ABSTRACT

Clean and safe water is essential for human health and recognized as a fundamental right. However, the degradation of water quality worldwide, due to pollution and inadequate water management, poses serious risks to public health, making it crucial to evaluate water potability effectively. This study aims to classify water potability more accurately by leveraging machine learning models. The research utilized the Water Potability Dataset, which contains 3,276 records and 10 parameters, such as pH, hardness, and solids, all of which were analyzed to ensure a comprehensive assessment of water quality. Missing values in the dataset were addressed through imputation techniques, while class imbalance was resolved with data balancing methods. The dataset was divided into 80% for training and 20% for testing, ensuring reliable model validation. Decision Tree and Random Forest models were applied due to their efficiency in managing complex datasets and providing interpretable outputs. The Random Forest model outperformed the Decision Tree, achieving higher accuracy and demonstrating superior reliability in classifying water as potable or non-potable. Evaluation results indicate that these models effectively classify water as potable or non-potable, showcasing the potential of machine learning to improve water quality analysis. The findings of this study provide a reliable, data-driven framework that can support policymakers, water management authorities, and public health organizations in addressing water quality challenges. By facilitating timely and accurate assessment, this research contributes to safeguarding public health and promoting sustainable water management practices. This study was also prepared as part of a final project and is currently in the process of being submitted for publication, to ensure the findings can contribute more broadly to both academic research and practical applications.

Keywords: Imbalanced Data, Missing Data Handling, Decision Tree Algorithm, Random Forest Model, Water Contamination Risk, Research