ENHANCING ENVIRONMENTAL AND HEALTH STATISTICS THROUGH ARTIFICIAL INTELLIGENCE: A COMPARATIVE STUDY OF IMPUTATION TECHNIQUES[1]

Simona Cafieri, Francesco Pugliese, Mauro Sodani

Abstract. In an increasingly globalized world, addressing health, environmental sustainability and social inequalities is crucial and requires an integrated approach involving national statistical offices. The latter is increasingly called upon to develop statistical frameworks to facilitate informed policy-making. However, incomplete or missing data in questionnaires or registers may compromise the accuracy and reliability of results.

The main objective of this study is to assess the effectiveness of different imputation methods using machine learning (ML) and artificial intelligence (AI) techniques in dealing with missing data in social surveys. To this end, a comparative analysis of different imputation techniques has been carried out, based on real datasets from the Istat Multi-purpose Household Survey, where missing data are common. Preliminary results suggest that ML/AI-based imputation methods outperform traditional statistical techniques in terms of performance and robustness.

The aim is to improve imputation techniques in official statistics to improve data quality on critical issues.

1 Introduction

The rise of artificial intelligence (AI) is having a significant impact on official statistics. AI methods provide solutions to data incompleteness and support informed decision-making (Sun et al., 2023). National Statistical Institutes have developed frameworks to support policy decisions, particularly in relation to environmental sustainability, health and social inequalities (Rigo, 2022).

In Italy, these issues are of particular importance as they form the basis of the BES (balanced and sustainable well-being) indicators that underpin the government's economic and financial planning document (Istat, 2024). It is worth noting that a significant proportion of these indicators are based on survey data, which are prone to inaccuracy and unreliability if incomplete or missing responses are not addressed. The application of AI is a promising solution to the problem of missing data in surveys. Such methods can be used to predict and impute missing values, thereby improving the overall quality of statistical datasets. Traditional techniques such as

[1] This work is the result of a close collaboration among the authors.

mean or median imputation often introduce bias, whereas AI-based methods can provide more accurate and impartial estimates. Several machine learning and deep learning models can be employed to impute missing data in official statistics. A substantial body of research has already been conducted to apply these methods, with positive results (De Fausti et al., 2023). This paper explores how AI can improve the quality of official statistics, focusing on the potential of machine learning (ML) and deep learning (DL) to improve the accuracy and reliability of health and environmental data.

2 Related Works

According to the literature, many deep learning and machine learning algorithms have been considered for 'data imputation'. Support vector machines (SVMs), a supervised learning method that identifies optimal hyperplanes for class separation in high-dimensional spaces, have been used to deal with missing values (Honghai et al., 2005). In addition, decision trees, which divide data into branches for decision-making, are evaluated for interpretability and used for data imputation (Nikfalazar, 2020).

XGBoost (Extreme Gradient Boosting) improves model performance by combining predictions from multiple estimators and is known for its efficiency (Mitchell, 2017; Rusdah & Murfi, 2020). The k-Nearest Neighbours (KNN) algorithm classifies data points based on their nearest neighbours and is popular for missing value imputation (Guo et al., 2003; Pujianto et al., 2019).

Linear regression models relationships between variables and is often used for predictive analysis (Montgomery et al., 2021). Random forest (RF) constructs multiple decision trees to improve model accuracy (Breiman, 2001) and has been used to impute missing values (Tang & Ishwaran, 2017). Long short-term memory (LSTM) networks capture long-term dependencies in sequential data and have been used to impute missing values (Hochreiter & Schmidhuber, 1997; Yuan et al., 2018). Gated recurrent units (GRUs) provide a simplified architecture for a similar task (Dey & Salem, 2017; Wang et al., 2022).

