Perbandingan Kinerja Algoritma *Unsupervised Machine Learning* untuk Deteksi Anomali dalam Proses ETL

Muhammad Faisal Ashshidiq[1)], Mohammad Nurkamal Fauzan[2)], RN Nuraini[3)]

[1)2)3)] Teknik Informatika, Universitas Logistik dan Bisnis Internasional
Jl. Sariasih No.54, sarijadi, kece.sukasari, Kota Bandung, Jawa Barat
[1)] sidiqfaisal30@gmail.com
[2)] m.nurkamal.f@ubli.ac.id
[3)] nurainisf@ulbi.ac.id

Abstrak

Penelitian ini melakukan perbandingan komprehensif terhadap tiga algoritma unsupervised machine learning untuk deteksi anomal dalam proses *Extract, Transformasi, Load* (ETL). Algoritma yang dibandingkan adalah *Isolation Forest, Local Outlier Factor*, dan *One-Class Support Vector Machine* (OC-SVM). Penelitian ini menggunakan dataset dengan struktur nested array yang umum ditemukan pada aplikasi berbasis web dan *Internet of Things* (IoT). Hasil penelitian menunjukkan bahwa *Isolation Forest* memberiikan performa terbaik dengan nilai F1-Score 0.723, accuracy 0.935, precision 0.567 dan recall 1.00. *Local Outlier Factor* menunjukkan performa terendah dengan F1-Score 0.221, dan One-Class SVM memberikan performa moderat dengan F1-Score 0.488. Hasil visualisasi menggunakan *Principal Component Analysis* (PCA) untuk memperkuat temuan dalam memisahkan data normal dan anomali. penelitian ini memberikan kontribusi penting dalam pemilihan algoritma deteksi anomaly yang tepat untuk menjaga kualitas data setelah proses ETL.

Kata kunci: deteksi anomali, ETL, *unsupervised machine learning*, kualitas data, *isolation forest*

Abstract

This study conducts a comprehensive comparison of three unsupervised machine learning algorithms for anomaly detection in the Extract, Transform, Load (ETL) process. The algorithms compared are Isolation Forest, Local Outlier Factor, and One-Class Support Vector Machine (OC-SVM). This study uses a dataset with a nested array structure commonly found in web-based applications and the Internet of Things (IoT). The results show that Isolation Forest provides the best performance with an F1-Score of 0.723, accuracy of 0.935, precision of 0.567, and recall of 1.00. Local Outlier Factor shows the lowest performance with an F1-Score of 0.221, and One-Class SVM provides moderate performance with an F1-Score of 0.488. Visualization results using Principal Component Analysis (PCA) reinforce the findings in separating normal and anomalous data. This study makes an important contribution to the selection of the appropriate anomaly detection algorithm to maintain data quality after the ETL process.

Keywords: *anomaly detection, ETL, unsupervised machine learning, data quality, isolation forest*

1. PENDAHULUAN

Di era digital, data telah berkembang menjadi aset strategis, penting bagi organisasi atau perusahaan[1]. Secara umum data merupakan sekumpulan angka, huruf dan simbol yang dapat membantu merepresentasikan suatu kondisi kemudian dapat diolah untuk menghasilkan informasi yang penting. Proses mengolah data menggunakan suatu metode tertentu untuk merapihkan dan membersihkan data supaya dapat digunakan oleh analisis[2]. keunggulan yang kompetitif, inovasi produk, dan efisiensi operasional tercipta melalui proses pengelolaan data yang dapat mempercepat waktu, karena proses mengelola ini mempunyai kemampuan untuk mengumpulkan, mengelola. proses pengelolaan data ini

DOI: https://doi.org/10.47111/jointecoms.v5i3 *Received: 01-09-2025*
Accepted: 10-09-2025

memiliki kemampuan dalam mengumpulkan data skala besar[3]. semakin banyak data yang memiliki struktur kompleks seperti *nested array*. data modern sering digunakan pada aplikasi berbasis web, selain itu digunakan pada *Internet of Things* (IoT) untuk menghasilkan data yang semi-struktur. data tersebut tidak sesuai untuk basis data relasional konvensional, Akibatnya teknologi NoSQL seperti MongoDB menjadi pilihan utama karena fleksibilitas skema data yang lebih baik dan sklabilitas, teknologi ini digunakan untuk menyimpan data dengan struktur *nested array* (larik bersarang).

Meskipun MongoDB fleksibel dalam struktur *nested array*, hal ini menciptakan tantangan bagi analisis data[3]. Tantangan bagi analisis adalah proses transformasi skema Platform atau aplikasi yang umum untuk analisis data adalah *Platform Business Intelligence* (BI), namun di dalam penelitian ini Metabase merupakan sebagai alat analisis.

Sebelum proses analisis, dibutuhkan data dengan struktur tabular untuk memudahkan analisis data. Sebelum proses analisis, dibutuhkan proses pengumpulan data yang siap digunakan untuk proses ETL[4]. Proses ETL mencakup tiga tahapan penting yaitu pengambilan data (*Extract*), transformasi (Transform), dan memuat data (ETL)[5]. Data yang dikumpulkan kemudian ditransformasi dengan metode *flattening*, metode untuk menjalankan proses meratakan struktur *nested array* menjadi struktur tabular[8]. *Flatenning* diperlukan di proses *Extract Transform Load* (ETL) karena proses tersebut rentan terhadap kegagalan, seperti masalah jumlah nilai null yang tinggi. Jumlah nilai null yang tinggi menimbulkan *data corrupt*, atau data terpotong (tidak lengkap) sehingga kualitas data tidak konsisten[6]. Walau kualitas data yang memiliki banyak nilai null tidak seluruhnya merupakan data kosong, terdapat anomali yang mengindikasikan kegagalan dalam proses ETL.

Kehadiran data anomali langsung mengancam integritas dan keandalan kualitas data. Penggunaan data berkualitas rendah untuk analisis akan menghasilkan informasi yang keliru dan mendapatkan kesimpulan yang tidak valid. Data tidak valid itu termasuk dari hasil bahwa data tersebut tidak akurat.

