleading decision making based on these results. Most of the work done by organizations when they implement their data warehouse includes data extraction, data cleaning and data transformation. The missing data can be discovered by searching for null values in a data set but also wrong data or outliers are considered and treated as missing data during the data pre-processing phase.

Missing or partially answered questions, in the field of statistics are referred to as item non-response in contrast to unit non-response. The unit non-response can happen when the person refuses to be part of a survey or when he cannot be contacted during the field work phase (Fig. 1). Weighting procedures are used for handling the unit non-response. Item non-responses occur frequently in large scale surveys and in this kind of surveys it can never be prevented totally. They occur in different domains due to

different reasons such as partial refusal of respondent, malfunctions or failures of systems or sensors used to collect and process data, hardware unavailability, etc.

Unit non response Item non response

v1	v2	v3	v4	v1	v2	v3	v4
X	X	X	X	X	X	.	X
X	X	X	X	.	X	X	.
X	X	X	X	X	.	X	.
X	X	X	X	X	X	.	.
.	X	X	X

Fig. 1. Unit non response and item non response

In this situation, researchers have to handle the missing data through understanding them and imputing them in a way to achieve unbiased parameter estimates and accurate standard errors. The amount of missing data is considered to be an important quality indicator for survey data processing [2]. Also, analysis of the effect of imputation is undertaken as part of assessing quality of the data collection in EUROSTAT Quality Assurance Framework [3]. Many techniques have been developed to handle the imputation process. Deletion techniques are the most traditional techniques used to handle missing data by statisticians and they are based on the approach to discard the missing cases from analyses. These techniques are easy and they are recommended to be used in the cases when the data set contains a small number of missing records. In contrast, most of real world data sets contain a considerable amount of missing records and applying these methods can lead to biased parameter estimates. So, in these data sets different techniques should be applied to predict the missing information based on the available information. Several statistical techniques are reported for imputation such as mean/mode, regression methods, and multiple imputation methods. Also, machine learning methods are used to impute the missing data such as decision trees, Bayesian networks, K-Nearest Neighbor, multilayer perceptron and clustering techniques. The decision about what methods to use is not easy and depends on many factors such as the data sets itself, the missing data mechanism, missing patterns and data types that are going to be imputed. This research presents a comparison of some statistical methods and machine learning techniques used for imputation in real world data sets with artificially created missing records. In this paper, we use different missing ratio in the data from 25% to 0.5% for the attributes of the selected data set. This paper is divided into following sections such as introduction, related work, system design, performance measurements and results, and conclusions.

2 Related Work

In a data set, there can be non-observed values or unknown values called as missing values. Most of the missing data occurs during the data collection phase where there can be a lack of information due to uncontrolled system failure or the data are not reported due to privacy or refusal issues. Also, in some cases the values can be lost during the

data processing phase. As the presence of missing data in knowledge discovery in databases process is harmful, leading to wrong knowledge extraction and wrong decision making, procedures to handle them should be in place.

The problem of missing data in an experimental design was firstly introduced in 1933 by Allan and Wishart [4]. Their solution was based on estimating the missing data based on iteration methods. This approach is not useful when the number of missing observations is high due to computational time needed during iteration process. In 1937, it was Bartlett who used the analyses of covariance among variables to impute the missing data based on the degree of relationship [5]. Since then, many researchers have proposed different methods to solve the problem of missing data. The selection of the methods is based on missing patterns, missing mechanisms (MCAR, MAR, MNAR) and data set characteristics (numerical or categorical).

The missing pattern of a data set gives information about the structure of the missing data. There are two missing patterns, monotone and arbitrary pattern. In the arbitrary pattern, the missing data are interspersed among full data values while in monotone patterns the missing data are at the end, from left to right and there are not gaps between the missing data and full data, illustrated in Fig. 2 [6]. A monotone missing pattern is easier and more flexible in the selection of the techniques for imputation of data. The first step that a researcher does when she wants to start an imputation process is to analyze the missing patterns via available procedures.

Missing monotone				Missing arbitrarily			
v1	v2	v3	v4	v1	v2	v3	v4
X	x	x	x	x	x	.	x
X	x	x	x	.	x	x	.
X	x	x	.	x	.	x	.
X	x	.	.	x	x	.	.
X	x	x	x

Fig. 2. Missing monotone and missing arbitrary (v: variable, missing, x: existing variable value)

Missing data mechanism, classified as Missing Completely at Random (MCAR), Missing At Random (MAR) or Not Missing At Random (NMAR), is seen as a key factor for developing a successful imputation procedure. MCAR is a mechanism where the probability of a record to have a missing value is independent of the existing observed data or the missing one. This case is the highest level of randomness and suggested approach is not to take into account the records with missing values during data processing. In case that the number of records with missing value is too high, using this approach can reduce the data set and end up with loss of information. MAR mechanism occurs when the missing data depends on the observed data but it does not depend on the missing value itself. Most of the real-world data sets follow this mechanism and different imputation methods exist to predict the missing information based on the existing data. NMAR occurs when the probability of having missing data depends on the value of the missing attribute [7, 8].