3 Methods

One of the main sources of social and household health data in Italy is the Aspects of Daily Life (AVQ) survey, carried out annually by Istat (Istat, 2022). AVQ represents an integral component of a unified system of social surveys. Indeed, collecting data is indispensable for understanding the daily lives of individuals and households. The survey provides information on the habits of citizens and the problems they face in everyday life through interviews with a sample of 20,000 households, representing approximately 50,000 individuals. Since 2018, the survey

has been carried out using a sequential CAWI/CAPI mixed-mode technique. The survey investigates a range of social aspects, including education, employment, family and social life, leisure time, political and social participation, health, lifestyles, access to services and other factors relevant to the study of quality of life. These topics are investigated from a social perspective, with particular consideration given to behaviors, motivations and opinions as key elements in the definition of social information. The survey is included in the National Statistics Plan, which collates the statistical investigations that are required for the Country. However, it is not uncommon for questionnaires to be incomplete, which can affect the precision and dependability of the resulting data.

To address this issue, we designed and implemented an imputation experiment by using the AVQ dataset from the 2021 survey comprising 735 variables. The presence of missing values in the dataset, frequently represented as blanks or NaN, is incompatible with scikit-learn estimators, which require all values to be numeric and significant. A fundamental approach is the complete case analysis, whereby rows (dropNA) or columns with missing values (list-wise deletion) are excluded. Nevertheless, this may result in a significant reduction of the available information. An effective strategy is to impute missing data by inferring it from the available data. Conversely, techniques such as the use of central tendency measures (mean, median, etc.) can be employed. This approach appears to be relatively straightforward and robust. However, there is a risk of underestimating or overestimating the true values, which could introduce bias into the resulting estimates. This phenomenon occurs when an algorithm produces results that are systematically biased due to incorrect assumptions, which are typically present in the data set or in the machine learning process.

In this study, we used missing data imputation techniques known as 'regression imputation'. Essentially, this method estimates missing values using a regressor (e.g. support vector regressor or random forest regressor), with the missing variable as the target and the other variables as inputs. Regression imputation is divided into 'deterministic' and 'stochastic'. The main difference between these two approaches is how the missing values are estimated and how uncertainty is taken into account. In deterministic regression imputation, missing values are estimated using a deterministic relationship between the variables. The trained regression model is used to predict the missing values for incomplete observations. The predicted value is used directly as an estimate for the missing value, hence the term deterministic. Deterministic imputation does not take into account the uncertainty associated with the estimate, so the predicted values are always the same for a given combination of input values. In contrast, stochastic regression imputation incorporates the uncertainty of the estimate into the imputation process. A stochastic noise term is added to the prediction. This noise term can be generated using the distribution of

residual errors from the regression model. For example, if the regression model has a residual variance of sigma squared, noise can be added extracting it from a normal distribution with a mean of 0 and a variance of sigma squared, resulting in a more realistic estimate. However, in our work we only used deterministic regression imputation, with the intention of exploring stochastic imputation adapted to machine learning and deep learning models in the future. As a final observation, we cannot directly impute regression values before preprocessing. The input predictor variables also contain missing data themselves, which would cause issues for the machine learning and deep learning models, as the libraries we used (i.e., scikit-learn or Keras) do not accept null values. Therefore, we imputed all input variables with missing values using a method called "Simple Random Imputation," which involves replacing the missing value with a random value. It is proven that this approach does not significantly affect the final estimate given the large number of variables present in our dataset. It is crucial to emphasize that the dataset for the subsequent machine learning models must adhere to specific rules and requirements to ensure optimal final imputation. Missing data should be random or at least analyzable; for example, it should be Missing Completely at Random (MCAR) or Missing at Random (MAR). In the event that the data is missing not at random (MNAR), the process of imputation becomes more complex. Regarding the percentage of missing data, datasets exhibiting a markedly high percentage of missing values (>50%) may potentially compromise the accuracy of the imputation models. It is optimal for the percentage of missing data to fall within the range of 12% to 50%. In the case of datasets with a missing data percentage lower than 12%, it would be preferable to utilize simpler imputation methods, such as setting the missing values to the mean of the variable or other constant values. Moreover, a significant proportion of the variables must remain complete, as these data points are indispensable for reconstructing the relationships between variables and predicting the missing ones. In particular, when employing deep learning models, it is crucial to ensure that the dataset is sufficiently large to enable effective training of the model, as these models typically require a significant number of samples to generalise effectively. Conversely, for a traditional machine learning model, the inclusion of a greater number of useful variables facilitates a more accurate prediction of the missing values. Additionally, the dataset must be structured in a way that meets the statistical production requirements. For example, existing values must be consistent with domain rules. In the case of a variable representing age, for instance, negative values or values exceeding a realistic threshold should be avoided. Imputation models must be capable of accommodating both qualitative variables, which should be accurately encoded using one-hot or ordinal encoding, and quantitative variables, which should be standardized or normalized in accordance with the models employed. The imputation process may be distorted by extreme outliers, particularly in the case of