Data yang tidak akurat dapat menimbulkan kerugian bagi organisasi atau perusahaan di saat pengambilan keputusan strategis baik secara finansial maupun reputasi. Oleh karena itu penelitian ini bertujuan memastikan kualitas data yang baik setelah proses ETL. Dibandingkan dengan proses ETL, proses validasi data manual akan tidak efisien untuk data berukuran volume besar (Big Data)[7]. Oleh karena itu, penelitian ini menggunakan metode Mahine Learning untuk deteksi anomali supaya dapat mengidentifikasi data bermasalah dengan cepat dan akurat. Berdasarkan dari tinjauan literatur yang dilakukan pada jurnal intership 2 terdapat tiga algoritma deteksi anomali yang menunjukkan potensi untuk menjaga integritas, menjaga kualitas data di antara lain: (1) Isolation Forest (IF), (2) Local Outlier Factor (LOF), (3) One-Class SVM (OC-SVM).

Isolation Forest merupakan algoritma yang dirancang untuk mendeteksi anomali dengan keunggulan dalam mencari pola dan tren yang tidak biasa[8]. LOF merupakan algoritma deteksi anomali yang di rancang untuk mengukur nilai keanehan berdasarkan dari kepadatan data. OC-SVM merupakan algoritma deteksi anomali yang dirancang untuk megklasifikasi data normal dan mengidentifikasi data point berasal dari luar zona normal.

Penelitian ini diharapkan dapat menyimpulkan hasil dari perbandingan untuk pemilihan algortima deteksi anomali yang sesuai dengan proses pengambilan keputusan bisnis di suatu perusahaan. penelitian ini bertujuan menjadi referensi bagi peneliti selanjutnya dalam kualitas data dan deteksi anomali memastikan kualitas data yang tinggi dalam proses pengambilan keputusan berbasis data.

2. TINJAUAN PUSTAKA

Deteksi anomali merupakan bidang yang penting dalam analisis data yang bertujuan untuk mengidentifikasi pola atau observasi yang secara signifikan berbeda dari mayoritas data[9], [10]. Proses mendeteksi anomali setelah dilakukan proses ETL yang menghasilkan kualitas data. Namun dari hasil akhir menghasilkan kualitas data yang tidak konsisten, sehingga memerlukan

tahapan deteksi anomali dengan menggunakan pendekatan *unsupervised machine learning*[11]. Tujuan deteksi anomali untuk menganalisis kualitas data supaya mengetahui performa kualitas data dan memastikan kesalahan dalam proses transformasi data.

2.1 Extract, Transform, Load (ETL)

Proses ETL merupakan komponen fundamental dalam arsitektur data warehouse dan sistem business intelligence modern[12]. ETL didefinisikan sebagai proses sistematis untuk mengekstrak data dari berbagai sumber heterogen, mentransformasikannya sesuai dengan kebutuhan bisnis, dan memuatnya ke dalam sistem target untuk analisis lebih lanjut[15]. Tahap ekstraksi melibatkan pengambilan data dari berbagai sumber seperti database relasional, file flat, API, dan sistem NoSQL. Setelah ekstraksi berhasil dilakukan selanjutya ke proses transformasi mencakup pembersihan data, normalisasi, agregasi, dan konversi format untuk memastikan konsistensi dan kualitas data. Tahapan loading bertanggung jawab untuk memasukkan data yang telah ditransformasi ke dalam *data warehouse*[13]. Proses ETL memilikii tantangan yaitu melakukan perubahan skema data, struktur data dan sebagainya. Tantanganya yaitu untuk menangani data *nested array* diperlukan teknik atau pendekatan khusus karena struktur hierarkis yang tidak dapat langsung digunakan untuk analisis data[16]. Teknik atau pendekatan yang digunakan adalah *flattening* supaya dapat mengubah struktur nested menjadi format relasional. Namun, proses ini sering menghasilkan anomali data seperti duplikasi, data yang terpotong, dan inkonsistensi tipe data.

2.2 Kualitas Data dan Anomali

Kualitas data merupakan aspek kritis yang menentukan keberhasilan sistem informasi dan pengambilan keputusan berbasis data. Hasil akan di identifikasi empat dimensi yaitu akurasi, kelengkapan, konsistensi, dan ketepatan waktu, karena berbagai faktor dapat terdegradasi. Hal tersebut disebabkan karena adanya anomali, sehingga diperlukan identifikasi pola secara signifikan dari perilaku di dalam dataset. Mengklasifikasikan anomali data menjadi beberapa kategori supaya dapat memperbaiki kesalahan. Dampak anomali terhadap kualitas data sangat signifikan karena menunjukkan bahwa penggunaan kualitas data rendap dapat menyebabkan kerugian finansial hingga 15% dari revenue tahunan oraganisasi.

2.3 *Machine Learning* untuk Deteksi Anomali

Deteksi anomali menggunakan *machine learning* telah menjadi area penelitian yang aktif dalam beberapa dekade terakhir. Berbeda dengan supervised learning yang memerlukan labeled data, unsupervised anomaly detection dapat bekerja dengan data unlabeled dan lebih sesuai untuk skenario real-world di mana anomali jarang terjadi dan sulit didefinisikan secara eksplisit. mengkategorikan metode deteksi anomali unsupervised menjadi beberapa pendekatan: statistical methods, proximity-based methods, clustering-based methods, dan ensemble methods[20]. Setiap pendekatan memiliki kelebihan dan kekurangan yang berbeda tergantung pada karakteristik data dan jenis anomali yang ingin dideteksi.

