

ABSTRAK

Pencemaran udara merupakan salah satu permasalahan lingkungan yang signifikan di DKI Jakarta dan membutuhkan metode klasifikasi yang akurat untuk memantau kualitas udara. Penelitian ini bertujuan untuk mengevaluasi performa algoritma *Decision Tree C4.5* dan *Random Forest* dalam mengklasifikasikan kualitas udara berdasarkan parameter utama Indeks Standar Pencemar Udara (ISPU), yaitu PM_{10} , SO_2 , CO , O_3 , dan NO_2 . Data penelitian diperoleh dari portal satudata.jakarta.go.id dengan jumlah 6.874 baris, yang kemudian diproses melalui tahapan data cleaning, normalisasi, *feature selection*, serta penyeimbangan data menggunakan *Synthetic Minority Oversampling Technique (SMOTE)*. Model dibangun dan divalidasi menggunakan teknik *K-Fold Cross Validation* dengan variasi $K=5, 7$, dan 9 . Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memberikan performa lebih baik dibandingkan *Decision Tree C4.5*, baik pada kondisi *non-balancing* maupun *balancing*. Pada data *balancing*, *Random Forest* mencapai akurasi tertinggi sebesar 99,79% dengan *precision*, *recall*, dan *F1-score* yang konsisten tinggi (99,79%), sedangkan *Decision Tree C4.5* mendapatkan akurasi 99,65%. Temuan ini mengindikasikan bahwa *Random Forest* lebih unggul untuk diterapkan dalam klasifikasi pencemaran udara di DKI Jakarta.

Kata Kunci: *Decision Tree C4.5*, *Random Forest*, *SMOTE*, *K-Fold Cross Validation*, Klasifikasi Pencemaran Udara

ABSTRACT

Air pollution is one of the most significant environmental issues in DKI Jakarta and requires accurate classification methods to monitor air quality. This study aims to evaluate the performance of the Decision Tree C4.5 and Random Forest algorithms in classifying air quality based on the key parameters of the Air Pollution Standard Index (ISPU), namely PM10, SO₂, CO, O₃, and NO₂. The dataset used in this research was obtained from the satudata.jakarta.go.id portal, consisting of 6,874 records, and was processed through several stages including data cleaning, normalization, feature selection, and data balancing using the Synthetic Minority Oversampling Technique (SMOTE). The models were built and validated using K-Fold Cross Validation with K values of 5, 7, and 9. The results indicate that the Random Forest algorithm outperforms the Decision Tree C4.5 under both non-balancing and balancing conditions. On the balanced dataset, Random Forest achieved the highest accuracy of 99.79%, with precision, recall, and F1-score consistently reaching 99.79%, while Decision Tree C4.5 obtained an accuracy of 99.65%. These findings suggest that Random Forest is more effective and reliable for air pollution classification in DKI Jakarta.

Keywords: Decision Tree C4.5, Random Forest, SMOTE, K-Fold Cross Validation, Air Pollution Classification