International Conference of Engineering, Science and Education

APPROACHES TO FILLING MISSING VALUES IN DATASETS TO IMPROVE THE RESULTS OF ARTIFICIAL INTELLIGENCE MODELS

Isaqulov To'lqin Mahmud ugli[1], Shaun Billones[2,3], Taylakova Dilnoza Norbekovna[1], Bayjonov Furqat Baxramovich[1], Abdurazzaqov O'ktam Abduqayumovich[1], Shomurodov Jamshid Olimboy ugli[1], Pulatova Nargiza Negmatovna[3] and Mamadiyorov Jamol Baxodirovich[1]

[1] *Oriental University, Tashkent, Uzbekistan*

[2] *Polytechnic University of the Philippines, Manila, Philippines*

[3] *Samarkand State Pedagogical Institute, Samarkand, Uzbekistan*

tolqinisaqulov188@gmail.com

Keywords: Missing data, imputation, artificial intelligence, machine learning, deep learning, MICE, GAN.

Abstract: This article explores modern approaches to filling in missing values in datasets to improve the accuracy and stability of artificial intelligence models. First, the theoretical foundations of the MCAR, MAR, and MNAR mechanisms, as well as their forecast quality and impact, are analyzed. Using real data from healthcare and education, artificial missingness scenarios are generated at various proportions, i.e. 6–31% and structures, and several statistical, machine learning, and deep learning-based imputation methods are compared using mean/median, regression, MICE, kNN, random forest, autoencoder, GAN, GAIN, GAMIN. The imputation quality was assessed using RMSE, MAE, and correct recovery rate, while the performance of SI models was assessed using metrics such as Accuracy, ROC–AUC, F1, and R^2. The results show that there is no universal best method, and the choice of method depends on the missingness mechanism, proportion, data structure, computational resources, and fairness and interpretability requirements. This work proposes a conceptual methodology and practical recommendations for method selection in practical scenarios, supporting decision-making in managing missing data in artificial intelligence-based systems. The impact of different imputation strategies on group outcomes, fairness indicators, and model interpretability is analyzed, and theoretical recommendations are made, taking into account precautions for regulated sectors.

1 INTRODUCTION

Missing values are always present in real-life datasets, and this directly affects the results of artificial intelligence and machine learning models. Due to various reasons, interruptions and shortcomings in the data collection phase, whether from sensors, questionnaires, online platforms, or corporate information systems, some observations are not fully recorded. As a result, empty cells, incorrect entries, or completely missing columns or rows appear in the tabular database. If such flaws are ignored, the sample size during model training will decrease, the variance of estimates will increase, and the forecasts will become biased and unreliable. In addition, failure to properly account for the structure of missing values can lead to systematic errors and unfairness for some categories, which can lead to a decrease in trust in artificial intelligence systems. According to the statistical literature, missing data mechanisms are divided into two classes: random and non-random. It is observed that classical statistical methods provide relatively reliable results for cases where data are completely missing at random. However, in most practical situations, the probability of loss depends on other factors, whether observed or unobserved. This revealed that approaches such as simply filling in with an average value or deleting missing rows were not sufficient. Especially in emerging industries such as medicine, finance, education analytics, IoT, and cybersecurity, where decisions are often based on the results of SI models, data quality and completeness are of critical importance. Therefore, scientifically based analysis and management of missing data is becoming a separate research area today. In recent years, many