# ABSTRAK

Proses analisis dan klasifikasi kualitas air sungai yang digunakan oleh DLH Kota Yogyakarta sangat panjang. Selain itu, data historis kualitas air sungai yang dimiliki oleh DLH Kota Yogyakarta hanya digunakan untuk pelabelan status mutu air, tanpa eksplorasi lebih lanjut mengenai pola-pola yang ada. Penelitian ini bertujuan untuk memberikan alternatif baru kepada DLH Kota Yogyakarta dengan membangun sistem prediksi kualitas air sungai menggunakan algoritma *Random Forest* dan kerangka kerja CRISP-DM, serta memberikan analisis terkait pola-pola data yang ditemukan selama proses pembuatan model prediksi.

Penelitian ini menggunakan kerangka kerja CRISP-DM yang terdiri dari tahapan *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Data yang digunakan dalam penelitian ini adalah data kualitas air sungai Kota Yogyakarta tahun 2022–2023, yang telah dilabeli dengan status mutu oleh DLH Kota Yogyakarta berdasarkan PermenLHK No. 27 tahun 2021. *Dataset* ini terdiri dari 342 baris data dengan delapan parameter sebagai fitur dan satu atribut sebagai label. Empat skenario *data preprocessing* diterapkan untuk menangani *missing values* dan *outlier*. Algoritma *Random Forest* digunakan untuk membangun model prediksi, dan hasil *modeling* dievaluasi menggunakan *confusion matrix*.

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* menghasilkan akurasi sebesar 95% dalam prediksi kualitas air sungai. *Fecal Coliform* adalah parameter yang paling signifikan dalam menentukan kualitas air, diikuti oleh pH dan *Chemical Oxygen Demand* (COD). *Dissolved Oxygen* (DO), *Biochemical Oxygen Demand* (BOD), Fosfat, *Total Suspended Solids* (TSS), dan Nitrat (N) juga berkontribusi terhadap prediksi kualitas air, namun dengan pengaruh yang lebih kecil. Model dengan delapan fitur sesuai dengan PermenLHK No. 27 tahun 2021 terbukti lebih optimal dibandingkan dengan model yang menggunakan pengurangan fitur. Sistem prediksi kualitas air sungai yang dibangun berbasis *website* memungkinkan DLH Kota Yovyakarta untuk melakukan prediksi kualitas air secara mandiri dan efisien tanpa memerlukan keahlian teknis yang mendalam dalam *machine learning* atau pemrograman.

Kata kunci: kualitas air sungai, CRISP-DM, *random forest,* prediksi

# *ABSTRACT*

*The problem addressed in this study is the lengthy process of analysis and classification of river water quality used by the Environmental Agency (DLH) of Yogyakarta City, as well as the lack of in-depth analysis of historical river water quality data. The historical data owned by DLH Yogyakarta is only used for labeling water quality status, without further exploration of the existing data patterns. This research aims to provide a new alternative for DLH Yogyakarta by building a river water quality prediction system using the Random Forest algorithm and the CRISP-DM Framework, and to provide an analysis of the data patterns found during the model development process.*

*This research uses the CRISP-DM Framework, consisting of the stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The data used in this study is the river water quality data of Yogyakarta City for the years 2022–2023, which has been labeled with quality status by DLH Yogyakarta based on PermenLHK No. 27 of 2021. This dataset consists of 342 rows of data with eight parameters as features and one attribute as the label. Four data preprocessing scenarios were applied to handle missing values and outliers. The Random Forest algorithm was used to build the prediction model, and the modeling results were evaluated using a confusion matrix.*

*The results of the study indicate that the Random Forest algorithm achieved an accuracy of 95% in predicting river water quality. Fecal Coliform is the most significant parameter in determining water quality, followed by pH and Chemical Oxygen Demand (COD). Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Phosphate, Total Suspended Solids (TSS), and Nitrate (N) also contribute to the prediction of water quality, but with lesser influence. The model with eight features as per PermenLHK No. 27 of 2021 proved to be more optimal compared to the model with reduced features. The river water quality prediction system built as a web-based application allows DLH Yogyakarta to predict water quality independently and efficiently without requiring deep technical expertise in machine learning or programming.*



*Keywords: river water quality, CRISP-DM, random forest, prediction*