Development and Evaluation of Metrics for Assessing Synthetic Tabular Data Quality

Nora Amama Ben Hassun[1], Jordi Cortés Martínez[1], Daniel Fernández[1]

[1]Department of Statistics and Operations Research(DEIO). Universitat Politècnica de Catalunya · BarcelonaTech(UPC), Spain

The growing reluctance to share original datasets and the increasing demand to comply with privacy regulations have motivated the adoption of synthetic data. Synthetic data replicates the statistical properties of the original datasets while ensuring that individual-level information or sensitive variables are not disclosed. However, to effectively evaluate the quality of synthetic data, the development and refinement of validation metrics based is required. This assessment ensures the usability and reliability of synthetic datasets.

This research aims to introduce some existing validation metrics implemented in tools such as the `synthpop` package. The focus is on synthetic tabular data, with an emphasis on showcasing a comprehensive list of validation metrics that hold statistical significance and serve as a foundation for the development of new metrics. To address the challenges of validating synthetic data, the research highlights tailored methodologies for specific domains, such as energy, where there are unique challenges. Synthetic data offers opportunities to accelerate model training while ensuring compliance with privacy regulations. By developing robust metrics, the goal is to provide a practical framework for validating high-quality synthetic datasets that meet the needs of sensitive fields. All these metrics will be illustrated through a case study to highlight their applicability and relevance, ultimately filling a considerable gap in the literature concerning synthetic data validation in the energy sector.

Validation metrics are examined on three key dimensions: resemblance, utility, and privacy. Resemblance metrics evaluate the similarity in the statistical distributions between the synthetic and original datasets. Utility assesses the suitability of synthetic data for specific analytical tasks, such as machine learning or statistical modeling. Privacy, meanwhile, ensure that sensitive information from the original data cannot be reconstructed or identified.

Keywords: Synthetic Tabular Data, Validation Metrics, Statistics, Resemblance, Utility, Privacy