UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna Austria

12 July 2024

Enhancing Environmental and Health Statistics through Artificial Intelligence: A Comparative Study of Imputation Techniques

Simona Cafieri, Francesco Pugliese, and Francesco Ortame (Istat, Italy)
cafieri@istat.it

Abstract

In an era of increasing global networking, it is imperative to vigorously address emerging challenges affecting health, environmental sustainability and social inequalities. These closely intertwined issues require an integrated approach involving National Statistical Institutes. They are increasingly called to develop statistical frameworks on these topics to contribute to informed policy decision-making, but incomplete or missing data in questionnaires or registers can affect the accuracy and reliability of the results, The main objective of this work is to assess the effectiveness of different imputation methods using Machine Learning (ML) and Intelligence (AI) techniques in dealing with missing data in social surveys. To achieve this goal, a comparative analysis of different imputation techniques, including traditional statistical methods and cutting-edge deep learning algorithms, has been carried out. These techniques include Linear Regression (LR), k-Nearest Neighbour (KNN), Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB), Support Vector Machines (SVMs) and Deep Learning models such as Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Long-Short Term Memories (LSTMs), Generative Adversarial Networks (GANs) and the recent Transformers. All these methods are implemented as regressors since want to investigate the regressive imputation framework. The comparisons are based on real datasets from Istat multipurpose survey on households, where missing data are common. Preliminary results suggest that ML/AI-based imputation methods outperform traditional statistical techniques in terms of performance and robustness, especially when dealing with complex datasets and high-dimensional features. Therefore, this work aims to explore innovative AI solutions to contribute to the advancement of imputation techniques in official statistics to have more complete and more accurate data on health, environment, inequality, and other social aspects. That will be the basis for evidence-based decision-making for a more equitable and sustainable future.