UNECE

STATISTICS AUSTRIA
The Information Manager

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Expert Meeting on Statistical Data Editing
7 to 9 October 2024, Vienna, Austria

Enhancing Official Statistics through Artificial Intelligence

Mauro Bruno, Francesco Ortame, Francesco Pugliese,
Istat | Central Directorate for Methodology and Design of Statistical Processes

Simona Cafieri
Istat | Central Directorate for Communication, Information and Services to Citizens and Users

Istat | Istituto Nazionale di Statistica

Background

- ✓ National Statistical Institutes are increasingly called to develop statistical frameworks on different topics, to contribute to informed policy decision-making.
- ✓ In particular, in an era of increasing global networking, it is imperative to vigorously address emerging challenges affecting environmental sustainability, health and social inequalities.
- ✓ These issues are of great importance as they are key elements of equitable and sustainable well-being (Bes), which is the basis of the economic and financial planning document of the Italian government
- ✓ The main problem is that incomplete or missing data in questionnaires or registers can affect the accuracy and reliability of the results

Data sources

One of the main sources of data on the social situation of households in Italy is the Aspects of Daily Life survey(AVQ), which is carried out every year by ISTAT

In this survey missing data in questionnaires are very common and can affect the accuracy and reliability of the results

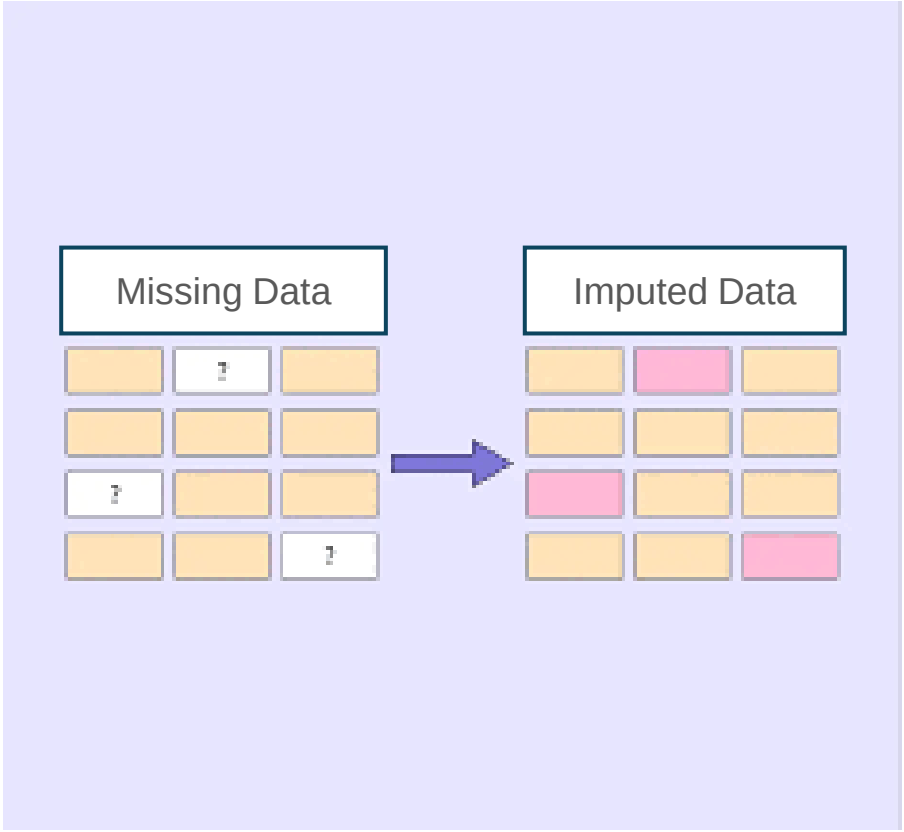
So we started from this survey to make an imputation experiment with the use of Artificial Intelligence-

We use a dataset related to 2021 survey, with 735 variables.

Regressive Imputation

- Datasets of AVQ survey, such as those in Official Statistics, contain missing values, often encoded as **blanks** or **NaNs**. These datasets are typically incompatible with *scikit-learn* estimators, which require all values to be numerical and significant.