traditional models such as KNN or regression. Outliers should be identified and managed, either through transformations or exclusion. Variables must have a stable and meaningful distribution, as this enables the models to accurately capture the patterns. Variables with highly imbalanced distributions may reduce the effectiveness of the imputation. Finally, for models such as linear regression or logistic regression, it is essential to reduce multicollinearity (strong correlation between independent variables), as this could negatively affect the imputation. This can be addressed using techniques such as the Variance Inflation Factor (VIF). In the literature, this method is more efficient than other methods of replacing predictor variables with a zero, the mean, etc. (Kalton & Kish, 1984).

In this work, we trained all the traditional machine learning models described in the previous section: SVM, DT, RF, XGBoost, KNN, and the most recent deep learning models: MLP, LSTM, GRU, CONV1D. The objective of the training was to create models for the imputation of the following health-related variables: The variables of interest were body mass index (BMI) for individuals aged 18 and over. The same models have been trained for the imputation of additional environmental variables, namely SODPOAP (resident satisfaction with household waste collection services). The ML and DL models were evaluated using a range of metrics that are appropriate for regression problems. The Root Mean Squared Error (RMSE), the Mean Absolute Percentage Error (MAPE), the Mean Absolute Error (MAE) and the R2 Score were employed for the assessment of the models. We observed that even though the RMSE yields absolute values, it was an adequate metric for comparing the performance of the different models and for their ranking. Accordingly, the trained models were ordered in descending order of RMSE, with the most effective model identified as the one with the lowest error. Furthermore, models that demonstrated minimal overfitting, as evidenced by a minimal discrepancy between the RMSE values for the training and test sets, were deemed the most optimal. Prior to the commencement of the training phase, a train-test split was conducted, with the test set comprising 20% of the entire dataset. No pre-processing was applied to the traditional machine learning models, given that the data is entirely numeric and lacks qualitative variables. For the deep learning models based on neural networks, all numerical inputs were normalized between 0 and 1. The training phase used Python's Hyperopt framework to optimize hyperparameters for scikit-learn (ML) and Keras (DL) models. Given the strong impact of initial hyperparameters, optimization was integrated directly into each model's training, adding computation time but ensuring optimal model selection, as Hyperopt returns the best model similarly to RandomizedSearch and GridSearch. Hyperopt is more efficient than these methods, using the Tree-structured Parzen Estimator (TPE) and Bayesian inference. TPE classifies parameter sets as "good" or "bad," focusing sampling on "good" regions. Iteratively, Hyperopt refines probability estimates, narrowing the search and

optimizing models faster. All results here were obtained using Hyperopt. The K-nearest neighbour (KNN) method exhibited the shortest training time, at 0.03 seconds, while the Gated Recurrent Units (GRUs) method demonstrated the longest training time, at 143 seconds. The remaining methods exhibited an average runtime of approximately 20 seconds, with the exception of Decision Trees and Linear Regression, which completed their execution in 0.4 seconds. Furthermore, XGBoost is optimized in terms of speed in comparison to traditional Gradient Boosting, which is known for its slow processing time and was therefore not included in this analysis. Regarding the deep learning models (MLP, Conv1D, LSTM, and GRU), the hyperparameters employed included a batch size of 32 (to ensure a sufficient degree of parallelism without unduly limiting the potential for parallelism or causing performance issues if the batch size were too high), a learning rate of 0.1, and 50 training epochs for all models. The deep learning models required an average training time of 100 seconds, except the MLP, which required only 21 seconds.