Another factor to take into consideration during the imputation process is the type of the variables that we are imputing. As there are many imputation methods, some of

them can handle only categorical variables (nominal or ordinal), others only numerical variables (discrete or continues) and some other methods can handle both type of variables, numerical and categorical ones. Also, during the imputation process an important aspect to take into account is the capability of methods to deal with complex data sets, in terms of number of variables and number of records.

In state of the art, various imputation techniques are available. Traditional imputation techniques are based on statistical methods and are model based. Handling missing data in surveys using statistical imputation techniques means defining models based on the non missing data sets and doing inferences based on probability distribution of the model [9]. The simplest method to impute missing values is the mean/mode approach [10]. Also, other methods such as Hot-Deck and machine learning are well used for the imputation. However, as the focus of this paper is to find out the best method to be used for imputing survey database such as census, the evaluation measurement should be done based on the accuracy of classification for each categorical target variable and the ability of each imputation strategy in restoring the original marginal distribution of the numerical target variable. After a good overview of some recent research in imputation field, we found out different machine learning algorithms used for imputation and the comparison among them is based on univariate analyses of data, putting apart the preservation of joint distribution.

Bayesian networks are considered as efficient methods based on some previous studies. In 2004, Di Zio suggested the use of Bayesian networks for imputing missing values. He applied this methodology in a subset of 1991 UK population census [11] and concluded that this method improves the consistency and preservation of joint distribution.

Recently, in [12] authors experimented with several clustering algorithms to impute a financial data set and concluded that the k means algorithm suit the imputation process in larger data sets better in terms of achieving a higher accuracy compared to other algorithms.

To the best of our knowledge, there is a lack of comparing supervised and unsupervised machine learning methods with statistical methods used in imputation process in terms of univariate data characteristics and preservation of marginal distribution. Despite the considerable efforts that have been done by researchers in the imputation field, there is not a broad consensus among them regarding the most appropriate training model or the parameters to use for these models, in order to improve the accuracy of the estimated values. From our review, we find out that there is not a machine learning algorithm that outperforms the others. This happens because the performance of imputation does not only depend on the amount of missing data, but also depends on the missing data structure, the missing data mechanism, the type of variables that are going to be imputed and the nature of the data set that is going to be imputed.

We conducted many experiments with population data estimating the missing values with most used models in this field, to find the most promising model for our situation.

3 System Design

As indicated previously, in this study, we impute the missing values of a census data set using different machine learning methods in order to compare their performance. In this section, we describe the processes of data collection, preparation and the machine learning algorithms we selected to test.

3.1 Data Source and Data Preparation

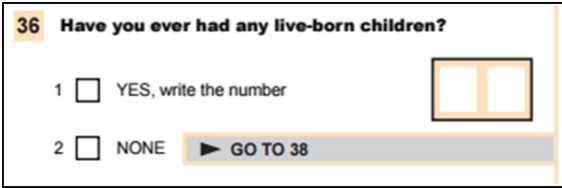
In this work, as an experiment dataset, we use the Albanian Population and Housing Census, data collected for the entire population of Albania in 2011. This is the main statistical activity carried out by a country with the aim to count all the residents collecting accurate demographical, economic and social data, providing a clear picture of the social-economic situation in the country [13]. This data set is very important for the policy making in the country because it gives changes in population from a social-economic perspective. Meanwhile, as it is a complex and large-scale activity, the reliability of the data depends on the whole data processing steps. As the scope of this paper is the process of imputation, we will be focused only on handling missing data and not on the other type of errors.

The data set in total contains 2,800,138 records which is the total population of the country in 2011, the latest enumeration process. In our study, we have selected only female aged over 15 years old and we have chosen two variables to impute, a categorical and a numerical one (Table 1). This paper deals with how to impute a numerical variable and a categorical variable through its dependence on other numerical and/or categorical variables.

Table 1. The number of records by gender and age

Total population	Female	Female >15 age old
2,800,138	1,397,079	1,094,429

We decided to impute two different type of variables due to the fact that most of the ML algorithms are data type dependent (Fig. 3). The categorical variable is “having or not any live-born children” and the numerical one is “number of children live-born”.