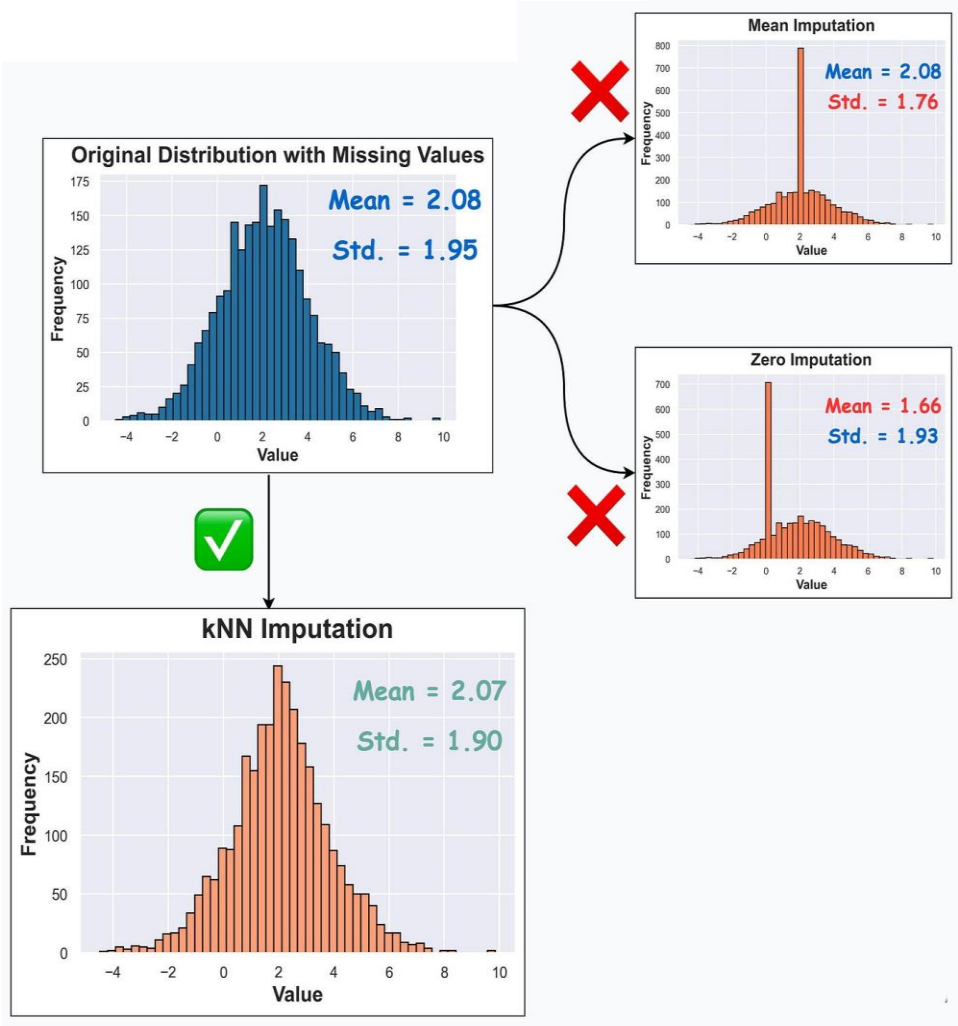
- A basic approach to handling missing data is Complete Case Analysis, where rows (**dropNA**) or columns with missing values (**List-wise deletion**) are discarded. However, this can lead to significant information loss. A more effective strategy is to impute the missing data by inferring it from the available data.



Regressive Imputation

- On the other hand, techniques such as using **Measures of Central Tendency** (mean, median, etc.) can be employed. This approach appears to be simple and safe. However, it is easy to under- or overestimate the actual values, thereby introducing bias into our estimates. For instance, consider a person who does not provide their salary data because they earn just enough to meet their daily needs. If we impute the mean salary for this individual, we overestimate their actual earnings, which introduces **bias** into our analysis.

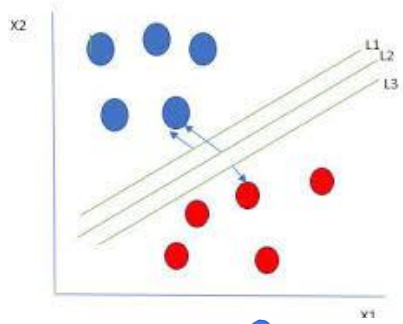
- It is important to remember that in **machine learning**, bias is a phenomenon that occurs when an algorithm produces systematically skewed results due to incorrect assumptions, usually present in the dataset or the machine learning process.



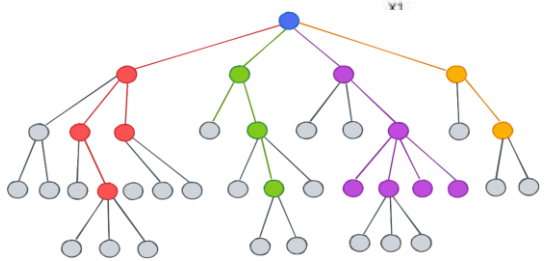
Machine Learning and Deep Learning Models

We conducted a comprehensive comparison of both Machine Learning and Deep Learning models.