4 Results

Upon completion of the training phase, it was observed that the combination of the most effective models varied depending on the variable being imputed. This finding may be related to the No-Free-Lunch Theorem (ADAM, et al. 2019), which sets a theoretical limit in machine learning. The No-Free-Lunch Theorem postulates that no single optimal machine learning model exists for every task. Consequently, the strength of our method lies in its capacity to identify the optimal model for each variable (for each variable, the task is completely distinct) to be imputed, which is meticulously selected from a vast array of models. About the health-related variable "BMI", the results for all the models are presented in Table 1.

Table 1 – *Table of metrics of the models trained to predict the variable "BMI".*

MODEL	Training RMSE	Test RMSE
LSTM	0.7489	0.7546
GRU	0.7423	0.7551
CONV1D	0.7637	0.7854
MLP	0.7866	0.8122
SVM	0.8524	0.8823
KNN	0.6954	0.7949
LR	0.6400	0.7936
XG Boost	0.3898	0.7318
RF	0.2666	0.7240
DT	0.0000	1.0143

It can be seen that the Deep Learning model is the most effective. Long Short-Term Memory (LSTM) model. Table 2 presents a comparison of the descriptive

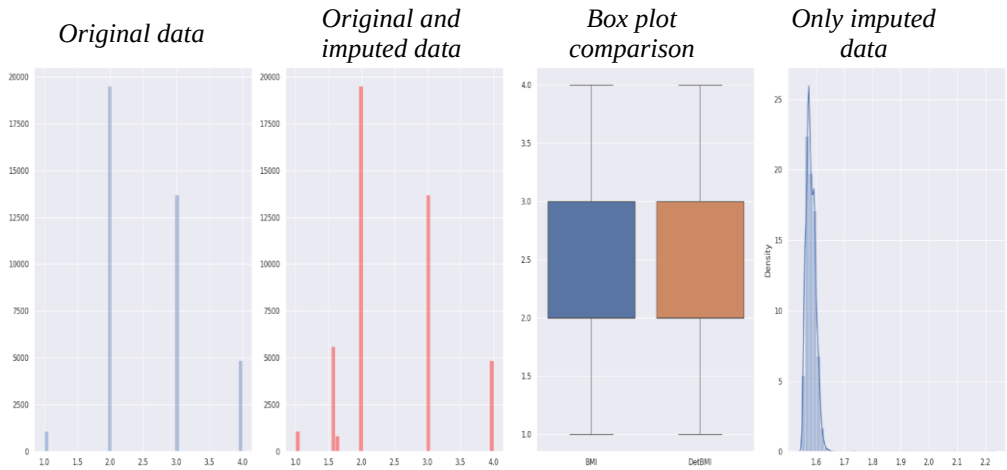
statistics (mean, standard deviation, and quartiles) calculated for the original variable prior to imputation and the imputed variable. This preliminary assessment indicates that the imputed distribution is not markedly disparate from the original distribution.

Table 2 – *Table of comparisons between descriptive statistics of BMI and Imputed BMI (DetBMI).*

MODEL	Mean	STD	MIN	25%	50%	75%	MAX
BMI	2.5695	0.7413	1	2	2	3	4
DetBMI	2.4296	0.7683	1	2	2	3	4

Figure 1 provides a comparison between the original distributions (omitting the nulls in the BMI column) and the distribution with imputed data (the BMI column without nulls plus imputed values).

