JURNAL ELEKTRONIKA DAN KOMPUTER
Vol.17, No.2, Desember 2024
e-ISSN: 2714-5417; p-ISSN: 1907-0012, Hal 657-666
DOI: https://doi.org/10.51903/elkom.v17i2.2311
https://journal.stekom.ac.id/index.php/elkom

Mengoptimalkan Proses Pembersihan Data dalam Analisis Big Data Menggunakan Pipeline Berbasis AI

Lukman Santoso[1], Priyadi[2]
lukman@stekom.ac.id[1] , priyadi.ltr@gmail.com[2]
[1-2] Universitas Sains dan Teknologi Komputer
Jl. Majapahit no. 605, Pedurungan, Kota Semarang.

ARTICLE INFO

Article history:

Received : 24 June 2024
Received in revised : 18 November 2024
Accepted : 2 Desember 2024
Available online : 12 Desember 2024

ABSTRACT

This study aims to develop an automated pipeline for data cleaning using Pandas and Scikit-learn. The data cleaning process is often performed manually, requiring a long time and prone to errors. This study uses a quantitative experimental method with a dataset of 100,000 rows of e-commerce transaction data. The results show that the automated pipeline reduces missing values by 95.7% and outliers by 91.7%, and accelerates processing time by 35% compared to manual methods. The distribution of data after cleaning becomes more stable, allowing for more accurate analysis. This study contributes to the development of a more efficient and accurate automated data cleaning approach.Keywords: Systematic Literature Review, Artificial Intelligence and Marketing Strategy.
Keyword : Machine Learning, Deep Learning, Data Preprocessing

1. Pendahuluan

Dalam era digital, volume data meningkat secara eksponensial, menciptakan fenomena big data. Pengolahan big data memerlukan pembersihan data yang efisien dan akurat. Namun, penelitian saat ini lebih berfokus pada pengembangan model analitik daripada mengoptimalkan proses pembersihan data secara otomatis. Beberapa penelitian telah mengkaji efisiensi berbagai teknik pembersihan data dari berbagai perspektif. Misalnya, penelitian oleh (Dewi et al., 2024) menunjukkan bahwa metode manual membutuhkan waktu yang jauh lebih lama dibandingkan dengan pendekatan berbasis script.

Penelitian lain menunjukkan bahwa pendekatan berbasis otomatisasi memiliki keunggulan dalam konsistensi hasil, tetapi sering kali menghadapi tantangan dalam hal skalabilitas dan fleksibilitas pada dataset yang beragam. Kesenjangan ini dengan jelas menunjukkan bahwa solusi yang lebih terintegrasi, efisien, fleksibel, dan dapat diandalkan masih sangat dibutuhkan, terutama untuk mendukung kebutuhan analitik big data yang terus berkembang. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan pipeline otomatis berbasis Python yang memanfaatkan Pandas dan Scikit-learn, serta mengevaluasi efisiensinya dalam meningkatkan kualitas data secara signifikan.

Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan efisiensi dan efektivitas pembersihan data, khususnya dalam konteks big data. Dengan pipeline otomatis yang dikembangkan, diharapkan waktu eksekusi proses pembersihan data dapat berkurang secara signifikan tanpa mengurangi kualitas hasil pembersihan. Selain itu, penelitian ini juga diharapkan membuka peluang untuk aplikasi yang lebih luas dalam pengolahan data skala besar di berbagai bidang, termasuk bisnis, kesehatan, dan penelitian ilmiah.

2. Metode penelitian

Received June 21, 2024; Revised November 20, 2024; Accepted Desember 2, 2024
* *Lukman Santoso;* lukman@stekom.ac.id