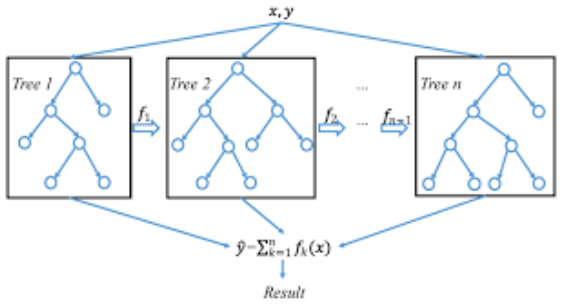
- **Support Vector Machines (SVM)**: A supervised learning algorithm that finds the optimal hyperplane to separate different classes in the feature space. It is effective in high-dimensional spaces and commonly used for classification and regression.



- **Decision Tree:** A non-linear model that splits data into branches based on feature values to make decisions. It is useful for its interpretability and is commonly applied in classification and regression tasks.



- **XGBoost (Extreme Gradient Boosting):** An advanced boosting algorithm that combines the predictions of several base estimators to improve model performance. It is known for its high efficiency and accuracy and is commonly used in machine learning competitions.



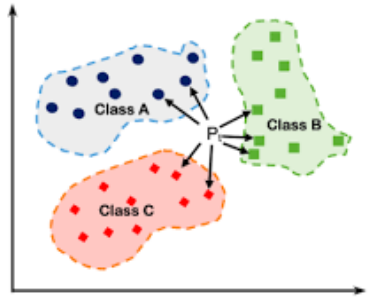
Istat

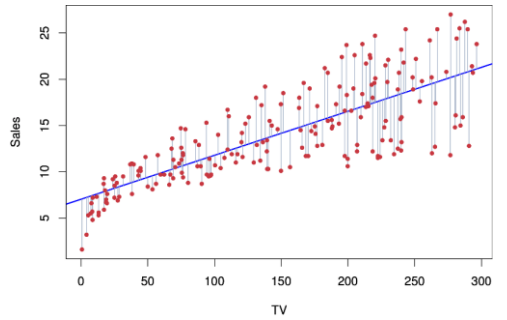
Machine Learning and Deep Learning Models

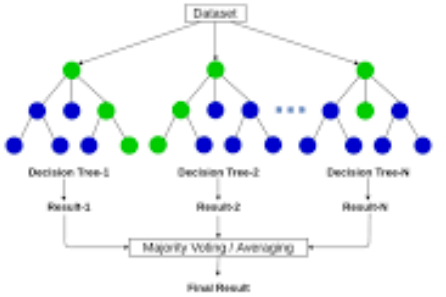
- **K-Nearest Neighbors (KNN):** A simple, instance-based learning algorithm that classifies data points based on the majority class among its nearest neighbors. It is suitable for classification tasks in pattern recognition and recommendation systems.

- **Linear Regression**: A linear approach to modeling the relationship between a dependent variable and one or more independent variables. It is commonly used for predictive analysis and trend forecasting.

- **Random Forest:** An ensemble learning method that builds multiple decision trees and merges their results to improve accuracy and robustness. It is used in applications such as feature selection and complex classification tasks.



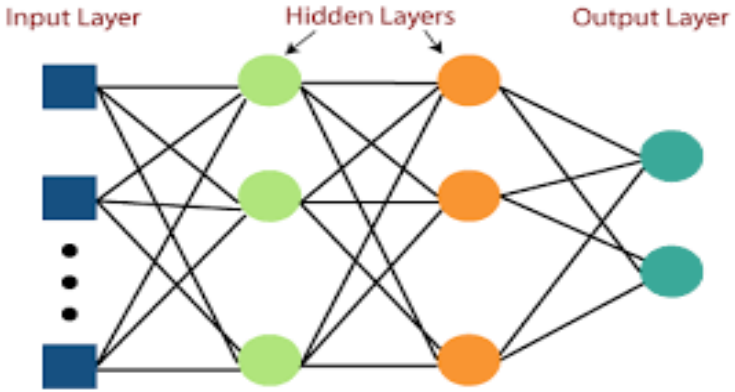


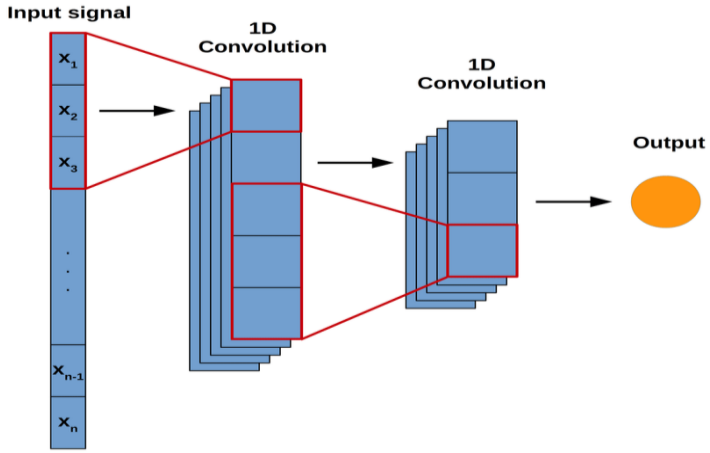


Machine Learning and Deep Learning Models

- **Multi-Layer Perceptron (MLP)**: A class of feedforward artificial neural networks with multiple layers of neurons, used for both classification and regression tasks. It is effective for various prediction problems where data is structured and requires non-linear modeling.