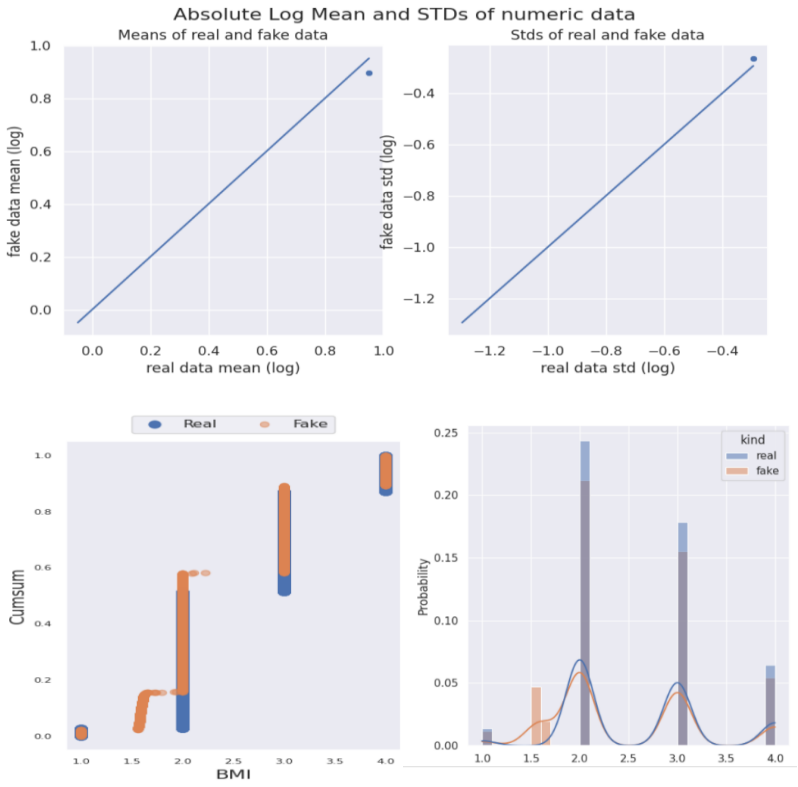
Figure 1 – *Comparisons among distributions charts for the variable BMI.*



Furthermore, an analysis of the box plots of both distributions and the distribution of only the imputed values that replace the nulls is provided. As can be observed, the two distributions are similar, although a slight margin of error is to be expected at this stage. Figure 2 presents a further comparison between the mean and standard deviation of the imputed data and the original data. Furthermore, a comparison between the original and imputed univariate and cumulative

distributions is presented. The distributions are highly similar, which reinforces the assertion that AI models can markedly enhance the handling of missing data.

Figure 2 – *Comparisons among means, standard deviations, cumulate and univariate distributions charts.*



With regard to the environmental variable SODPOAP, Table 3 provides a comparison of the metrics for all models, indicating that the Multi-Layer Perceptron (MLP) is the optimal model in this context. The most recent models exhibit a proclivity for overfitting, as evidenced by the markedly lower error rate on the training set in comparison to the test set. It is also noteworthy that, as anticipated, deep learning models demonstrate superior performance compared to machine learning models on these high-dimensional imputation datasets. Indeed, deep learning models consistently rank among the top performers, irrespective of whether they are recurrent or not. This suggests that the longitudinal (temporal) aspect of the data does not influence the models' performance in this particular dataset. In

comparison to traditional machine learning models, simpler models such as SVM and LR appear to demonstrate superior performance, whereas more sophisticated ensemble models like XGBoost and RF tend to exhibit a higher propensity for overfitting

Table 3 – *Table of metrics of the models trained to predict the variable "SODPOAP".*

MODEL	Training RMSE	Test RMSE
MLP	0.6843	0.6970
GRU	0.6048	0.6190
CONV1D	0.6169	0.6313
LSTM	0.6020	0.6165
SVM	0.6031	0.6189
KNN	0.5347	0.6115
LR	0.5093	0.6195
XGBoost	0.2915	0.5315
RF	0.1964	0.5337
DT	0.0000	0.7211

In Table 4, we present comparisons between descriptive statistics.

