Implementing K-Nearest Neighbor (KNN) Using Grid Search CV to Predict the Severity of Acute Respiratory Infections (ARI)

(Case Study: Mlati II Community Health Center, Sleman)

Nofrianus Adiputra Mokodompit, Rr. Hajar Puji Sejati

Program Studi Informatika Medis Program Sarjana, Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Jl. Ringroad Utara Jombor Sleman Yogyakarta
E-mail: nofriamokodompit3@gmail.com,

ABSTRACT

Acute Respiratory Tract Infection (ARI) is the disease with the highest prevalence at the Mlati II Community Health Center in Sleman, Yogyakarta. The high number of patient visits due to ARI necessitates a decision support system capable of identifying risk early and accurately. However, the current information system at the community health center is static and lacks predictive analytics features. This study aims to develop an ARI risk prediction model based on the K-Nearest Neighbor (KNN) algorithm, with hyperparameter optimization using Grid Search Cross Validation (GridSearchCV), and to implement it as a web-based prediction system for use by medical personnel. Initial data were collected from the medical records of 200 patients at the Mlati II Community Health Center. After comprehensive preprocessing, the dataset comprised 120 patients with 12 clinical features, including body temperature, blood pressure, respiratory rate, and symptoms such as cough, runny nose, shortness of breath, and sore throat. The preprocessing steps involved imputing missing values, label encoding, normalization using StandardScaler, and class balancing with the SMOTE technique. The dataset was split into 80% training data (96 samples) and 20% test data (24 samples). The KNN model was optimized by testing various combinations of n_neighbors, weights, and distance metrics using GridSearchCV with 5-fold cross-validation. The results showed that the KNN model with optimal parameters (n_neighbors = 9, weights = 'distance', metric = 'manhattan') achieved an accuracy of 97.50% on the test data, with a weighted average F1-score of 0.98 and a cross-validation score of 97.69%. The model successfully classified all severe ARI cases correctly (recall = 1.00), demonstrating very high sensitivity to critical cases. Model validation was conducted using three approaches: technical validation with a Python model, which showed 100% accuracy on 10 new test samples; mathematical validation through manual calculations, confirming 100% consistency with the model results; and clinical validation using expert diagnosis, which achieved 90% accuracy, with one discrepancy in sample S6, but maintained 100% recall for severe case detection. The model was then integrated into a web-based system to facilitate its use by medical personnel in diagnosis and decision-making. Based on the evaluation results, the optimized KNN model proved effective and reliable in classifying ARI risk. The implementation of this system is expected to improve the quality of primary healthcare services, particularly by enhancing the speed and accuracy of ARI risk identification.

Keywords: ARI, K-Nearest Neighbor, Grid Search CV, Risk Prediction, Decision Support System