- **Convolutional 1D Neural Network**: A variant of Convolutional Neural Networks (CNNs) specialized for one-dimensional data, such as time series or sequences. It is used in applications like signal processing and text classification.

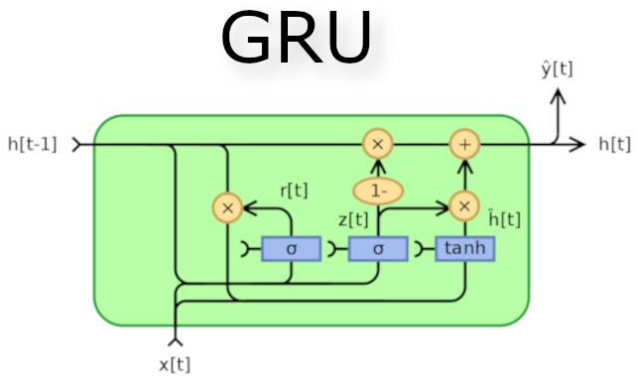


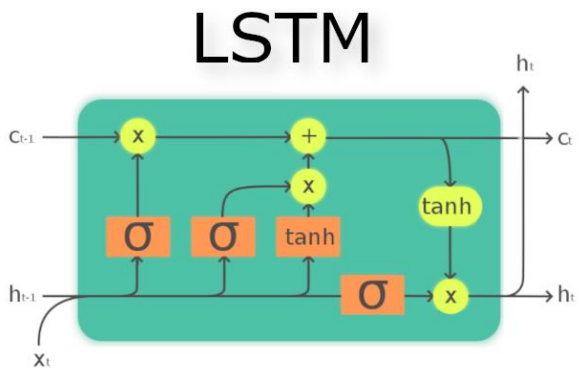


Machine Learning and Deep Learning Models

- **Long-Short Term Memory (LSTM)**: A type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data using memory cells. It is commonly used for tasks like speech recognition and time-series forecasting.

- **Gated Recurrent Unit (GRU)**: A type of RNN similar to LSTM but with a simplified architecture, making it faster and more efficient. It is used in similar applications as LSTM, such as natural language processing and sequential data analysis.





Istat

Results of Imputing BMI

- We rank the **best models** based on the **root mean squared error** (RMSE) in descending order, with the best model having the lowest error. Additionally, we consider the models with the least possible overfitting, indicated by a low difference between the training set and the test set RMSE. In this case, **LSTM** proves to be the best.

BMI-Body Mass Index

Model	Root Mean Squared Error on the Trainining Set	Root Mean Squared Error on the Test Set
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

Results of Imputing SODPOAP

- **Training** both **Machine Learning** and **Deep Learning** models reveals that the boundary conditions change for each column, and therefore the best model is always different. In this case, SVMs are the best.

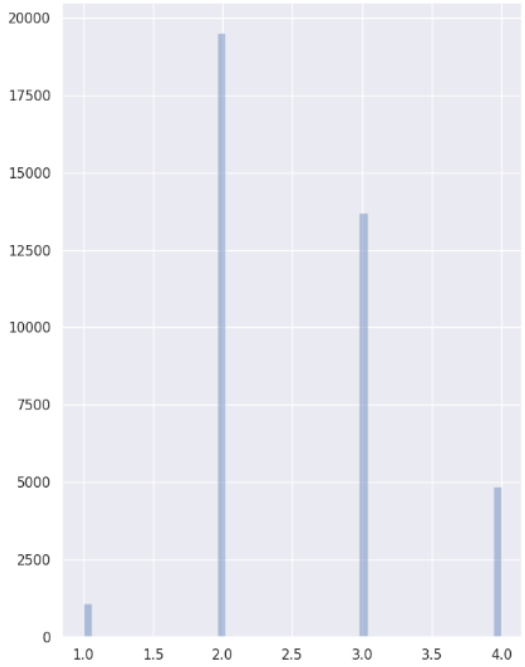
SODPOAP (Door to door waste collection)

Model	Root Mean Squared Error on the Trainining Set	Root Mean Squared Error on the Test Set
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