36 Have you ever had any live-born children?

1 YES, write the number

2 NONE **▶ GO TO 38**

Fig. 3. Question 36 of the population and housing census questionnaire

In surveys, the missing data follows a missing at random mechanism (MAR) and not missing completely at random (MCAR) due to the fact that some people are more

or less likely to have a valid response. For example, in censuses some questions had a higher probability to be left blank for example number of children live-born for younger women. The selection of variables that are used during the imputation phase are crucial for the entire process, as the main desirable features of the process is to reduce biases in the final results arising from non responses in the data sets. A set of variables that have a good explanatory power with respect to fertility are used in all the selected machine learning algorithms. They are: woman age, marital status, urban/rural area, employment status, level of education, woman disability status. The data set is anonymized to prevent direct/indirect identifications of persons. All the simulations are conducted using the same data set and the same variables. The data set which contains 1,094,430 records is a final data set without missing values. For our experiment, we have created artificially missing values which later are imputed with different methods. Since the original value of the artificially missing data created are known values, we can evaluate the performance of the machine learning algorithms and do a comparison among them. As the imputation procedure depends not only on the missing mechanism, but also on the amount of missing, data sets with different missing ratios (50%, 20% and 1%) are created as shown in Table 2. These data sets are generated as follows: the original data set is ordered randomly and random missing values are generated for the two variables that are going to be imputed. The missing mechanism of the data sets is MCAR. The missing model used is Uniformly Distributed (UD) where each variable has equal number of missing values.

Table 2. Settings of missing data simulations

Missing mechanism	No of attributes having missing value	Missing ratio	Missing model	No of records per data set
MCAR	2	50%–25% per variable	Uniformly Distributed	1,094,429
		20%–10% per variable		
		1%–0.5% per variable		

Usually in the problem of class prediction, the real-world data set are unbalanced which has an impact in the performance of the classifier during the training phase. The problem of unbalanced data is a crucial problem in many domains such as in medical data sets when rare disease is predicted, in fraud detection, network intrusions etc.

To handle the unbalance class distribution data set problem, external techniques are required during the pre-processing phase. In literature, there are two strategies to deal with the class imbalance problem, data-level strategy and algorithm level strategy [15].

The methods at data level approach adjust the data sets with adding or removing records in order to reduce the discrepancy among classes, while the algorithm level strategy is focused on tuning classifier algorithms to improve the learning process in respect to the minority class. The most known data-level strategy are over-sampling and down-sampling which both has the focus to reduce the imbalance ratio of the data set

used in the training phase. When we apply down-sampling we remove part of the records belonging to the majority class until each category is represented by the same number of records, and when we apply up-sampling we duplicate records from the minority class [16]. In the case when we apply down-sampling approach, we reduce the chance of over-fitting but there is a high risk to lose potential information from the training data set as considerable amount of records are excluded during the learning phase.

When we apply the second approach, the duplication of the records of minority class do not provide additional information about the class, so it is not dealing with the lack of the data from minority class.

For doing the experiment, we have split the data set into two, one with missing record and one without missing records. For taking into account the problem of unbalance classes for the imputation of categorical variables, we have used three different data sets, down sampling, no sampling and over sampling for each of the missing patterns.

The principle of analysis is shown in Fig. 4.