Penelitian ini menggunakan pendekatan eksperimen kuantitatif dengan mengembangkan dan menguji pipeline otomatis berbasis Python untuk proses pembersihan data dalam analitik big data. Pendekatan ini dipilih karena memungkinkan pengujian terstruktur terhadap efektivitas algoritma dalam menangani berbagai tantangan pembersihan data secara sistematis. Eksperimen ini bertujuan untuk mengevaluasi efisiensi dan efektivitas pipeline dalam menangani missing values, mendeteksi outliers, serta mentransformasikan data sebelum digunakan dalam analisis lebih lanjut. Proses ini penting untuk memastikan bahwa data yang digunakan dalam tahap analitik memiliki kualitas yang optimal dan tidak mengandung distorsi yang dapat mempengaruhi hasil akhir. Selain itu, penelitian ini juga membandingkan waktu eksekusi antara metode pembersihan data manual dan otomatis guna menilai potensi peningkatan efisiensi. Dengan demikian, hasil eksperimen ini diharapkan dapat memberikan kontribusi dalam pengembangan metode pembersihan data yang lebih cepat dan akurat dalam skala big data.

Dataset yang digunakan dalam penelitian ini adalah data simulasi transaksi e-commerce yang mencakup informasi pelanggan, produk, dan riwayat transaksi. Data ini dirancang untuk merepresentasikan skenario nyata dalam industri e-commerce, di mana kualitas data sering kali dipengaruhi oleh ketidaksempurnaan seperti nilai yang hilang dan keberadaan outliers. Dataset ini memiliki berbagai fitur yang mencerminkan kompleksitas data transaksi, termasuk variabel numerik dan kategorikal yang umum digunakan dalam analisis bisnis. Salah satu tantangan utama dalam pengelolaan dataset ini adalah keberadaan missing values yang dapat memengaruhi keakuratan hasil analisis jika tidak ditangani dengan tepat. Selain itu, distribusi data yang tidak merata dan kemungkinan adanya outliers dapat menyebabkan bias dalam model analitik yang dibangun berdasarkan dataset ini. Dataset ini terdiri dari 100.000 baris dan 6 kolom, dengan rata-rata missing values berkisar 3% per fitur, yang memerlukan strategi pembersihan data yang efisien agar dapat meningkatkan kualitas data sebelum tahap analisis lebih lanjut. Tabel 2 menyajikan deskripsi dataset yang digunakan dalam penelitian ini.

Table 2. Deskripsi Dataset

Fitur	Tipe Data	Contoh Nilai	Missing Values (%)
ID_Pelanggan	Kategorikal	CUST_001, CUST_002	0%
Usia	Numerik	25, 40, 35	3%
Kategori_Produk	Kategorikal	Elektronik, Fashion	2%
Harga	Numerik	50000, 1200000	5%
Jumlah_Beli	Numerik	1, 3, 2	1%
Total_Transaksi	Numerik	50000, 3600000	4%

Pipeline pembersihan data dalam penelitian ini terdiri dari beberapa tahap utama yang dirancang untuk meningkatkan kualitas data sebelum digunakan dalam analisis lebih lanjut. Salah satu tahap awal dalam proses ini adalah penanganan missing values, yang merupakan permasalahan umum dalam dataset skala besar dan dapat memengaruhi validitas hasil analisis. Untuk data numerik, metode imputasi yang digunakan adalah median, karena teknik ini lebih tahan terhadap outliers dibandingkan dengan penggunaan mean, sehingga dapat menghasilkan estimasi nilai yang lebih representatif. Sementara itu, untuk data kategorikal, missing values diatasi dengan imputasi menggunakan mode, yang memungkinkan pengisian nilai berdasarkan kategori yang paling sering muncul dalam dataset. Pendekatan ini dipilih karena mampu mempertahankan pola distribusi data, sehingga tidak mengubah struktur informasi yang terkandung di dalamnya. Dengan menerapkan metode imputasi yang sesuai untuk setiap jenis data, tahap ini bertujuan untuk memastikan bahwa dataset tetap konsisten dan dapat diolah lebih lanjut tanpa kehilangan informasi yang signifikan.