Body Mass Index– BMI Distribution Analysis

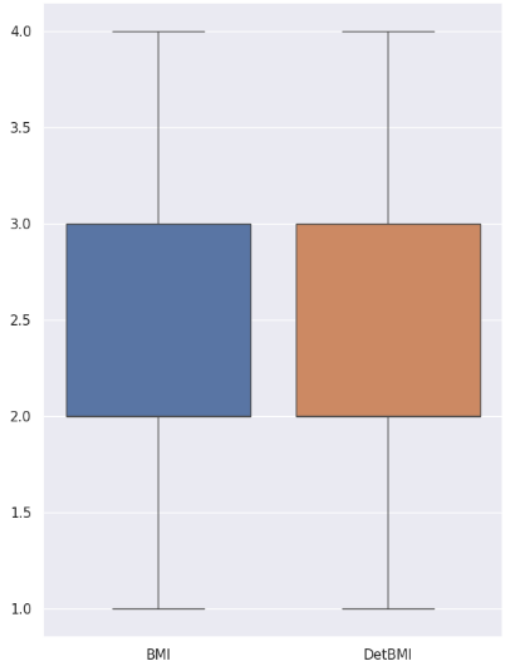
As we can see, the **imputed** distribution is similar to the original one (first two graphs), reinforcing the claim that AI models can significantly improve the handling of missing data.