2.4 Isolation Forest

Isolation Forest merupakan algoritma deteksi anomali yang menggunakan prinsip isolasi untuk mengidentifikasi outlier[21]. Algoritma ini bekerja berdasarkan asumsi bahwa anomali lebih mudah diisolasi dibandingkan dengan data normal karena mereka memiliki karakteristik yang berbeda dan jumlahnya sedikit. Proses kerja Isolation Forest melibatkan konstruksi ensemble dari isolation trees. Setiap tree dibangun dengan memilih secara random sebuah feature dan split value, kemudian membagi data secara rekursif hingga setiap instance terisolasi. Anomali akan memiliki *path length* yang lebih pendek karena lebih mudah dipisahkan dari mayoritas data. Keunggulan algoritma ini adalah efisiensi komputasional yang tinggi dengan kompleksitas waktu dan kemampuan untuk menangani dataset berdimensi tinggi.

Penelitian ini menunjukkan bahwa Isolation Forest efektif untuk deteksi anomali pada data tabular dengan performa yang konsisten [22]. Namun, algoritma ini dapat mengalami penurunan

performa pada data dengan dimensi yang sangat tinggi atau ketika distribusi anomali mirip dengan data normal.

2.5 Local Outlier Factor

Local Outlier Factor merupakan algoritma deteksi anomali berbasis densitas yang mengukur *local deviation* dari sebuah *data point* terhadap tetangganya[23]. LOF menghitung anomali skor berdasarkan rasio *local density* suatu point terhadap *local density* dari k-nearest neighbors.

Konsep kunci dalam LOF adalah *local reachability density* (LRD) dan *local outlier factor*. LRD mengukur kepadatan lokal suatu point berdasarkan jarak *reachability* ke tetangganya, sedangkan LOF membandingkan LRD suatu point dengan rata-rata LRD tetangganya. Point dengan LOF score tinggi dianggap sebagai anomali karena memiliki kepadatan yang lebih rendah dibandingkan tetangganya. Keunggulan LOF adalah kemampuannya untuk mendeteksi local outlier yang mungkin tidak terdeteksi oleh metode global. Algoritma ini sangat efektif untuk data dengan kluster yang memiliki density berbeda-beda. Namun, LOF memiliki kompleksitas komputasional yang tinggi $O(n^2)$ dan sensitif terhadap parameter k dan metrik jarak yang digunakan. Penelitian ini menunjukkan bahwa LOF dapat diadaptasi untuk berbagai jenis data dan memberikan hasil yang baik pada dataset dengan struktur yang kompleks[24]. LOF dapat efektif untuk mendeteksi record yang memiliki pola atribut yang tidak konsisten dengan mayoritas data.

2.6 *One-Class Support Vector Machine* (OC-SVM)

One-Class Support Vector Machine merupakan adaptasi dari *Support Vector Machine* untuk masalah deteksi anomali dengan konsep *unsupervised*[25]. OC-SVM bekerja dengan mempelajari *decision boundary* yang memisahkan data normal dari anomali dalam *feature space* berdimensi tinggi. Algoritma ini menggunakan kernel trick untuk memetakan data ke *feature space* berdimensi tinggi dan mencari hyperplane yang memaksimalkan margin dari origin. Data yang berada di sisi yang salah dari *hyperplane* atau di luar margin diklasifikasikan sebagai anomali. Parameter utama OC-SVM adalah nu (ν) yang menentukan fraksi anomali yang diharapkan dalam dataset dan gamma yang mengontrol kompleksitas model.

Keunggulan OC-SVM adalah kemampuannya untuk menangani data berdimensi tinggi dan *non-linear pattern* melalui penggunaan kernel function. Algoritma ini juga memiliki theoretical foundation yang kuat berdasarkan statistical learning theory. Namun, OC-SVM memiliki kompleksitas komputasional yang tinggi dan sensitif terhadap pemilihan parameter dan kernel function. Penelitian ini menunjukkan bahwa OC-SVM efektif untuk deteksi anomali pada data dengan *boundary* yang kompleks[26]. OC-SVM dapat digunakan untuk mengidentifikasi *record* yang memiliki kombinasi atribut yang tidak biasa atau melanggar *business rules* implisit.

2.7 Evaluasi Performa Deteksi Anomali

Evaluasi performance algoritma deteksi anomali memiliki tantangan khusus karena nature dari masalah yang imbalanced dan seringkali tidak memiliki ground truth yang jelas. Metrik evaluasi tradisional seperti accuracy kurang sesuai karena bias terhadap *majority class*. Oleh karena itu, diperlukan metrik yang lebih sensitif terhadap *minority class* (anomali).
Precision, recall, dan F1-score merupakan metrik yang umum digunakan untuk evaluasi deteksi anomali. Precision mengukur proporsi anomali yang terdeteksi dengan benar dari semua deteksi, sedangkan recall mengukur proporsi anomali yang berhasil dideteksi dari total anomali yang sebenarnya ada. F1-score memberikan harmonic mean dari precision dan recall.

Area Under the ROC Curve (AUC-ROC) dan Area Under the Precision-Recall Curve (AUC-PR) juga sering digunakan untuk evaluasi. AUC-PR umumnya lebih informatif untuk imbalanced dataset karena lebih fokus pada performance terhadap minority class. Selain itu, metrik seperti *Average Precision* (AP) dan *Matthews Correlation Coefficient* (MCC) dapat memberikan insight tambahan tentang quality deteksi anomali.

2.8 Metrik Model Evaluasi

Model evaluasi yaitu confusion matrix merupakan sebuah tabel yang merangkum hasil prediksi dari sebuah model klasifikasi. Tabel ini berfungsi untuk membandingkan label actual dengan label prediksi, sebagai berikut:

Tabel 1. Deskripsi Metrik

Nama	Prediksi	Prediksi
Anomali	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)/

Penjelasan pada tabel yang diatas terbagi jadi beberapa komponen:
TP: model memprediksi dengan benar sebagai anomali.
FN: model salah memprediksi sebagai data normal.
FP: model salah prediksi sebagai data normal.
TN: model prediksi dengan benar sebagai normal.