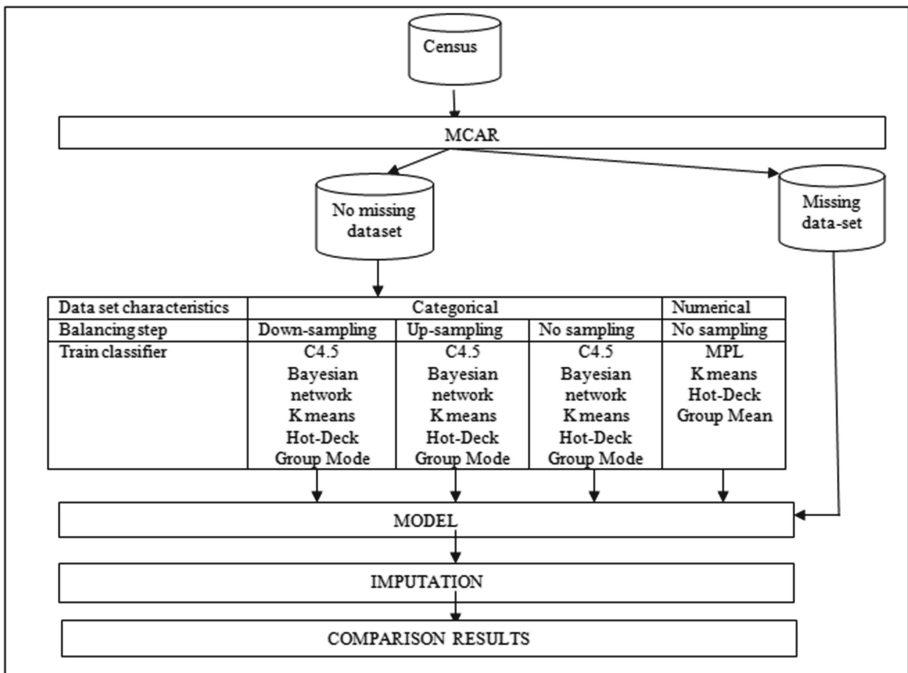


Fig. 4. Comparison architecture

3.2 Methods for Imputation of Missing Values

In literature, there exist various imputation methods starting from traditional one to sophisticated model-based imputation techniques. The methods used in this paper are three unsupervised imputation algorithms: Mean/Mode, Hot-Deck and clustering, and

three supervised machine learning algorithms: decision trees, Bayesian Network and neural networks. All these methods are briefly introduced in this section.

Mean/Mode imputation

When this method is applied, the missing values are imputed with the mean/mode of the corresponding variable [19]. So, the imputed value y_i^* for a missing value of y_i is calculated by the observed mean for the numerical variables:

$$y_i^* = \frac{\sum_{k \in \text{obs}} y_k}{p} \tag{1}$$

where y_k is the observed value of k^{th} record and p is the total number of non missing values of target variable y . The main limitation of this method is that it does not take into account auxiliary information from other variables during the imputation. For that reason, in this paper we applied Group Mean/Mode imputation which takes into account the variation among groups. When this method is applied, the data are grouped into similar groups based on auxiliary variables and a missing value is replaced by the mean/mode value of the corresponding group.

Hot-Deck imputation

It is based on the nearest neighbor imputation (NNI) and we will introduce this method with a simple case. Let's consider a data set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and let's suppose that from n y -values there are p observed values and the other values are missing ($m = n - p$). Let's suppose that the missing ones are $y_{p+1}, y_{p+2}, \dots, y_n$. The Hot-Deck method imputes the missing value y_i , where $p + 1 \leq i \leq n$ by y_j where $1 \leq j \leq p$. The nearest neighbor methods consider y_j as the nearest neighbor of y_i , if j satisfies:

$$|x_j - x_i| = \min |x_k - x_i| \tag{2}$$

where $1 \leq k \leq p$. In the case where there is more than one nearest neighbor of i , one of them is randomly selected.

This procedure is repeated until all the missing values are imputed.

The most widely used method for imputing statistical surveys is donor Hot-Deck imputation which for each missing value m in a donor record d with similar characteristics is searched in the dataset [20]. This method can be applied for numerical and categorical variables and allows multiple imputations of all the missing variables of a record with information provided by the corresponding donor. The donor pool is defined based on minimizing a distance function (nearest neighbor Hot-Deck).

Decision tree

A decision tree composed by nodes and leaves is a supervised classifier that can handle both numeric and categorical variables used to split the instance space. In this study, we used C4.5 algorithm which is an evolution of ID3 and this algorithm is considered to be a good one for treating missing values using gain ratio which is calculated using entropy based measurement [21]. Using this method, the original data set is divided into two

sets, one with missing values and the other one without missing values. A decision tree is created with the data set without missing values and a missing variable is considered a target variable (class variable). The main limitation is that the target variable has to be a categorical variable and the tree is sensitive to the number of records used to split the nodes.

Bayesian network

A Bayesian network is a directed acyclic graph containing nodes and arcs. A node corresponds to a variable and an arc to the casual relationship among variables. Bayesian networks are used in imputation of missing values because of their ability to deal with the problem of preservation of the variable distribution in large scale data sets. This network can be seen as a hierarchical ordering of variables associated with their conditional probability. This hierarchical ordering allows us to impute a missing value conditionally based only on the variables that are directly related to it.

Clustering

Clustering consists in grouping similar objects into clusters, where objects belonging to a class are similar among them and dissimilar to the objects belonging to the other cluster. As the amount of data to be processed is increased rapidly, clustering is becoming a powerful technique for drawing useful patterns. There has been many clustering algorithms proposed in literature but k means is the most popular. K means used in our study, is a centroid based algorithm [22]. At the initial phase of the algorithm, the k number of clusters should be specified and for each cluster k centroids are randomly selected as initial centers.