Original Data

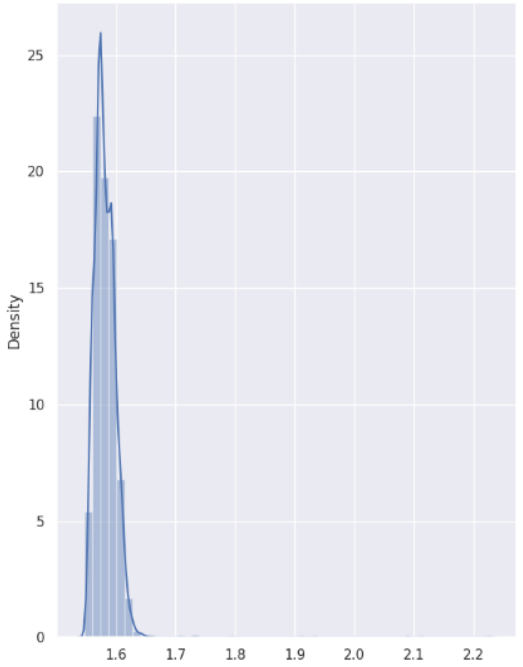


Original + Imputed Data

BoxPlots Comparisons



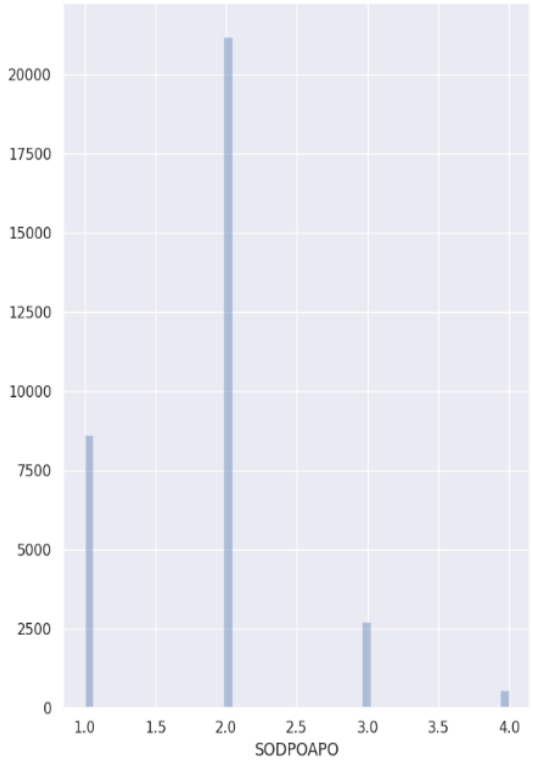
Only Imputed Data



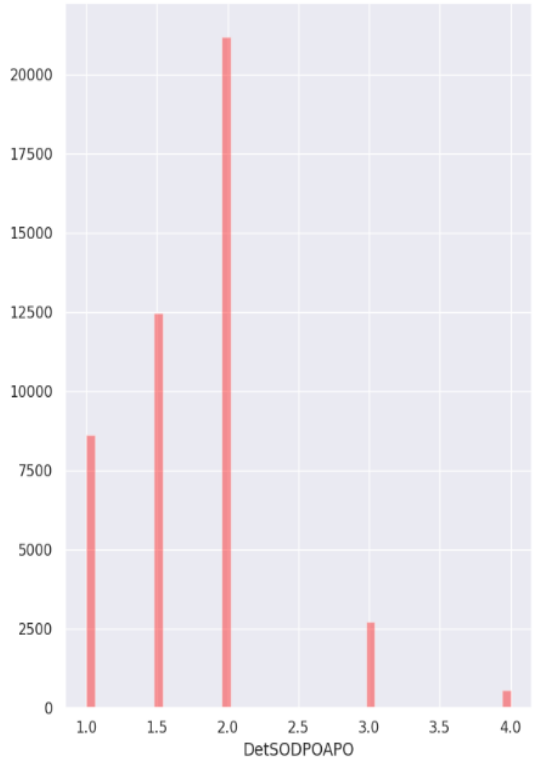
Door to door waste collection – SODPOAP Distribution Analysis

- With the variable SODPOAP, 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