Berdasarkan komponen di atas memiliki empat, maka dapat menurunkan metrik evaluasi yang lebih bermakna:

1. Precision (presisi)

   Precision merupakan sebuah metrik untuk mengukur tingkat ketepatan atau akurasi dari prediksi positif yang dibuat oleh model. Metrik ini penting untuk memahami seberapa andal model dalam memprediksi anomali. Nilai precision yang tinggi menunjukkan bahwa model memiliki tingkan False Positive yang rendah.

$$Precision = \frac{TP}{TP + FP}$$

2. Recall

   Recall merupakan sebuah metrik untuk mengukur kemampuan model dalam mengidentifikasi semua kejadian positif pada dataset. metrik ini sangat krusial dalam mengevaluasi seberapa komprehensif mendeteksi yang dilakukan. nilai recall yang tinggi menunjukkan bahwa model memiliki tingkat kegagalan dalam mendeteksi False Negative yang rendah.

$$Recall = \frac{TP}{TP + FN}$$

3. F1-Score

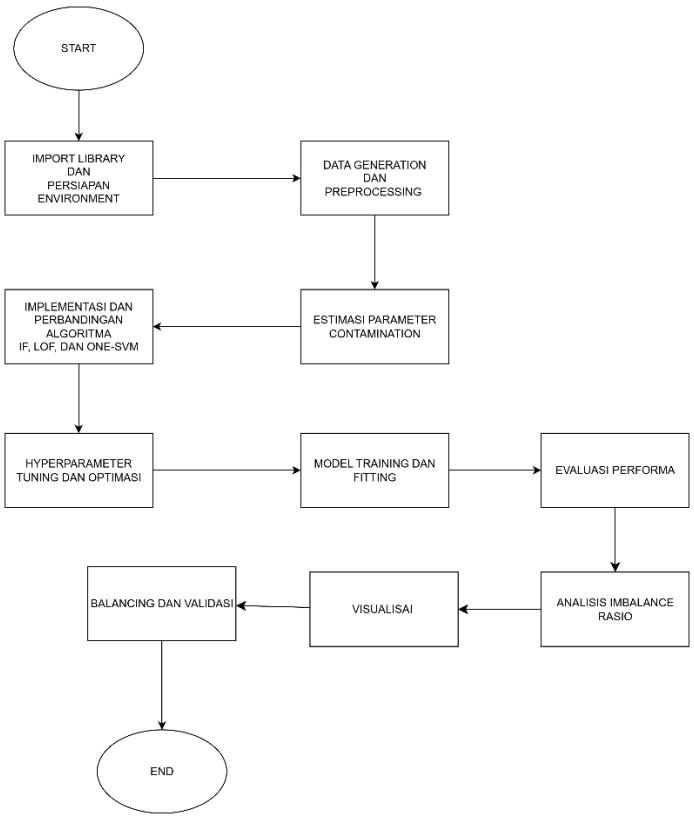
   F1-Score merupakan sebuah metrik yang menyajikan keseimbangan antara precision dan recall. metrik ini rata-rata harmonik dari kedua metrik menjadikan sebagai pilihan ideal dalam evaluasi kinerja suatu model pada dataset yang tidak seimbang.

$$F1 - Score = \frac{Precision * Recall}{Precision + Recall}$$

F1-Score penting karena ada pertukran antara precision dan recall. metrik ini dapat memberikan satu angka tunggal yang merangkum kinerja suatu model secara kesulurahan dengan mempertimbangkan kedua aspek. nilai F1-Score yang tinggi menunjukkan bahwa suatu model memiliki keseimbangan yang baik antara ketepatan prediksi dan kelengakapan deteksi.

3. METODE PENELITIAN

Penelitian ini menggunakan alur kerja dengan terstruktur untuk memastikan proses analisis data secara sistematis, iteratif, dan terdokumendasi. Alur kerja metodologi yang digunakan adalah *Unsupervised Machine Learning*, model ini mempelajari dari pola tersembunyi dalam dataset tanpa memerlukan labl atau target variabel yang telah ditentukan sebelumnya. Pendekatan ini digunakan bertujuan untuk mengidentifikasi anomali atau outlier dalam dataset berdasarkan pola yang dipelajari secara otomatis dari data, sehingga dapat mengklasifikasi observasi ke dalam kategorti "normal", "anomali" tanpa supervisi. Ada beberapa tahapan metodologi penelitian yang dilakukan.



Gambar 1. Alur yang digunakan di penelitian ini

3.1 Tahapan-Tahapan Diagram Alur Metodologi Penelitian

Ada beberapa tahapan Metode Penelitian yang dilakukan sebagai serangkaian yang saling berhubungan dengan menyesuaikan dengan Alur Metodologi Penelitian diatas, yaitu:

A. Persiapan awal dan pengolahan data

Pada tahapan ini diawali dengan menyiapkan bahan seperti import library dan persiapan environtment untuk keperluan analisis. Persiapan environtment merupakan proses awal sebelum bereksperimen untuk analisis llebih lanjut. Hal ini memastikan seluruh kebutuhan komputasi tersedia.

Setelah persiapan yang diperlukan untuk kebutuhan telah diselesaikan tahap selanjutnya yaitu Data generation dan preprocessing. Pada tahapan ini diperlukan data dummy dari organisasi atau perusaah di Bandung dengan struktur *nested array* yang sudak dikumpulkan. Proses selanjutya yaitu melakukan preprocessing untuk mempersiapkan data. Data yang disiapkan akan dilakukan eksplorasi untuk memahami karakteristik, distribusi, dan pola dalam dataset. Analasisi EDA penting supaya pada tahapan Analisis rasio ketidakseimbangan dapat memahami proporsi antara data normal dan data anomali. Proporsi kedua data akan ditentukan secara strategi dalam meangani data supaya menghasilkan metrik evaluasi yang tepat. Metrik evaluasi yang akan dilakukan untuk memastikan kualitas data sebagai fondasi performa model.

