CLASSIFICATION SYSTEM OF AIR CONDITIONS IN DKI JAKARTA USING RANDOM FOREST ALGORITHM WITH SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

RIYAN ZAENAL ARIFIN

Data Science Study Program, Faculty of Science & Technology
University of Technology Yogyakarta
Jl. Ringroad Utara Jombor Sleman Yogyakarta
E-mail : riyan.5201811011@student.uty.ac.id

ABSTRACT

Air quality in big cities like DKI Jakarta is often a special concern for the public and especially stakeholders. This study aims to create an air condition classification system in DKI Jakarta using the Random Forest algorithm with the Synthetic Minority Over-sampling Technique (SMOTE) sampling technique. The Random Forest algorithm was chosen because of its ability to handle high-dimensional data and its tolerance to overfitting. Meanwhile, the SMOTE technique is used to overcome the problem of class imbalance in air quality data. This study uses the 2023 DKI Jakarta Air Pollution Standard Index dataset of 1825 data and 12 attributes. The data preparation process includes label selection, data normalization, feature selection, deleting missing, label encoding, SMOTE, and deleting duplicate data. From this process, 3237 balanced data were obtained. Testing data was taken from the dataset of 50 data for model test samples if deployment has been carried out, while the rest was used for model training data using 10 fold cross validation. The results of the metric evaluation show the same accuracy, precision, recall, f1 score values, which are 100% for each K in cross validation, both evaluations on validation data and training data. The deployment process uses the first model of cross validation. The results of this model have better performance than previous studies. This shows the good potential of using the Random Forest algorithm with SMOTE in monitoring and predicting air quality. Researchers suggest creating a more complex deployment model, such as connecting to a third party and using your own server to run the application and trying the case in another city.

Keywords: ISPU, Random Forest, SMOTE, Cross Validation, Deployment