Neural networks

Neural networks are created by small units called neurons which try to imitate the neural brain system. All the neurons of a network are connected among them via connections that are weighted. This is considered to be an adaptive system by learning to estimate the parameters based on the training records. During the training phase the network tries to adjust the connection weights to improve the class prediction based on the input signal. The neural networks are widely used when the data sets are high dimensional, when they contain noisy data or when there are complex and hidden patterns among variables. So, because of their ability to present non linear models, they are recently used in imputation of missing data. The neurons of a neural network are organized in layers. The first layer is known as the input layer and the last one as the output layer. There are different neural networks such as perceptron, multi layer perceptron, Kohonen features maps, SOM etc. In this study, we selected the multilayer perceptron as it is the most popular type of neural network. It contains multi layers and is a feed-forward network because cycles or loops within the same layer are not allowed and during the training phase; the weights are updated by mapping inputs to outputs [22].

As a summary, Table 3 shows a summary of methods used in imputation process illustrating their main benefits and limitations.

Table 3. Methods used for imputation process

Methods	Main benefit	Main limitation	Variable type	Computational cost
Mean/Mode	Valid when the amount of missing is low	Variance is reduced artificially	Categorical/ Numerical	Low
Hot-Deck	Variance is not reduced artificially	All possible combinations need to be computed	Categorical/ Numerical	Medium
Decision tree (C4.5)	Use all information for construction of the tree	Does not offer a complete data table	Categorical	High
Bayesian network	Use all information for construction of the tree	Does not offer a complete data table	Categorical	High
Artificial neural network	Maintain complex non linear relationship	Poor generalization in complex data sets	Categorical/ Numerical	Very high
Clustering (k means)	Performs well in large data sets	Parameters in the initialization phase	Categorical/ Numerical	High

4 Performance Measurements and Results

The proposed approach is implemented in Weka 3.8 and executed in a PC with Intel® Core i5 processor with 2.7 GHz speed and 8 GB of RAM. SAS software, Version 9.2 has also been used for processing. Hot-Deck imputation is done using CONCORD JAVA (CONtrollo e CORrezione dei Dati version with Java interface) software developed for data editing and imputation, and IDEA (Indices for Data Editing Assessment) software is used for calculating similarity indexes. These are open source software developed by Italian National Institute of Statistics (ISTAT) [18].

4.1 Algorithms Tuning

All the above mentioned methods were evaluated with 10-fold cross-validation within the Weka data mining library. A ten-way cross-validation approach was selected where the dataset was partitioned randomly into ten subsets of equal size. Nine of these datasets were used as the training set, and the induced tree was used for predicting on the tenth subset (test set).

C4.5 was used to learn and predict values for the target variables. Since this method is a supervised classifier, during the training phase the missing records were not included. We learn a pruned tree, using a confidence threshold of 0.25 with a minimum number of 2 instances per leaf. For K-means algorithm, we select Manhattan distance metric to

compute the distance between any two data objects, and the numbers of clusters k is specified using Elbow rule ($k = 10$).

Regarding the neural networks, we used the default heuristic values in all of our experiments because tuning them seemed to have little impact on the final results.

4.2 Evaluation Measures

The evaluation of all the machine learning algorithms has been done by comparing the original value and the imputed value by each algorithm. Imputation performance for the categorical variable is evaluated using different evaluation metrics such as classifier accuracy and root mean square error (RMSE). Accuracy and RMSE are defined as follow:

Accuracy or recognition rate: percentage of test set records that are correctly classified

$$\text{Accuracy} = \frac{(\text{TruePositive} + \text{TrueNegative})}{\text{All}} \quad (3)$$

The RMSE explores the average difference of actual values with the imputed values

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (P_i - O_i)^2}{N}} \quad (4)$$

Where P_i is the imputed value of i^{th} missing value ($1 \leq i \leq N$), O_i is the actual value of i^{th} artificially created missing value, and N number of artificially created missing values.

As imputing a numerical variable is not a classification problem like categorical variables, the evaluation measurements should be done at two levels: micro level preserving single values for each variable and at macro level preserving marginal distribution of variables. In our study, we used the following measurements:

- micro level: univariate analyses of the variable before and after imputation (mean, mode, first quartile, median, third quartile, minimum, maximum);
- macro level: Kolmogorov-Smirnov Distance(KS) index for measuring similarities of distribution of variable before and after imputation.

KS index is calculated based on the following cumulative distribution functions:

$$F_{\bar{O}_n}(t) = \frac{1}{N} \sum_{i=1}^N I(\bar{O}_i \leq t) \quad (5)$$

$$F_{P_n}(t) = \frac{1}{N} \sum_{i=1}^N I(P_i \leq t) \quad (6)$$

Before and after imputation are computed the value:

$$KS = \max_t \left(\left| F_{O_n}^-(t) - F_{P_n}(t) \right| \right) \tag{7}$$

The KS distance is equal to 0 when the distributions of the variable are equal before and after imputation and takes the maximum value to 1 when there is a maximum difference between the two distributions.

