International Conference on Engineering, Trends in Engineering, Science, Humanities and Management (ICETESHM – 2024)
31st March, 2024, Karol Bagh, New Delhi, India.

CERTIFICATE NO: ICETESHM /2024/ C0324338

Responsible Artificial Intelligence Through Synthetic Tabular Data Generation: Addressing Bias, Privacy, and Data Imbalance

Harer Savita Laxman
Research Scholar, Department of Computer Science Engineering, Vikrant University, Gwalior, M.P., India.

ABSTRACT

Artificial Intelligence (AI) has become a transformative technology across domains such as healthcare, finance, education, cybersecurity, and public administration, where data-driven decision-making plays a crucial role. However, the effectiveness and trustworthiness of AI systems are often compromised by challenges associated with tabular datasets, including algorithmic bias, class imbalance, missing values, and stringent privacy requirements. These issues can lead to unfair predictions, reduced model generalization, discrimination against minority groups, and limitations in data sharing due to privacy regulations. Consequently, developing Responsible Artificial Intelligence (RAI) solutions that ensure fairness, transparency, accountability, and privacy has become a significant research priority. This study presents a comprehensive framework for Responsible Artificial Intelligence through Synthetic Tabular Data Generation to address bias, privacy, and data imbalance in machine learning models. The proposed framework employs advanced synthetic data generation techniques, including Conditional Tabular Generative Adversarial Networks (CTGAN), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and differential privacy mechanisms to generate realistic and privacy-preserving synthetic datasets. These datasets are used to train and evaluate machine learning models while maintaining statistical fidelity and minimizing disclosure risks. The framework further integrates fairness-aware preprocessing, synthetic oversampling, explainability methods, and Responsible AI evaluation metrics to improve predictive performance and ensure equitable outcomes across diverse demographic groups. Experimental validation is performed using benchmark tabular datasets and widely adopted machine learning algorithms, including Random Forest, XGBoost, LightGBM, Logistic Regression, Support Vector Machine, and deep neural networks. Performance is assessed using predictive, fairness, and privacy metrics. The findings demonstrate that synthetic tabular data significantly enhances minority class representation, mitigates algorithmic bias, preserves sensitive information, and maintains high predictive accuracy. The proposed framework offers a scalable and practical approach for developing trustworthy, privacy-preserving, and ethically responsible AI systems for real-world tabular data applications.

Keywords: *Responsible Artificial Intelligence (RAI), Synthetic Tabular Data, Synthetic Data Generation, Tabular Machine Learning, Machine Learning.*