PLAGIAT MERUPAKAN TINDAKAN TIDAK TERPUJI

ABSTRAK

Penelitian ini mengevaluasi unjuk kerja algoritma *Gradient Boosting Decision Tree* (GBDT), yakni CatBoost dan LightGBM, dalam mengestimasi Indeks Standar Pencemar Udara (ISPU) perkotaan berdasarkan kerangka *Knowledge Discovery in Databases* (KDD). Tantangan utama dalam peramalan kualitas udara spasial-temporal ini adalah tingginya volatilitas polutan gas serta risiko pergeseran tren struktural mendadak (*regime shift*). Penelitian ini menggunakan 1.186 rekaman observasi harian dari stasiun Air Quality Monitoring System (AQMS) Gondokusuman, Yogyakarta, yang mencakup periode 2022–2024 (pelatihan) dan awal 2026 (pengujian) untuk enam parameter utama (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO) yang dipartisi secara kronologis dengan rasio 70:20:10. Metodologi yang diimplementasikan secara rinci meliputi pra-pemrosesan dengan mempertahankan *outlier* sebagai representasi sinyal *regime shift* serta penanganan *missing values* via *K-Nearest Neighbors* (KNN) *Imputer* dengan kalibrasi k=7. rekayasa fitur komprehensif yang menginkorporasikan fitur kelambanan (*lag* 1–13 hari), rata-rata bergerak (*rolling mean* 7 dan 14 hari), fitur temporal kalender, dan *cyclical encoding* (*sin/cos*), optimasi *hyperparameter* melalui *Grid Search*, dan pengujian model menggunakan *5-Fold Time-Series Cross-Validation*. Hasil evaluasi menunjukkan bahwa algoritma CatBoost, khususnya pada skenario jendela waktu historis 14 hari, mendemonstrasikan unjuk kerja paling stabil dan akurat untuk polutan partikulat (PM_{10} dan $PM_{2.5}$). Hal ini dibuktikan dengan capaian tingkat kesalahan terendah pada *test set*, yakni nilai *Root Mean Squared Error* (RMSE) sebesar 6,530 dan 6,826. Struktur *Symmetric Trees* pada CatBoost terbukti tangguh memitigasi fluktuasi tajam akibat *regime shift* dibandingkan LightGBM. Sebaliknya, kedua algoritma menghadapi limitasi prediktif pada parameter gas (CO, NO_2, SO_2) yang ditandai dengan perolehan R^2 negatif, menegaskan sifat gas yang sangat stokastik di atmosfer. Disimpulkan bahwa algoritma CatBoost dengan *lookback window* 14 hari sangat direkomendasikan untuk implementasi taktis sistem prediksi partikulat kualitas udara.

Kata Kunci: Kualitas Udara, AQMS, CatBoost, LightGBM, *Time-Series Cross-Validation*, *Feature Engineering*.

PLAGIAT MERUPAKAN TINDAKAN TIDAK TERPUJI

ABSTRACT

This study evaluates the predictive performance of Gradient Boosting Decision Tree (GBDT) algorithms, namely CatBoost and LightGBM, in estimating the urban Air Pollutant Index (API/ISPU) based on the Knowledge Discovery in Databases (KDD) framework. The primary challenges in spatial-temporal air quality forecasting are the high volatility of gaseous pollutants and the risk of sudden structural trend alterations (regime shifts). This study utilizes 1,186 daily observation records from the Gondokusuman Air Quality Monitoring System (AQMS) in Yogyakarta*, covering the* 2022–2024 period (training) and early 2026 (testing) *for six main parameters* (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO)*, partitioned chronologically at a* 70:20:10 ratio*. The detailed methodology encompasses: preprocessing that retains outliers as representative signals of regime shifts, and missing values imputation via K-Nearest Neighbors (KNN) Imputer with a calibration of k=7, comprehensive feature engineering incorporating lag features (1–13 days), rolling means (7 and 14 days), calendar temporals, and cyclical encoding (sin/cos), hyperparameter optimization via Grid Search, and model testing using 5-Fold Time-Series Cross-Validation. Evaluation results indicate that CatBoost, particularly with a 14-day historical lookback window scenario, demonstrates the most stable and accurate performance for particulate pollutants (PM_{10} and $PM_{2.5}$). This is evidenced by achieving the lowest error rates on the test set, with Root Mean Squared Error (RMSE) values of 6.530 and 6.826, respectively. CatBoost's Symmetric Trees structure proves to be highly robust in mitigating sharp fluctuations caused by regime shifts compared to LightGBM. Conversely, both algorithms face predictive limitations on gaseous parameters (CO, NO2, SO2), marked by negative R^2 scores, affirming the highly stochastic nature of gases in the atmosphere. It is concluded that CatBoost with a 14-day lookback window is highly recommended for the tactical implementation of particulate air quality prediction systems.*

Keywords: *Air Quality, AQMS, CatBoost, LightGBM, Time-Series Cross-Validation, Feature Engineering.*