In cases when the experiment is large, which means that there is a huge number of records to be processed from real world data sets, Kovahi in [14] proposed to use 10-fold cross-validation. In the following supervised algorithm, we used this method, which randomly partitions the data set in 10 equal size subsamples, where each of the subsamples is used as a validation data.

As our dataset is not balanced (Table 4) within classes, another objective of these experiments is to discover the effect of balancing the training data set in the performance of classifiers.

Table 4. Frequency of classes

		Frequency	Percent
Class	1	798972	73.0
	2	295457	27.0
	Total	1094429	100.0

4.3 Results and Considerations

In this section, we are discussing the experimental results of the census data set, and some considerations are drawn.

For both variables imputed and for each missing data simulation conducted in the experiments Tables 3 and 4 give the performance measurements for all the algorithms applied in the imputation phase. Table 3 illustrates the performance measurements of the categorical variable whereas Table 5, the univariate characteristics and KS index are given.

Table 5. Performance of algorithms on census data set

Missing ratio	Balance approach	C4.5		Bayesian		K means		Hot-Deck		Group Mode	
		Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE
25%	no sampling	93.8552	0.2264	93.4441	0.2306	95.92721	0.2017	86.88	0.3622	83.84	0.2481
	down sampling	92.2622	0.2442	92.2322	0.2469	96.8758	0.1919	86.99	0.3606	83.77	0.2451
	up sampling	90.0247	0.2813	89.9771	0.2872	96.3174	0.1965	85.45	0.3814	82.99	0.2647
10%	no sampling	93.9672	0.2259	93.4229	0.2312	96.4695	0.1878	87.25	0.3570	87.93	0.2463
0.50%	no sampling	93.8526	0.2269	93.29	0.2348	91.5057	0.2914	88.91	0.3330	93.54	0.2541

It is clear that the best overall results based on accuracy and RMSE measurements are achieved when we use clustering algorithm K means where the accuracy is higher and RMSE is lower. The worst case is obtained when we apply traditional statistical methods for imputation such as Hot-Deck. Regarding mode imputation method, we did not apply it because the data set is unbalanced. If this method is applied in an unbalanced

data set, all the records are imputed with the value of the class that has the highest frequency. To avoid this problem, we applied the group mode method when we impute the missing class based on the available information from other variables. Due to the fact that we have selected variables that are highly correlated to fertility, the group mode method outperforms the Hot-Deck method in all the cases of missing ratios. Usually the Hot-Deck technique is considered as a better statistical method compared to the group mode because it allows multiple imputation based on the information coming from donors and also it is better in preserving the condition distribution of the imputed variable. When the record missing rate is very low, the mean method fits well in comparison to the other methods. In the other case, where the missing rate is 25% (273,449 records in our data set), the machine learning algorithms performs better in terms of accuracy of prediction. To test the effect of balancing the data over unbalanced data, we applied all the methods in three datasets, down sampling, up sampling and no sampling. We selected the case with highest missing ratio because in this case it makes more sense to test the balancing approach. When we compare the results of the methods used to impute it is evident that the model on unbalance data is more accurate compared to the balanced data sets for mostly of the methods. Also, the confusion matrix of all the machine learning algorithms shows that false positive and false negative values are lower when we use the unbalanced data set to predict the missing values. Only when we apply clustering methods, a very slight improvement in accuracy is evident in the down sampling approach. Our experiment concludes the fact that a 50:50 (down-sampling or over-sampling) balance ratio between classes in the training dataset does not improve the classifier performance. Also, other studies discovered the same finding, when the data set is large, balancing the data is not beneficial as a consequence of the fact that the size of samples for both classes is big enough containing significant information for each of them [17]. Also, as a conclusion from the results, we can find out that down-sampling approach may be more suitable in comparison to up-sampling due to the fact that the false positive cases are minimized. In particular, the k means imputation provides superior results across all the methods when the missing ratio is 25%. This method improved the classification accuracy by 2% compared to the Bayesian network and by 12% compared to group mode imputation (Table 6).

