Contribution ID: **33** Type: **Talk**

Data mining, management and modelling in advancing water and sanitation systems

Wednesday, 3 July 2024 16:00 (20 minutes)

Data mining and management play an important role in advancing water and sanitation systems, ensuring the sustainable delivery of essential services. The application of data mining encompasses the collection, processing, and analysis of vast datasets derived from various sources such as sensor networks, satellite imagery, and public health records. These techniques facilitate the identification of patterns, trends, and anomalies, which are important for informed decision-making and strategic planning.

By leveraging predictive analytics, water management authorities can anticipate demand fluctuations, optimise resource allocation, and enhance the efficiency of distribution networks. Similarly, in sanitation, data mining assists in monitoring system performance, detecting potential failures, and mitigating health risks by providing early warnings of contamination events. Moreover, the adoption of robust data management frameworks ensures the integration, storage, and accessibility of diverse datasets, supporting real-time monitoring and long-term strategic initiatives. Challenges such as data privacy, accuracy, and the need for interdisciplinary collaboration need to be addressed to ensure the reliability and efficacy of these systems. The convergence of data mining and management in water and sanitation sectors holds significant promise for enhancing operational efficiency, ensuring resource sustainability, and safeguarding public health.

Data integration poses a significant hurdle due to varying formats and structures across different sources. The main challenge is data quality and accuracy with issues like missing values and outliers. Water databases may contain diverse types of data, including spatial, temporal, and multi-dimensional information. Integrating and reconciling these different types of data can be challenging, especially when they come from various sources with distinct formats and structures.

Machine learning is a rapidly expanding field of computer science with diverse applications. Understanding seasonal rainfall changes is crucial for both academic and societal objectives. Current data on surface and groundwater, including water quality and quantity was reviewed. Collecting real-time data, using automated sensors, and integrating remote sensing technologies helped understand water quality dynamics. Data from the Geographical Information System (GIS) was collected to better understand the spatial distribution of water quality and quantity factors. Integrating GIS data enhances our understanding of water resources.

After collection, data was cleaned for accuracy and dependability. After analysing the water datasets, it was found that the random forest (RF) method outperforms all the water quality classification models tested in this project, with a good combination of precision and recall across both classes. Although support vector machines are good at identifying negative classes, they struggle with positive ones. Linear regression, also known as logistic regression, has limits when separating water quality groups. Although decision tree models have balanced performance, there is still potential for development. When estimating water volume, both linear regression and RF models do moderately well, but the latter struggles to capture the underlying patterns. A negative R2 score for the RF model implies a lack of substantial predictive potential, necessitating additional research or evaluation of alternative models.

Primary author: Dr SULIMAN, Ridhwaan (CSIR)

Co-authors: Ms MALEKA, Prettier Morongoa (CSIR-Next-Gen Enterprises and Intuitions); Dr TSHWANE, David (CSIR/University of Limpopo)

Presenter: Dr SULIMAN, Ridhwaan (CSIR)

Session Classification: Session