B. Penegembangan dan Pelatihan Model

Pada tahapan ini dilakukan proses analisis lebih lanjut dengan menggunakan algoritma *unsupervised machine learning* untuk deteksi anomali. Setiap Algoritma menyesuaikan dengan karakteristik data dan pola yang di dalam dataset[14]. Penelitian ini menggunakan tiga algoritma untuk mendeteksi anomali, yaitu isolation forest, local outlier factor (LOF) dan one-class SVM. Ketiga algoritma akan pemrosesan konfigurasi parameter berdasarkan dari training model tanpa menggunakan label target. Model akan mempelajari distribusi normal data secara otomatis dan mengidentifikasi yang menyimpang dari pola mayoritas sebagai anomali berdasrkan skor anomali yang dihitung oleh algoritma.

Selain itu, pada penelitian ini dilakukan pemodelan dengan menggunakan algoritma *unsupervised machine learning* untuk mendeteksi anomali. Algoritma yang kan digunakan ada tiga, yaitu Isolation Forest (IF), Local Outlier Factor (LOF), dan One-Class SVM. Ketiga algoritma ini bertujuan untuk membandigkan performa dalam mengidentifikasi yang menyimpang dari mayoritas.

Sebelum melakukan pelatihan data perlu dengan teknik yang efektif. Teknik yang digunakan yaitu estimasi parameter contamination untuk memperkirakan proporsi anomali dalam. Selanjutkan proses hyperparameter tuning dan optimasi dilakukan untuk menemukan kombinasi parameter terbaik bagi setiap algoritma. Kombinasi parameter ada beberapa Teknik untuk dapat diterapkan dengan validasi silang (cross-validation) supaya memastikan model tidak overfitting dan dapat digeneralisasi[15]. Dengan proses hyperparameter optimal telah dilakukan dengan hasil yang sesuai akan dilanjutkan ke tahapan model training dan fitting untuk melatih dengan menggunakan dataset tanpa label target. Tahapan hperparameter tuning merupakan proses optimal parameter model untuk mencapai performa terbaik. Setiap algoritma akan mempelajari pola dan struktur intrinsic dari data normal untuk dapat membedakan antara normal dan anomali. Proses fitting dipantau bertujuan untuk memastikan konvergensi algoritma dan ekstraksi pola yang representatif.

C. Validasi, Evaluasi dan Visualisasi

Pada tahapan ini dilakukan pengukuran kinerja dan penyajian hasil secara komprehensif. Penyajian yang dilakukan dari proses analisis rasio ketidakseimbangan untuk identifikasi adanya ketidakseimbangan kelas yang signifikan. Analisis rasio menggunakan beberapa metode dengan mencakup oversampling (SMOTE) untuk memperbanyak sampel kelas minoritas (anomaly) atau undersampling untuk mengurangi sampel kelas mayoritas (normal). Analisis ini bertujuan untuk meningkatkan kemampuan model dalam mendeteksi anomali tanpa mengorbankan performa secara keseluruhan, serta memvalidasi pendekatan ini supaya dapat memberikan hasil yang terbaik. Nilai dari validasi akan dilakukan evaluasi performa untuk mengukur kualitas dan validitas hasil deteksi anomali. Karena penelitian ini menggunakan pendekatan unsupervised learning, metrik evaluasi yang digunakan tidak memerlukan label kebenaran (ground truth). Hal ini akan menampilkan metrik yang relevan, sebagai berikut:

a. Silhouette Score: mengukur seberapa mirip dengan kluster yang dibandingkan dengan kluster lain.
b. Davies-Bouldin Index: mengevaluasi kualitas klustering berdasrkan rasio antara jarak dalam kluster dengan jarak antar kluster.
c. Calinski-Harabasz Index: mengukur rasio antara varians antar kluster dengan varians di dalam kluster.

Metrik yang diatas bertujuan untuk membantu menilai model berdasrkan struktur data yang terbentuk.

Visualisasi merupakan hasil akhir dari proses analisis dan evaluasi yang telah dilakukan. Pada tahapan ini mencakup penyajian grafik distribusi data, plot yang menunjukkan kumpulan titik data yang terdeteksi sebagai anomali oleh setiap model. Kumpulan titik akan dibandingkan supaya mengetahui informasi performa dari ketiga algoritma secara berdampingan. Hasil visual menjadi landasan utama untuk menarik, sehingga akan merumuskan rekomendasi dari penelitian.

Berdasarkan dari keseluruhan metodologi yang telah dirancang bertujuan untuk memastikan proses deteksi anomali dilakukan secara sistematis dan menghasilkan model yang robust dan dapat diandalkan untuk penelitian selanjutnya. Setiap tahapan dilakukan sangat berkaitan dan berkontribusi terhadap kualitas hasil akhir penelitian.

4. PEMBAHASAN

Penelitian ini diawali dengan tahapan memuat dan menganalisis dataset dengan format csv untuk dilakukan preprocessing supaya dapat menangani nilai yang hilang (missing values) dengan imputasi median dan melakukan pensakalaan fitur menggunakan RobustScaler untuk meminimalisir pengaruh outlier. Sebelum menerapkan algoritma deteksi anomaly untuk estimasi proporsi anomaly dalam data menggunakan beberapa metode statistic. Berikut tahapan untuk menentukan parameter contamination untuk model unsupervised, sehingga menghasilkan estimasi disajikan pada table di bawah.

Tabel 2. Hasil Estimasi Proporsi Anomali

Metode	Jumlah	Persentase	Contamination
IQR	1053	10.5	0.105
Z-Score	585	5.9	0.59
Modified Z-Score	734	7.3	0.73
Rata-rata estimasi	-	-	0.85

Berdasrkan dari analisis pada tabel 2 merupakan nilai contamination sebesar 0.15 di rekomendasikan sebagai baseline yang representatif. Nilai tersebut digunakan untuk membuat label ground truth dengan menggunakan algoritma Isolation Forest sebagai acuan standar suoaya dapat menentukan berdasarkan kelas “normal“ dan “anomali“.