Table 6. Preservation of distribution and aggregates, Kolmogorov-Smirnov Index

MPL	25%					10%					0.5%				
	Original	MPL	K means	Hot-Deck	Group Mean	Original	MPL	K means	Hot-Deck	Group Mean	Original	MPL	K means	Hot-Deck	Group Mean
Mean	2.34	1.89	2.38	2.5	2.347	2.35	2.95	2.31	2.7	2.3473	2.34	2.7616	2.35	3.1	2.3724
STD	2.174	1.616	1.762	3.45	1.676	2.172	1.611	1.835	3.4	1.67649	2.178	1.82487	1.769	3.3	1.6780
Max	16	6	6	20	6	20	6	6	20	6	16	6	6	20	6
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q1	0	1	2	1	1	0	1	1	1	1	0	2	1	1	1
Q2	2	2	2	2	2	2	3	2	2	2	2	3	2	2	2
Q3	3	3	3	5	4	3	4	3	5	4	3	4	4	5	4
Mode	0	0	2	2	2	0	1	3	2	2	0	3	2	2	2
KS	-	0.151	0.013	0.167	0.066	-	0.157	0.075	0.155	0.068	-	0.188	0.104	0.099	0.073

According to the KS index, which measures the ability to restore the original marginal distribution, it seems that all the methods give good results (the index values

do not exceed 0.188). The lowest index value is achieved using the K means algorithms (0.013) when the missing ratio is 25%. We observe that K means performs better on a larger data set. When the missing data set is small, the traditional method of mean imputation performs better in terms of a smaller KS index and a mean closer to the original mean. Related to the mean preservation, multi layer perceptron algorithm produces negative bias for all the different missing percentage. The Group Mean method and K-means produce a slight bias on means.

An important aspect to analyze during imputation is the variability of the variable distribution measured by the standard deviation (STD). From the results, it is very evident the fact that all the methods produce consistent reduction of variability, except the Hot-Deck method. This can be explained by the fact that the other methods were not able to reproduce extreme original values. The maximum value imputed from the methods is 6 while the original data set has higher maximum values. All the imputation methods preserve variability better when the data sets are smaller due to the fact that, it is easier to control variations in these data sets. Regarding the use of neural networks for imputing the numerical variables, based on our experiment, we can conclude that they come with a greater computation and training complexity compared to all the other methods.

5 Proposed Imputation Approach

Based on the results, the proposed imputation approach is as shown in figure below (Fig. 5):

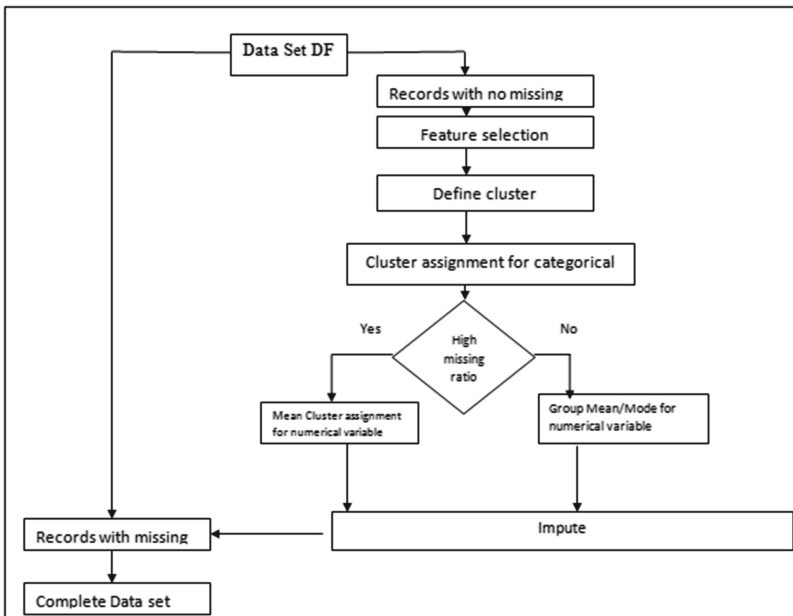


Fig. 5. The proposed imputation method

When dealing with data sets where there are missing values first a detailed analyze should be done regarding the missing mechanism and missing patterns. The data set has to be split in two sets, one without missing records and the other one with missing records. The data set without missing records will be used in the learning phase of the machine learning algorithms. The data sets with missing records will be used in the testing phase. Important step when dealing with imputation is the feature selection, which consists in analyzing the correlation among variables to find out the ones that have higher correction with the variables that are going to be imputed. The selected variables ate this step will be used during the clustering process. Similar records are grouped into same cluster and the cluster information is used to categorical variables. Predicting the missing value of a categorical variable is seen as classification process. When it comes to numerical variable imputation, it has to be taken into account the missing ratio in the data set. When this ratio is high based on our experiment results, imputing the missing values with cluster mean when the missing records belongs to, the imputed value is closer to the original value. In case that the missing ratio is low, imputing with group mean method we achieve better results. More specifically, the clustering approach we apply shows better results in terms of RMSE and KS index. In terms of RMSE, our clustering method exhibits better the missing value prediction for categorical variables, where the RMSE is 0.201. On the other hand, imputing the numerical variables when the amount of missing data is considerable, using clustering approach we can achieve a KS index 0.013, a value closer to 0 is achieved when the distributions of the variable are almost equal before and after imputation.