Table 4 – *Table of comparisons between descriptive statistics of SODPOAP and Imputed SODPOAP (DetSODPOAP).*

MODEL	Mean	STD	MIN	25%	50%	75%	MAX
SODPOAP	1.8563	0.6260	1	1	2	2	4
DetSODPOAP	1.7666	0.5532	1	1.5	2	2	4

Figure 3 presents a comparison of the original and imputed distributions of SODPOAP, while Figure 4 provides a comparison of the means, standard deviations, univariate, and cumulative distributions. As with the previous results, excellent results are obtained. However, it can be observed that the outcome and behavior of the models change depending on the difficulty level of the variable. The principal advantage of our methodology is the construction of a bespoke model for each variable to be imputed. The results demonstrate that deep learning models, such as long short-term memory (LSTM) and gated recurrent unit (GRU) networks, exhibit high performance in terms of root mean square error (RMSE) on every task, indicating their suitability for handling sequential data. Random Forest and XGBoost also demonstrated satisfactory performance, however, they tended to over-fit, rendering them unsuitable for the task of the imputation of missing values. The support vector machine models demonstrated reliable performance in imputation,

although they exhibited slightly higher root mean square error (RMSE) compared to the top-performing models.

Figure 3 – *Comparisons among distributions charts for the variable SODPOAP.*

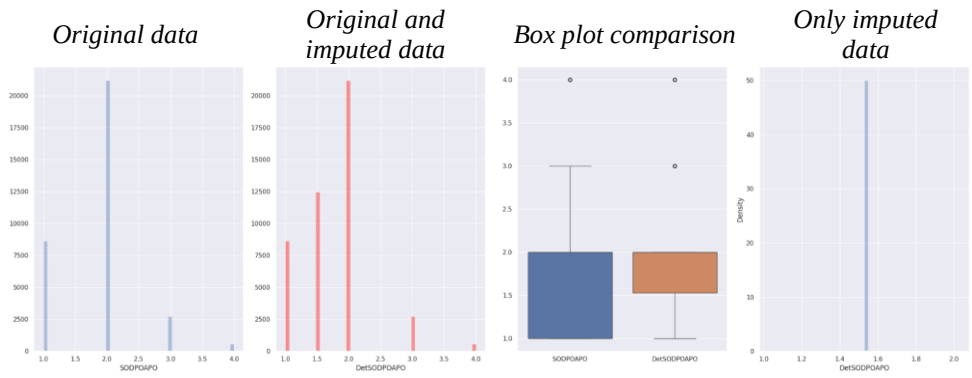
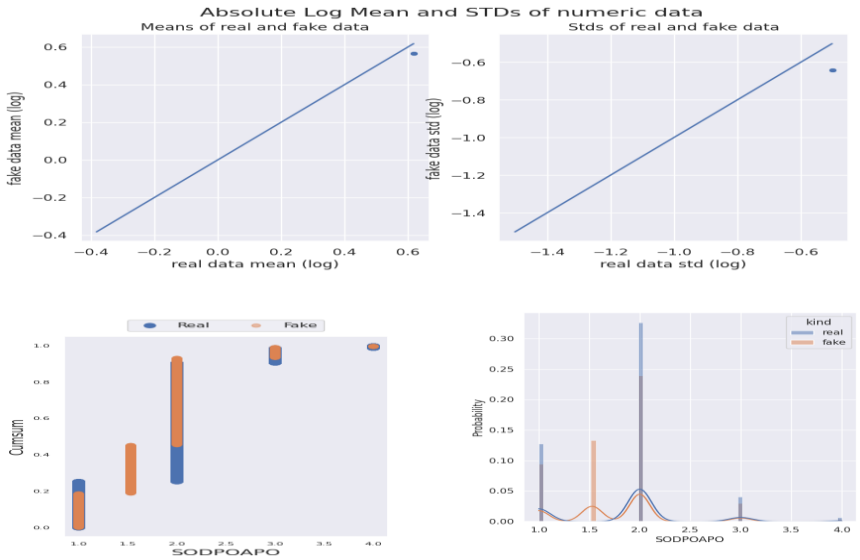


Figure 4 – *Comparisons among means, standard deviations, cumulate and univariate distributions charts.*



