ABSTRAK

Data leakage merupakan permasalahan metodologis yang sering terjadi pada *pipeline machine learning*, khususnya pada tahap *preprocessing*, dan berpotensi menginflasi performa model secara tidak valid. Penelitian ini menganalisis dampak strategi *preprocessing* terhadap risiko *data leakage* serta pengaruhnya terhadap evaluasi kinerja model pada tugas klasifikasi klinis. Pendekatan kuantitatif eksperimental diterapkan menggunakan *dataset* klinis *pediatric appendicitis* yang terdiri atas fitur numerik dan kategorikal. Dua skenario *preprocessing* dibandingkan, yaitu *leakage-free pipeline*, di mana pemisahan data dilakukan sebelum seluruh transformasi, dan *leakage-prone pipeline*, di mana *preprocessing* diterapkan sebelum *data splitting*. Tahapan yang dianalisis meliputi imputasi *missing values*, *encoding* fitur kategorikal, penanganan ketidakseimbangan kelas, seleksi fitur, dan *feature scaling*. Model *Logistic Regression* dan XGBoost dievaluasi menggunakan ROC-AUC sebagai metrik utama dan F1-score sebagai metrik klasifikasi pendukung. Hasil eksperimen menunjukkan bahwa skenario *leakage* menghasilkan nilai ROC-AUC *test* yang lebih tinggi, berada pada rentang 0,966–0,972, dibandingkan skenario *leakage-free* yang hanya mencapai 0,941–0,948. Selain itu, konfigurasi *leakage* juga menunjukkan nilai F1-score yang lebih tinggi (0,884–0,907) dibandingkan skenario *leakage-free* (0,802–0,834). Meskipun tampak unggul secara evaluasi, peningkatan performa pada skenario *leakage* tidak merefleksikan kemampuan generalisasi yang valid. Analisis lebih lanjut mengidentifikasi bahwa *feature selection* dan *feature scaling* yang dilakukan sebelum *data splitting* merupakan kontributor utama inflasi kinerja model. Temuan ini menegaskan pentingnya perancangan *preprocessing pipeline* yang ketat untuk memastikan evaluasi model *machine learning* yang valid dan andal secara ilmiah.

Kata Kunci: *data leakage*, *preprocessing pipeline*, evaluasi *machine learning*, *roc-auc*, *f1-score* klasifikasi medis.

ABSTRACT

Data leakage is a methodological issue that frequently occurs in machine learning pipelines, particularly during the preprocessing stage, and can lead to artificially inflated model performance. This study investigates the impact of preprocessing strategies on data leakage risk and their effects on model evaluation in a clinical classification task. A quantitative experimental approach was conducted using a pediatric appendicitis clinical dataset comprising numerical and categorical features. Two preprocessing scenarios were compared: a leakage-free pipeline, in which data splitting is performed prior to all transformations, and a leakage-prone pipeline, in which preprocessing steps are applied before data splitting. The analyzed stages include missing value imputation, categorical feature encoding, class imbalance handling, feature selection, and feature scaling. Logistic Regression and XGBoost models were evaluated using ROC-AUC as the primary metric and F1-score as the supporting classification metric. The experimental results show that leakage-prone preprocessing yields higher test ROC-AUC values (0.966–0.972) compared to the leakage-free scenario (0.941–0.948). Similarly, higher F1-scores are observed under leakage conditions (0.884–0.907) than in leakage-free configurations (0.802–0.834). Despite the apparent performance gains, the inflated results produced by leakage-prone pipelines do not reflect valid generalization capability. Further analysis identifies feature selection and feature scaling performed prior to data splitting as the primary contributors to performance inflation. These findings highlight the critical importance of rigorously designed preprocessing pipelines to ensure valid, reliable, and scientifically sound evaluation of machine learning models.

Keywords: *data leakage, preprocessing pipeline, machine learning evaluation, roc-auc, f1-score, medical classification.*