Proses penentuan dari setiap baris data (record) dalam dataset akan digunakan 2000 baris data untuk mengidentifikasi pola data. Identifikasi pola Degnan cara oversampling atau undersampling supaya, sebagai berikut sebagai parameter data sample.

a. Sekitar 300 baris data sebagai anomali.
b. Sekitar 1700 baris data sebagai norma

Pembahasan merupakan bagian terpenting dari naskah publikasi. harus mengandung hasil-hasil simulasi atau pengukuran sebagai validasi metode. Pembahasan dapat berupa tabel hasil, narasi yang didapat dari perhitungan suatu rumus maupun prosentase dari grafik perhitungan.

4.1 Data Generation dan Preprocessing

Pada tahapan ini dilakukan persiapan dataset yang akan digunakan untuk melatih dan mengevaluasi model deteksi anomali. Dataset ini dibuat secara sintetis (data generation) untuk merepresentasikan skenario data anomali muncul dalam proporsi yang lebih kecil dibandingkan data normal. Berikut hasil dari preprocessing data generation

Tabel 3. Hasil Data Generation dan Preprocessing

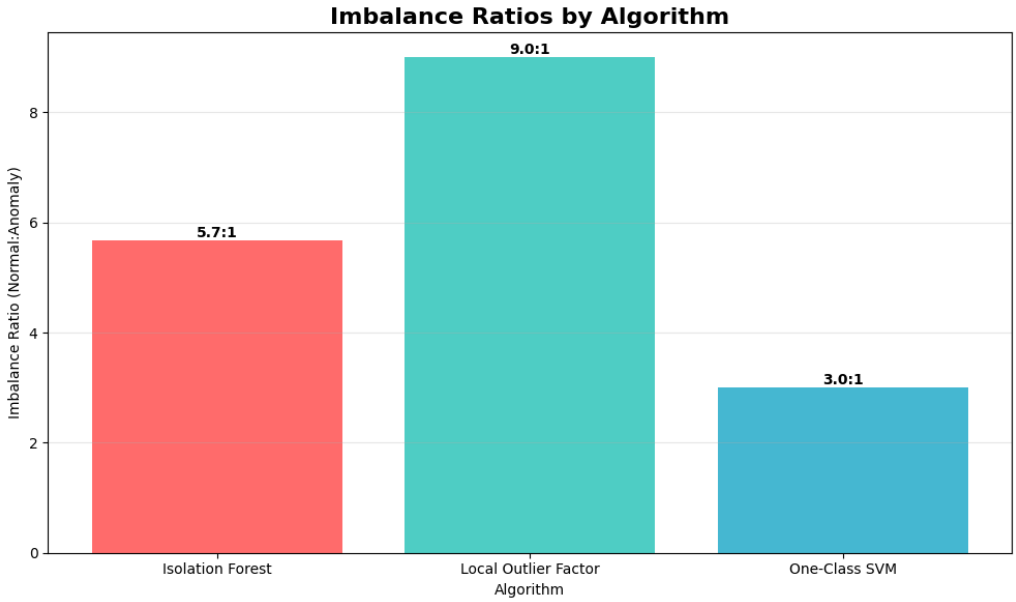
Total Sampel	10000
Jumlah fitur	5
Data normal	9000 (90%)
Data anomali	1000(10%)

Amati tabel 4. Bahwa dataset yang dihasilkan terdiri dari 10000 total sampel dengan 5 fitur numerik. Komposisi data ini dirancang tidak seimbang karena terdiri dari 9000 sampel (90%) sebagai data normal dan 1000 sampel(10%) sebagai data anomali. Rasio 10% bertujuan untuk mensimulasikan kondisi nyata anomali pada kejadian langka supaya dideteksi.

4.2 Analisis Rasio Ketidakseimbangan data

Hasil dengan menggunakan Algoritma Isolation Forest digunakan sebagai target untuk melatih model klasifikasi. Karena dari rasio menunjukkan bahwa eksperimen yang dilakukan dengan perlu teknik penyeimbangan data untuk menghasilkan metrik yang terbaik. Teknik penyeimbangan digunakan, sebagai berikut

a. Radom Oversampling
b. Random Undersampling
c. SMOTE



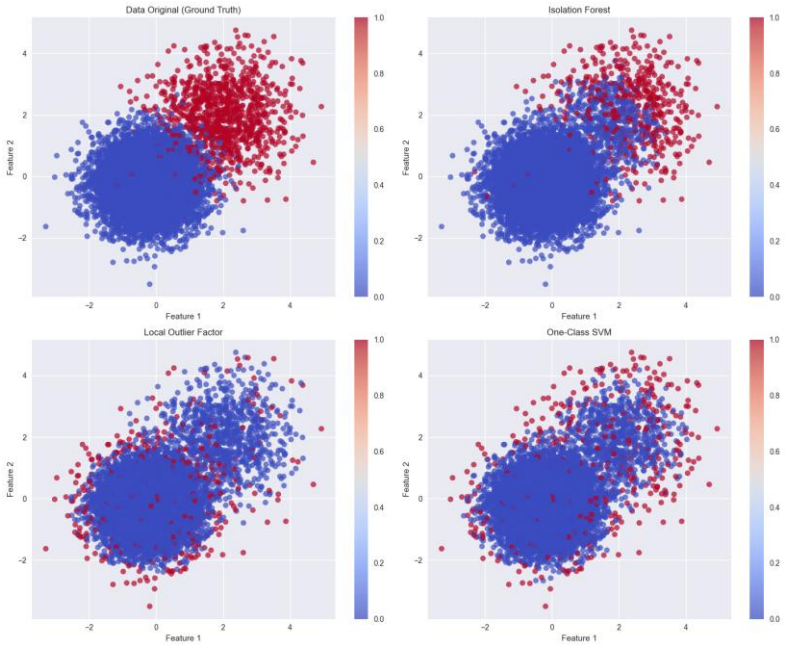
Gambar 2. Rasio Ketidakseimbangan Data Hasil Deteksi Anomali

