1. Comput Intell Neurosci. 2022 Sep 20;2022:9283293. doi: 10.1155/2022/9283293. eCollection 2022.

A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI.

Patel J(1), Amipara C(1), Ahanger TA(2), Ladhva K(1), Gupta RK(1), Alsaab HO(3)(4), Althobaiti YS(4)(5), Ratna R(6).

Author information:
(1)Department of Computer Science and Engineering Pandit Deendayal Energy University, Gandhinagar, Gujarat, India.
(2)College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Saudi Arabia.
(3)Department of Pharmaceutics and Pharmaceutical Technology, Taif University, Taif 21944, Saudi Arabia.
(4)Addiction and Neuroscience Research Unit, Taif University, Taif 21944, Saudi Arabia.
(5)Department of Pharmacology and Toxicology, College of Pharmacy, Taif University, Taif 21944, Saudi Arabia.
(6)Gedu College of Business Studies, Royal University of Bhutan, Bhutan.

During the last few decades, the quality of water has deteriorated significantly due to pollution and many other issues. As a consequence of this, there is a need for a model that can make accurate projections about water quality. This work shows the comparative analysis of different machine learning approaches like Support Vector Machine (SVM), Decision Tree (DT), Random Forest, Gradient Boost, and Ada Boost, used for the water quality classification. The model is trained on the Water Quality Index dataset available on Kaggle. Z-score is used to normalize the dataset before beginning the training process for the model. Because the given dataset is unbalanced, Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset. Experiments results depict that Random Forest and Gradient Boost give the highest accuracy of 81%. One of the major issues with the machine learning model is lack of transparency which makes it impossible to evaluate the results of the model. To address this issue, explainable AI (XAI) is used which assists us in determining which features are the most important. Within the context of this investigation, Local Interpretable Model-agnostic Explanations (LIME) is utilized to ascertain the significance of the features.

DOI: 10.1155/2022/9283293

PMCID: PMC9514946
PMID: 36177311 [Indexed for MEDLINE]

Conflict of interest statement: The authors declare that they have no conflicts of interest.