6 Conclusions

In this paper, the performance of some statistical and machine learning imputation methods has been evaluated through an experimental application on a real-life census data set. Starting from a complete data set we artificially created missing data sets which gives us the possibility to evaluate the performance of the imputation process by comparing the original value with the estimated one. The analyses of results suggested that all methods produce good results in terms of preservation of the marginal distribution for the numerical target variable, but the best result is achieved when we apply the clustering technique in the data set where the missing ratio is higher. As expected, we found that simple imputation methods such as group mean imputation performed just as well as or more than the machine learning methods in data sets where the missing amount of records is low. We found similar performance with the use of decision trees and Bayesian networks for imputing categorical variables. K means is the method which performs better even in predicting the missing values of categorical and numerical variables.

In general, results confirm that when the data sets have a higher missing ratio, the machine learning algorithms perform better compared to the statistical methods such as group mean or Hot-Deck. The results confirm as well that these methods produce satisfactory results related to univariate characteristics of data and preservation of marginal distribution. So we can conclude that machine learning algorithms provide better

performance in reproducing the original value, producing significantly better estimates for statistical parameter such as mean, mode, standard deviation. Based on the proposed approach, in surveys for a data set containing a large amount of missing data, clustering methods should be the first methods to be considered. We discover that the basic K-means algorithm outperforms the mean substitution method or other statistical methods, which are considered as simple and common approaches for missing data imputation. As the imputation process is much related to the data set characteristics, missing patterns and mechanism, different methods should be considered for finding the method that better fits the needs. As this process is very crucial, after it there is a high need to always check the result of the imputation.

Further studies are needed in order to measure the performance of these methods when more covariates containing missing data are used, when the number of variables containing missing data is higher, and where the missing mechanism is not missing completely at random. The experiments should also be performed in other domain data sets such as health or financial ones.

References

1. Mannila, H.: Data mining: machine learning, statistics, and databases. In: 8th International Conference on Scientific and Statistical Database Management (SSDBM 1996), p. 2 (1996)
2. Groves, R.M.: Survey Errors and Survey Costs. Wiley, New York (1989)
3. <http://ec.europa.eu/eurostat/documents/64157/4372717/Eurostat-Quality-Assurance-Framework-June-2013-ver-1-1-EN.pdf/352234ca-77a0-47ca-93c7-d313d760bbd6>
4. Allan, F.G., Wishart, J.: A method of estimating the yield of missing plot in field experiments. *J. AgricSci.* **20**, 399–406 (1930)
5. Barlett, M.S.: Some examples of statistical methods of research in agriculture and applied biology. *J. R. Stat. Soc. B* **4**, 137–185 (1973)
6. Berglund, P.A.: An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2. SAS Global Forum 2010-Statistics and Data Analysis (2010)
7. Rubin, D.B.: Inference and missing data. *Biometrika* **63**, 581–592 (1976)
8. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (1987)
9. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley, New York (2002)
10. Schneider, T.: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* **14**, 853–871 (2001). American Meteorological Society
11. Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, P.: Bayesian networks for imputation. *J. R. Stat. Soc. Ser. A* **167**(2), 309–322 (2004a)
12. Purwar, A., Singh, S.K.: Empirical evaluation of algorithms to impute missing values for financial dataset. IEEE (2014)
13. <http://www.instat.gov.al/en/census/census-2011.aspx>
14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143 (1995)
15. Li, Y., Sun, G., Zhu, Y.: Data imbalance problem in text classification. In: 2010 Third International Symposium on Information Processing (ISIP). IEEE (2010)

16. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. arXiv preprint [arXiv:1106.1813](https://arxiv.org/abs/1106.1813) (2002)
17. Visa, S.: Fuzzy classifiers for imbalanced data sets. Department of Electrical and Computer Engineering and Computer Science, University of Cincinnati, Cincinnati (2006)
18. <https://www.istat.it/en/tools/methods-and-it-tools/processing-tools/concordjava>
19. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2001)
20. Liu, Y., Salvendy, G.: Visualization support to better comprehend and improve decision tree classification modeling process: a survey and appraisal. *Theor. Issues Ergon. Sci.* **8**(1), 63–92 (2007)
21. Fujikawa, Y., Ho, T.B.: Cluster-based algorithms for dealing with missing values. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, pp. 549–554. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47887-6_54
22. Westin, L.: Missing data and the preprocessing perceptron. Department of Computing Science, Umeå University (2002)