Amati gambar 3 bahwa menunjukkan rasio ketidakseimbangan data yang dihasilkan dari setiap algoritma deteksi anomali. Setiap algoritma deteksi anomali terdapat variasi signifikan dalam Tingkat ketidakseimbangan yang dihasilkan oleh masing-masing algoritma, sebagai berikut:

a. Local Outlier Factor: menunjukkan hasil Tingkat ketidakseimbangan paling ekstrem dengan rasio 9.0:1 yang artinya terdapat 9 data normal untuk setiap 1 data anomali. Ketidakseimbangan yang tinggi dengan performa LOF yang rendah dalam mendeteksi anomali. LOF menghasilkan nilai F1-Score 0.221 pada evaluasi sebelumnya, sehingga perlu analisis rasio yang ekstrem untuk mengindikasikan bahwa LOF cenderung konservatif dalam mengklasifikasikan data sebagai anomali.

b. Isolation Forest: menunjukkan Tingkat ketidakseimbangan moderat dengan rasio 5.7:1. Hal ini terdapat ketidakseimbangan yang substansial, rasio ini relative lebih seimbang dibandingkan LOF sehingga performa dari algoritma ini dianggap superior. Karena algoritma ini mampu mengidentifikasi lebih banyak kasus anomaly tanpa terlalu konservatif.
c. One-Class SVM: menunjukkan hasil Tingkat ketidakseimbangan paling rendah dengan rasio 3.0:1 bahwa algoritma ini liberal dalam mengklasifikasikan data sebagai anomali. Rasio yang relatif seimbang karena kodisi nilai recall tinggi (96.1%) yang diamati sebelumnya, Namun nilai precision rendah (32.7%) karena cenderung over-detection.

Jadi analisis rasio ketidakseimbangan memiliki perbedaan implikasi penting untuk tahap klasifikasi, karena tingkat ketidakseimbangan yang berbeda akan memerlukan strategi penyeimbangan yang berbeda untuk mengoptimalkan performa model klasifikasi.



Gambar 3. Visualisasi kluster penyebaran data

Amati gambar 3 bahwa menyajikan isualisasi sebaran data dalam bentuk scatter plot melalui proses deteksi anomaly menggunakan algoritma isolation forest. Dalam plot tersebut terdapat dua titik yang memiliki warna yang berbda untuk klasifikasikan data sebagai “normal” (titik berwarna biru) dan “anomali” (ditandai titik berwarna merah). Hasil grafik menunjukkan bahwa data normal membentuk sebuah kluster normal yang padat bertujuan untuk mengindikasikan kesamaan karakteristik. Dari kumpulan titik data yang menonjol adalah data normal sehingga di bagian luar kluster normal cenderung tidak mengalami masalah. Pola ini memvisualisasikan utnuk mengkonfirmasi efektivitas algoritma, karakteristik anomali, dan rasio ketidakseimbangan.

Visualisasi ini memberikan pemahaman intuitif tentang perbedaan model antara data normal dan anomali di dalam ruang fitur.

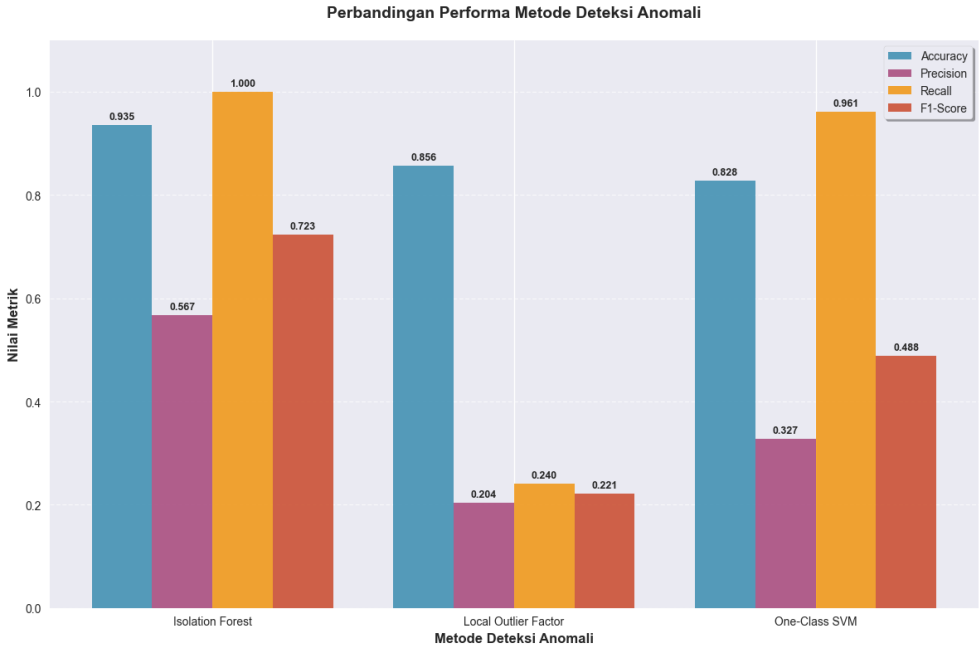
4.3 Kinerja Algoritma Deteksi Anomali

Penggunaan algoritma adalah tiga yaitu Isolation Forest, Local Outlier Factor (LOF) dan One-Class SVM diterapkan pada data yang telah di proses. Kinerja setiap algoritma untuk dievaluasi mempunyai kelebihan dan kekurangan tersendiri. Metrik evaluasi disebut confusion matrix yang terdiri dari beberapa nilai, sebagai berikut.

Tabel 4. Perbandingan Kinerja Algoritma Deteksi Anomali

Metode	Accuracy	Precision	Recall	F1-Score
Isolation Forest	0.935	0.567	1.00	0.723
Local Oulier Factor	0.856	0.204	0.240	0.221
One-Class SVM	0.828	0.327	0.961	0.488

Amati pada tabel 3 bahwa menunjukan kinerja terbaik dengan F1-Score tertinggi sebesar 0.723.



