

Studi Komparatif Model Unsupervised Outlier Detection dengan Pembuatan Data Sintetis Realistik: Studi Kasus Data Finansial Tidak Berlabel = Comparative Study of Unsupervised Outlier Detection Models with Realistic Synthetic Data Generation: A Case Study of Unlabeled Financial Data

Rendya Yuschak, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920553369&lokasi=lokal>

Abstrak

Dalam era digital saat ini, banyaknya data finansial yang melimpah dan tidak berlabel menimbulkan tantangan dalam pemilihan teknik pendeteksian outlier (outlier detection) yang optimal. Penelitian ini bertujuan untuk menangani tantangan tersebut dengan membandingkan model unsupervised outlier detection pada data sintetis yang dirancang untuk meniru karakteristik data finansial nyata. Sebagai studi kasus, penelitian ini menggunakan data Laporan Harta Keuangan Penyelenggara Negara (LHKPN) tahun 2022. Proses penelitian mencakup pengumpulan data, pemrosesan, pembuatan data sintetis, pengujian sepuluh algoritma outlier detection, dan penerapan model terbaik pada data LHKPN tahun 2022. Dari proses ini, model Median Absolute Deviation (MAD) dengan threshold 7.8 teridentifikasi sebagai yang paling efektif pada data sintetis yang meniru data LHKPN. Penelitian ini juga menemukan hyperparameter terbaik untuk model lain dan melakukan analisis skor outlier pada data nyata. Hasilnya memberikan wawasan baru dan menunjukkan potensi investigasi lanjutan dalam outlier detection pada data finansial tidak berlabel, dengan pendekatan yang menyeluruh mulai dari analisis distribusi data hingga pengujian model pada data sintetis dan asli.

.....In the current digital era, the abundance of unlabeled financial data poses challenges in selecting optimal outlier detection techniques. This research aims to address these challenges by comparing unsupervised outlier detection models on synthetic data, designed to mimic real financial data characteristics. Using 2022 data from the Laporan Harta Keuangan Penyelenggara Negara (LHKPN) as a case study, the research process includes data collection, processing, creating synthetic data, testing 10 outlier detection algorithms, and applying the most effective model, identified as Median Absolute Deviation (MAD) with a threshold of 7.8, on synthetic data based on LHKPN data. The study also finds the best hyperparameters for other models and conducts real data outlier score analysis, providing new insights and demonstrating further investigation potential in outlier detection for unlabeled financial data.