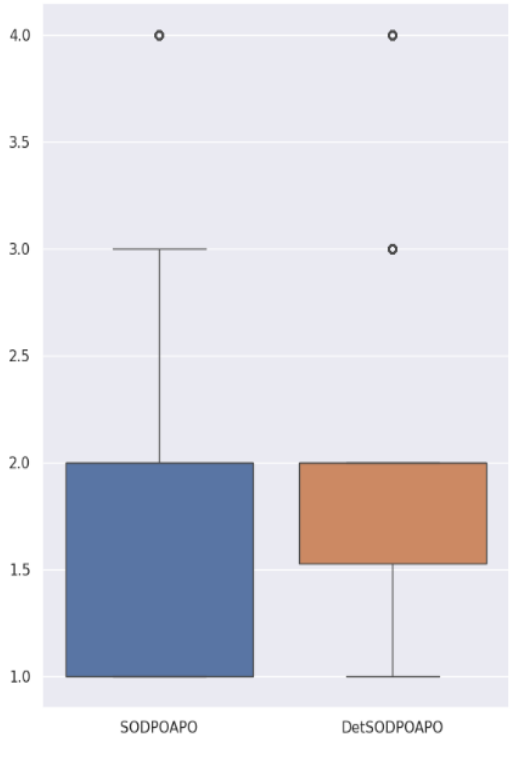
Original Data



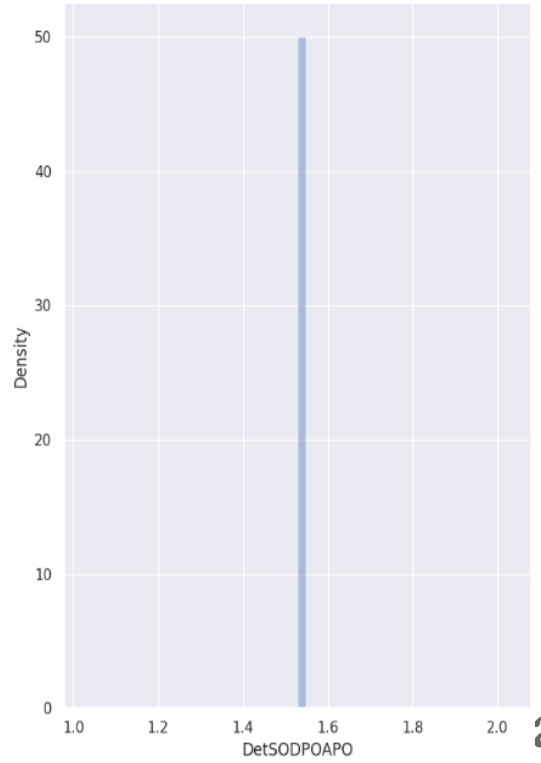
Original + Imputed Data



BoxPlots Comparisons



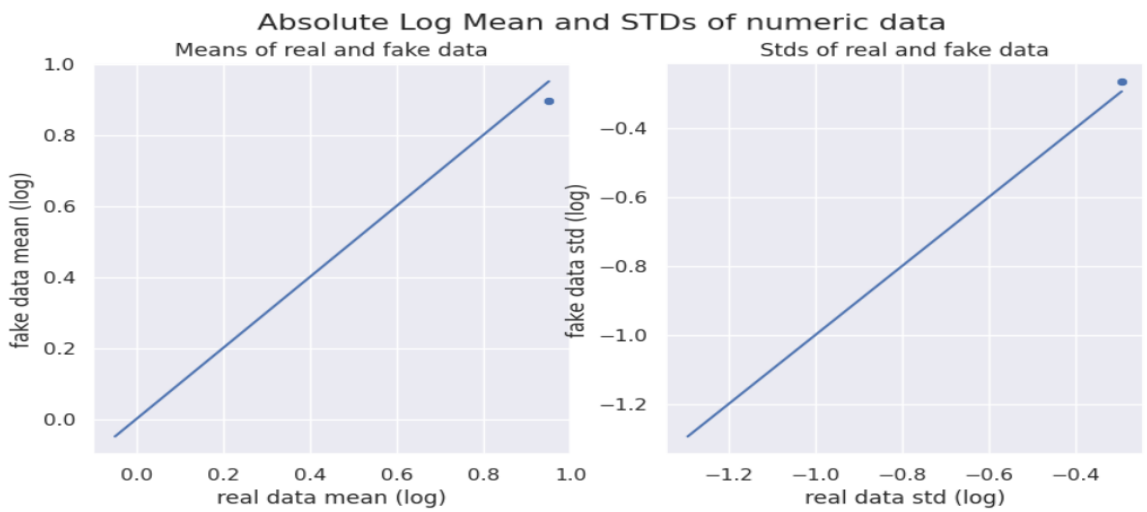
Only Imputed Data



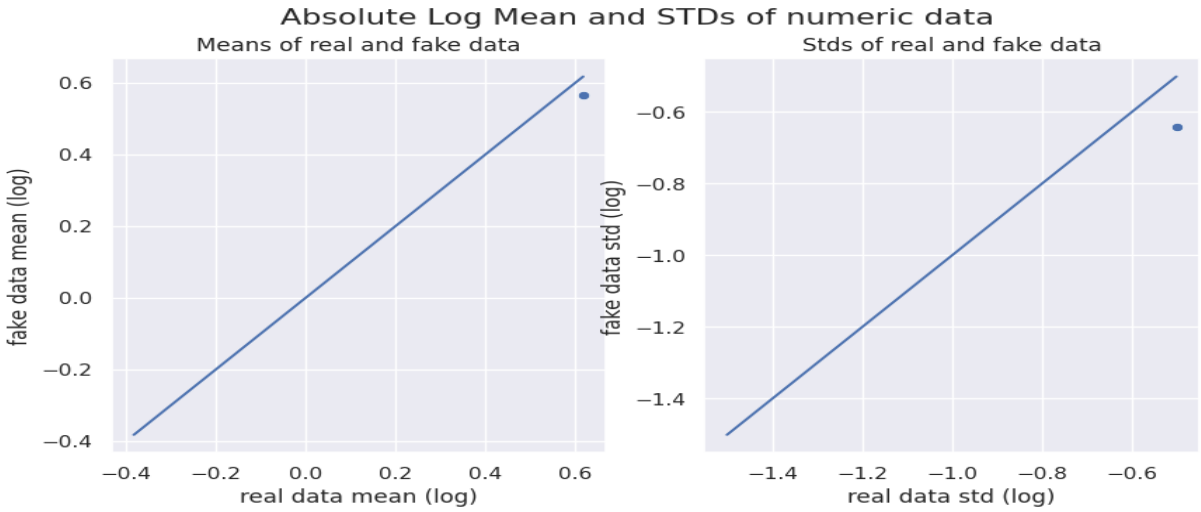
Results: Means and Standard Deviation Comparison

- The means for both BMI and SODPOAP correlate, while the variances are better preserved in the second variable.