Gambar 4. Visualisasi perbandingan kinerja algoritma

Amati gambar 4. Menyajikan visual perbandingan kinerja ketiga algoritma deteksi anomali melalui diagram bar yang merepresentasikan metrik evaluasi. Visualisasi ini menunjukkan bahwa performa paling unggul secara keseluruhan adalah *Isolation Forest*.

5. KESIMPULAN

Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa algoritma isolation forest menunjukkan kinerja terbaik untuk deteksi anomaly dalam proses ETL dibandignkan dengan algoritma lainnya. Isolation forest mencapai performa superior dengan nilai F1-Score tertinggi sebesar 0.723, akurasi 0.935, presisi 0.567 dan recall 1.00. Hal ini bahwa kemampuan mengidentifikasi seluruh anomaly dengan menjaga akurasi yang tinggi, sebalinknya LOF memiliki performa trendah karena memiliki nilai F1-Score 0.221 sehingga menandakan kesulitan dalam mendeteksi anomaly yang bersifat konservatif. One-Class SVM menunjukkan hasil moderat dengan nilai F1-Score 0.488 dan recall yang tinggi (0.961), namun cenderung mengalami over detection yang menyebabkan presisi rendah.

Penelitian ini menyoroti bahwa algoritma dapat menghasilkan Tingkat ketidakseimbangan data yang berbeda-beda dan perlu strategi penyeimbangan data yang spesifik untuk tahap selnajutnya. Visualisasi data memperkuat bahwa pola deteksi anomaly oleh isolation forest konsisten dengan ground truth.

DAFTAR PUSTAKA

[1] Q. Yang and Y. Tang, "Big Data-based Human Resource Performance Evaluation Model Using Bayesian Network of Deep Learning," *Appl. Artif. Intell.*, vol. 37, no. 1, p. 2198897, Dec. 2023, doi: 10.1080/08839514.2023.2198897.

[2] J. Ye, "Modeling of performance evaluation of educational information based on big data deep learning and cloud platform," *J. Intell. Fuzzy Syst.*, vol. 38, no. 6, pp. 7155–7165, June 2020, doi: 10.3233/JIFS-179793.

[3] F. F. Hasan and M. S. A. Bakar, "Data Transformation from SQL to NoSQL MongoDB Based on R Programming Language," in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2021, pp. 399–403. doi: 10.1109/ISMSIT52890.2021.9604548.

[4] W. Bulowski *et al.*, "Optimization of the ETL titanium dioxide layer for inorganic perovskite solar cells," *J. Mater. Sci.*, vol. 59, no. 17, pp. 7283–7298, May 2024, doi: 10.1007/s10853-024-09581-w.

[5] L. Dinesh and K. G. Devi, "An efficient hybrid optimization of ETL process in data warehouse of cloud architecture," *J. Cloud Comput.*, vol. 13, no. 1, p. 12, Jan. 2024, doi: 10.1186/s13677-023-00571-y.

[6] G. Hannák, G. Horváth, A. Kádár, and M. D. Szalai, "BILATERAL-WEIGHTED Online Adaptive Isolation Forest for anomaly detection in streaming data," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 16, no. 3, pp. 215–223, June 2023, doi: 10.1002/sam.11612.

[7] S. Vats, B. B. Sagar, K. Singh, A. Ahmadian, and B. A. Pansera, "Performance Evaluation of an Independent Time Optimized Infrastructure for Big Data Analytics that Maintains Symmetry," *Symmetry*, vol. 12, no. 8, p. 1274, Aug. 2020, doi: 10.3390/sym12081274.

[8] H. Xiang, J. Wang, K. Ramamohanarao, Z. Salcic, W. Dou, and X. Zhang, "Isolation Forest Based Anomaly Detection Framework on Non-IID Data," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 31–40, May 2021, doi: 10.1109/MIS.2021.3057914.

[9] A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, S. Pérez-Díaz, and J. J. Martínez-Herráiz, "Applied Machine Learning to Anomaly Detection in Enterprise Purchase Processes: A Hybrid Approach Using Clustering and Isolation Forest," *Information*, vol. 16, no. 3, p. 177, Feb. 2025, doi: 10.3390/info16030177.

[10] M. Nalini, B. Yamini, C. Ambhika, and R. Siva Subramanian, "Enhancing early attack detection: novel hybrid density-based isolation forest for improved anomaly detection," *Int. J. Mach. Learn. Cybern.*, Nov. 2024, doi: 10.1007/s13042-024-02460-5.

[11] N. Biswas, A. S. Mondal, A. Kusumastuti, S. Saha, and K. C. Mondal, "Automated credit assessment framework using ETL process and machine learning," *Innov. Syst. Softw. Eng.*, vol. 21, no. 1, pp. 257–270, Mar. 2025, doi: 10.1007/s11334-022-00522-x.

[12] H. Azgomi and M. K. Sohrabi, "A novel coral reefs optimization algorithm for materialized view selection in data warehouse environments," *Appl. Intell.*, vol. 49, no. 11, pp. 3965–3989, Nov. 2019, doi: 10.1007/s10489-019-01481-w.

[13] H. Fu, "Optimization Study of Multidimensional Big Data Matrix Model in Enterprise Performance Evaluation System," *Wirel. Commun. Mob. Comput.*, vol. 2021, no. 1, p. 4351944, Jan. 2021, doi: 10.1155/2021/4351944.

[14] T. Mendes, P. J. S. Cardoso, J. Monteiro, and J. Raposo, "Anomaly Detection of Consumption in Hotel Units: A Case Study Comparing Isolation Forest and Variational Autoencoder Algorithms," *Appl. Sci.*, vol. 13, no. 1, p. 314, Dec. 2022, doi: 10.3390/app13010314.

[15] D. Ribeiro, L. M. Matos, G. Moreira, A. Pilastri, and P. Cortez, "Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection," *Computers*, vol. 11, no. 4, p. 54, Apr. 2022, doi: 10.3390/computers11040054.