

ABSTRAK

Kualitas air merupakan aspek penting dalam kesehatan masyarakat, di mana Total Coliform digunakan sebagai indikator utama pencemaran mikrobiologis. Pengukuran konvensional melalui metode laboratorium memerlukan waktu yang lama dan tidak adaptif terhadap perubahan kondisi lingkungan secara cepat. Penelitian ini bertujuan menganalisis dan membandingkan kinerja algoritma ensemble tree-based Random Forest dan Gradient Boosting dalam memprediksi nilai Total Coliform pada Indian Water Quality Dataset dari Kaggle.

*Dataset terdiri dari 1.599 data bersih dengan 6 fitur fisikokimia sebagai masukan. Seleksi fitur dilakukan secara manual berbasis pertimbangan domain dan divalidasi menggunakan Mutual Information Regression sebagai pembanding. Untuk menangani distribusi target yang sangat tidak simetris ($skewness = 28,42$), diterapkan transformasi logaritmik $\log_{1p}$ sehingga menghasilkan empat varian model: RF_raw, RF_log, GB_raw, dan GB_log. Optimasi parameter setiap model dilakukan menggunakan GridSearchCV dengan 3-fold cross-validation**

Model RF_log memperoleh performa terbaik dengan $R^2 = 0,7369$, $MAE = 3.917,49$ MPN/100ml, dan $RMSE = 14.653,55$ MPN/100ml, diikuti GB_raw dengan $R^2 = 0,4126$. Model RF_raw dan GB_log menghasilkan R^2 negatif yang mengindikasikan kegagalan prediksi. Hasil feature selection menggunakan Mutual Information Regression mengidentifikasi B.O.D., Temp, dan Nitrat+Nitrit sebagai fitur paling informatif, dan model RF_log yang dilatih ulang dengan ketiga fitur tersebut menghasilkan $R^2 = 0,7604$ dengan selisih performa yang kecil, mengkonfirmasi bahwa seleksi fitur manual sudah memadai secara kuantitatif. Penelitian ini menyimpulkan bahwa kombinasi transformasi $\log_{1p}$ dengan algoritma Random Forest merupakan pendekatan paling efektif untuk memodelkan distribusi Total Coliform yang ekstrem pada dataset kualitas air ini.

Kata kunci: *Total Coliform, Random Forest, Gradient Boosting, machine learning, kualitas air*