INTISARI

EXTENDED ISOLATION FOREST SEBAGAI METODE PREPROCESSING PADA LIGHT GRADIENT BOOSTING MACHINE UNTUK DATA YANG TERKONTAMINASI OUTLIER DAN TIDAK SEIMBANG

Oleh

Meilin Budiarti

23/530378/PPA/06732

Data yang terkontaminasi outlier dan memiliki distribusi kelas tidak seimbang dapat menurunkan kualitas data serta mengganggu stabilitas kinerja model klasifikasi. Isolation Forest (IF) umum digunakan untuk deteksi outlier, namun mekanisme *axis-parallel split* berpotensi menghasilkan skor anomali yang kurang konsisten pada data berdimensi tinggi. Penelitian ini bertujuan mengevaluasi Extended Isolation Forest (EIF) sebagai metode pre-processing untuk meningkatkan kualitas data serta menganalisis pengaruhnya terhadap performa dan stabilitas klasifikasi menggunakan Light Gradient Boosting Machine (LightGBM).

Metode penelitian meliputi perbandingan IF dan EIF dalam mendeteksi outlier, di mana hasil deteksi digunakan sebagai tahap pra-pemrosesan sebelum klasifikasi LightGBM. Evaluasi kinerja dilakukan menggunakan metrik Accuracy, F1-Score, dan ROC-AUC pada enam dataset publik, yaitu Ionosphere, Satellite, Cardio, BreastW, Annthyroid, dan Pima, yang mewakili variasi karakteristik data dan tingkat ketidakseimbangan kelas.

Hasil penelitian menunjukkan bahwa EIF menghasilkan skor anomali yang lebih stabil dan terkonsentrasi dibandingkan IF. Penggunaan EIF sebagai tahap pre-processing menghasilkan data yang lebih homogen dan mendukung kestabilan kinerja LightGBM, terutama pada dataset dengan struktur kompleks. Dengan demikian, EIF efektif digunakan sebagai metode pre-processing untuk menjaga kualitas data dan meningkatkan kestabilan pipeline klasifikasi berbasis LightGBM pada data yang terkontaminasi outlier dan tidak seimbang.

Kata Kunci: Extended Isolation Forest, Isolation Forest, Light Gradient Boosting Machine, Deteksi Outlier, Data Tidak Seimbang

ABSTRACT

EXTENDED ISOLATION FOREST AS A PREPROCESSING METHOD FOR LIGHT GRADIENT BOOSTING MACHINE ON OUTLIER-CONTAMINATED AND IMBALANCED DATA

By

Meilin Budiarti

23/530378/PPA/06732

Data contaminated with outliers and having an unbalanced class distribution can reduce data quality and disrupt the stability of classification model performance. Isolation Forest (IF) is commonly used for outlier detection, but the axis-parallel split mechanism has the potential to produce inconsistent anomaly scores in high-dimensional data. This study aims to evaluate Extended Isolation Forest (EIF) as a pre-processing method to improve data quality and analyze its effect on classification performance and stability using Light Gradient Boosting Machine (LightGBM).

The research method includes comparing IF and EIF in detecting outliers, where the detection results are used as a pre-processing stage before LightGBM classification. Performance evaluation was conducted using Accuracy, F1-Score, and ROC-AUC metrics on six public datasets, namely Ionosphere, Satellite, Cardio, BreastW, Annthyroid, and Pima, which represent variations in data characteristics and class imbalance levels.

The results show that EIF produces more stable and concentrated anomaly scores compared to IF. The use of EIF as a pre-processing stage produces more homogeneous data and supports the stability of LightGBM performance, especially on datasets with complex structures. Thus, EIF is effective as a pre-processing method to maintain data quality and improve the stability of LightGBM-based classification pipelines on data contaminated with outliers and imbalances.

Keywords: Extended Isolation Forest, Isolation Forest, Light Gradient Boosting Machine, Outlier Detection, Imbalanced Data