

ABSTRAK

Kualitas udara merupakan isu penting di Jakarta yang mengalami polusi tinggi akibat aktivitas kendaraan dan industri. Penelitian ini bertujuan untuk mengklasifikasikan kualitas udara secara akurat dengan mengimplementasikan algoritma *Gradient Boosting* dalam kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*), yang terdiri dari enam tahapan utama, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Data yang digunakan merupakan data dari Dinas Lingkungan Hidup Provinsi DKI Jakarta yang diperoleh melalui Open Data Jakarta, dengan periode waktu 2010 hingga 2021 sebanyak 21.915 data. Parameter utama yang digunakan dalam penelitian ini meliputi lima fitur pencemar udara, yaitu PM₁₀, SO₂, CO, O₃, dan NO₂. Tahap *preprocessing* melibatkan proses data *cleaning*, transformasi data, serta penyeimbangan kelas menggunakan teknik *undersampling* untuk mengatasi ketidakseimbangan distribusi kelas. Berdasarkan analisis *feature importance*, fitur PM₁₀, SO₂, CO, O₃, dan NO₂ memiliki kontribusi terhadap hasil klasifikasi, dengan O₃ (ozon) sebagai fitur yang paling berpengaruh. Hasil evaluasi menunjukkan bahwa akurasi model mengalami peningkatan sebesar 0,25% setelah dilakukan *balancing*, dari 96,87% menjadi 97,11%. Proses optimasi model dilakukan menggunakan *Grid Search* untuk menemukan kombinasi hyperparameter terbaik. Evaluasi akhir menggunakan *confusion matrix* serta metrik akurasi, presisi, *recall*, dan *F1-score* menunjukkan bahwa algoritma *Gradient Boosting* mampu melakukan klasifikasi kualitas udara secara akurat dan efektif, dengan akurasi tertinggi mencapai 97,25%.

Kata kunci: Kualitas Udara, Klasifikasi, *Gradient boosting*, CRISP-DM, Jakarta

ABSTRACT

Air quality is a critical issue in Jakarta, which experiences high levels of pollution due to vehicle and industrial activities. This study aims to accurately classify air quality by implementing the Gradient Boosting algorithm within the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which consists of six main stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The data used in this research was obtained from the Environmental Agency of DKI Jakarta Province through Open Data Jakarta, covering the period from 2010 to 2021, with a total of 21,915 records. The main parameters used include five air pollutant features: PM_{10} , SO_2 , CO , O_3 , and NO_2 . The preprocessing stage involved data cleaning, data transformation, and class balancing using the undersampling technique to address class imbalance. Based on feature importance analysis, PM_{10} , SO_2 , CO , O_3 , and NO_2 were found to contribute to the classification results, with O_3 (ozone) being the most influential feature. Evaluation results showed an accuracy improvement of 0.25% after balancing, from 96.87% to 97.11%. Model optimization was carried out using Grid Search to find the best hyperparameter combination. Final evaluation using the confusion matrix and metrics such as accuracy, precision, recall, and F1-score demonstrated that the Gradient Boosting algorithm can classify air quality accurately and effectively, achieving a highest accuracy of 97.25%

Keywords: Air Quality, Classification, Gradient Boosting, CRISP-DM, Jakarta