5 Conclusion

The examination of health and environmental statistics offers a promising avenue for improving the quality and reliability of data. It should be noted that this research was conducted on a single dataset; however, it could have been performed on multiple datasets. However, we encourage you to consider a few points. The dataset is not a simple one; rather, it is essential for Istat's production needs. The objective was to concentrate on this particular dataset, comprising over 700 variables, to ascertain the limits of AI capabilities in the context of this specific imputation problem. Furthermore, the No-Free-Lunch theorem indicates that no optimal model or AI exists that can perform equally well on all tasks and datasets. All such applications are domain-specific or task-specific, and thus dataset-oriented. Consequently, even if the AI had been tested on all imputation datasets, it would not be possible to guarantee that it would always work. It can be stated with certainty that AI-based imputation methods can effectively address the issue of missing data, thereby enhancing the overall integrity of statistical surveys. However, it is essential to note that the most suitable model must be carefully designed and implemented for each imputation problem, as there is currently no general AI (AGI) that can be applied to all types of problems and datasets. A comprehensive analysis allows for the selection of the optimal model, the imputation of missing data using this model, and the subsequent evaluation of the quality of the imputed data. Future research should prioritize the development of a stochastic regression imputation method that more effectively preserves the variance between the original and imputed distributions, as well as the investigation of advanced models, such as Transformers and Generative Adversarial Networks (GANs), to further enhance the imputation process.

References

ADAM S P., ALEXANDROPOULOS S. A. N., PARDALOS P. M., VRAHATIS M. N. 2019 No free lunch theorem: A review. *Approximation and optimization: Algorithms, complexity and applications*, pp. 57-82.

BREIMAN L. 2001. Random forests. *Machine learning*, Vol. 45, pp. 5-32.

DE FAUSTI, F., DI ZIO M., FILIPPINI R., TOTI S., ZARDETTO, D. 2023. A study of MLP for the imputation of the “Attained Level of Education” in Base Register of Individuals. In: *WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS*. p. 69.

DEY R., SALEM F. M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* (pp. 1597-1600), IEEE.

GUO G., WANG H., BELL D., B, Y., GREER K. 2003. KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986-996). Springer Berlin Heidelberg.

HOCHREITER S., SCHMIDHUBER J. 1997. Long short-term memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780.

HONGHAI, F, GUOSHUN, C., CHENG, Y., BINGRU, Y., & YUMEI, C. 2005. A SVM regression based approach to filling in missing values. In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer Berlin Heidelberg, 2005. p. 581-587.

ISTAT. 2022. *Indagine Aspetti della vita quotidiana 2021*.

ISTAT 2024. *Rapporto BES 2023*.

KALTON G., KISH L. 1984. Some efficient random imputation methods, *Communications in Statistics-Theory and Methods*, Vol. 13, No.16, pp. 1919-1939.

MITCHELL R., FRANK E. 2017. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, Vol 3: e127.

MONTGOMERY D. C., PECK E. A., VINING G. G. 2021. *Introduction to linear regression analysis*. John Wiley & Sons.

NIKFALAZAR S., YEH C. H., BEDINGFIELD S., KHORSHIDI H. A. 2020. Missing data imputation using decision trees and fuzzy clustering with iterative learning, *Knowledge and Information Systems*, Vol. 62, pp. 2419-2437.

RIGO A. 2022. *Programmazione e innovazione: il percorso verso l'efficienza interna delle Pubbliche Amministrazioni*.

SUN Y, LI J., XU Y., ZHANG T., WANG X. 2023. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, Vol. 227: 120201.

TANG F., ISHWARAN H. 2017. *Random forest missing data algorithms*. Statistical Analysis and Data Mining: ASA Data Science Journal, Vol. 10, No. 6, pp. 363-377.

Simona CAFIERI, Istat, cafieri@istat.it
Francesco PUGLIESE, Istat, frpuglie@istat.it
Mauro SODANI, Istat, sodani@istat.it