BMI

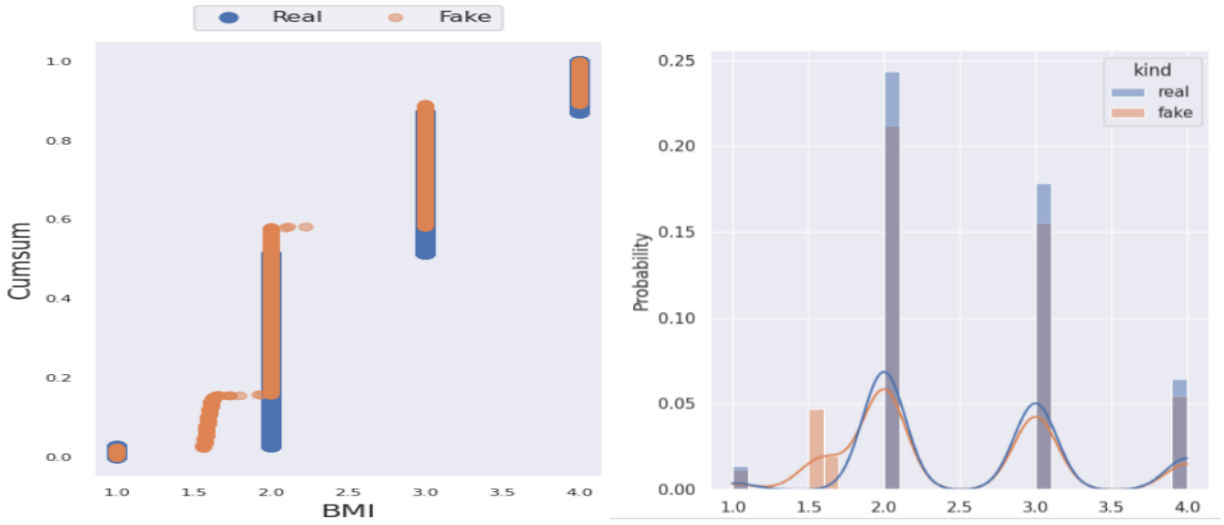


SODPOAP

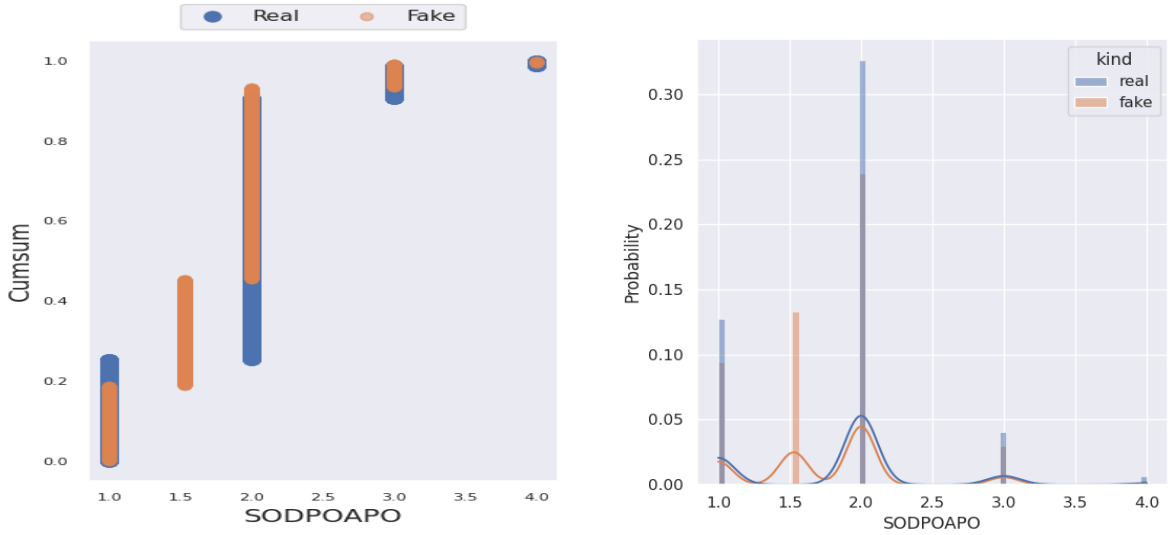


Results: Distribution Comparison

BMI



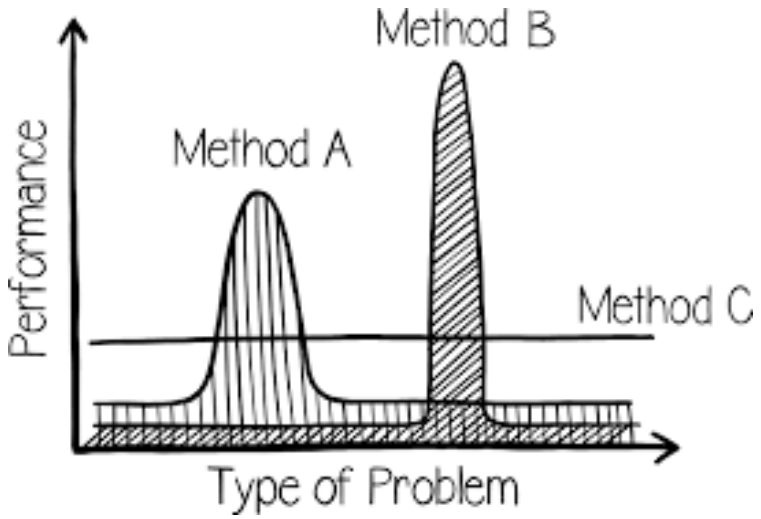
SOADPOP



Istat

Conclusions

- According to the **No-Free-Lunch Theorem**, there is no single best machine learning model for every column.
- The strength of this method lies in always finding the **best model for each variable** to be imputed, carefully selected from a wide range of models.
- A thorough analysis allows us to choose the best model, impute the missing data using it, and subsequently evaluate the **quality** of the imputed data.
- In the future, we will use other models such as **Transformers** and **GANs** and work on a stochastic regression imputation method that better preserves the variance between the original and imputed distributions.



Istat

Thank you!

mbruno@istat.it
cafieri@istat.it
francesco.ortame@istat.it
frpugliese@istat.it

Istat | Istituto Nazionale di Statistica