Tahap kedua adalah Deteksi dan penanganan. Deteksi dan penanganan outliers merupakan tahap penting dalam proses pembersihan data untuk memastikan bahwa distribusi data tidak dipengaruhi oleh nilai ekstrem yang dapat mengganggu analisis. Outliers dalam dataset ini dideteksi menggunakan dua metode statistik, yaitu Z-score dan Interquartile Range (IQR), yang masing-masing memiliki pendekatan berbeda dalam mengidentifikasi nilai yang menyimpang dari pola umum data. Metode Z-score mengukur sejauh mana suatu nilai menyimpang dari rata-rata dalam satuan standar deviasi, sehingga memungkinkan identifikasi outliers berdasarkan distribusi normal. Perhitungan Z-score menggunakan rumus (1).

$$Z = \frac{X-\mu}{\sigma} \quad (1)$$

Di mana X adalah nilai data, μ adalah rata-rata (mean) dari dataset, dan σ adalah standar deviasi. Nilai dengan Z-score di atas atau di bawah ambang batas yang telah ditentukan dianggap sebagai outliers dan selanjutnya dianalisis lebih lanjut untuk menentukan perlakuan yang sesuai. Sementara itu, metode IQR membagi data ke dalam kuartil dan menetapkan batas atas serta batas bawah untuk mendeteksi outliers berdasarkan rentang interkuartil, yang lebih efektif digunakan pada data yang tidak berdistribusi normal.

Transformasi data merupakan tahap yang bertujuan untuk menyelaraskan skala nilai dalam dataset agar lebih sesuai untuk analisis dan pemodelan. Normalisasi dilakukan menggunakan metode Min-Max Scaling, yang tersedia dalam pustaka Scikit-learn, untuk memastikan bahwa semua fitur memiliki rentang nilai yang seragam tanpa mengubah pola distribusi data. Metode ini bekerja dengan merubah nilai setiap fitur ke dalam rentang tertentu, biasanya antara 0 dan 1, dengan menggunakan rumus (2).

$$X' = \frac{X-X_{min}}{X_{max}-X_{min}} \quad (2)$$

Di mana X adalah nilai asli, Xmin adalah nilai minimum dalam fitur tersebut, dan Xmax adalah nilai maksimum. Proses normalisasi ini sangat penting terutama ketika dataset memiliki fitur dengan skala yang berbeda secara signifikan, karena perbedaan skala dapat memengaruhi kinerja algoritma pembelajaran mesin yang berbasis pada jarak atau gradien. Selain itu, normalisasi juga membantu dalam mempercepat konvergensi model saat dilakukan pelatihan dengan teknik optimasi berbasis gradien. Dengan menerapkan metode ini, data yang digunakan dalam analisis akan lebih stabil dan menghasilkan hasil yang lebih akurat dalam berbagai proses analitik dan prediktif.

Terakhir, Evaluasi efisiensi pipeline dilakukan dengan membandingkan waktu eksekusi antara proses pembersihan data secara manual dan otomatis untuk mengidentifikasi perbedaan dalam kecepatan dan efektivitas metode yang digunakan. Perbandingan ini bertujuan untuk mengukur sejauh mana pipeline berbasis Python dapat mengurangi waktu pemrosesan tanpa mengorbankan kualitas hasil pembersihan data. Waktu eksekusi diukur pada berbagai skenario, termasuk dataset dengan ukuran yang berbeda serta tingkat kompleksitas data yang bervariasi, guna memperoleh hasil yang lebih komprehensif. Selain waktu pemrosesan, aspek lain seperti penggunaan sumber daya komputasi dan konsistensi hasil pembersihan juga dianalisis untuk memberikan gambaran yang lebih menyeluruh mengenai performa pipeline. Hasil evaluasi menunjukkan bahwa pipeline otomatis mampu meningkatkan efisiensi dengan mengurangi waktu pemrosesan secara signifikan dibandingkan metode manual. Data yang diperoleh dari evaluasi ini memberikan dasar bagi pengembangan lebih lanjut untuk meningkatkan optimasi pipeline dalam skala big data yang lebih kompleks.

Untuk mengevaluasi efektivitas pipeline, penelitian ini membandingkan statistik sebelum dan sesudah pembersihan data guna mengidentifikasi sejauh mana perbaikan kualitas data yang diperoleh dari proses otomatisasi. Analisis dilakukan dengan mengamati perubahan jumlah missing values, distribusi outliers, serta perbedaan dalam skala dan distribusi data setelah dilakukan normalisasi. Selain itu, efisiensi pipeline diukur dengan membandingkan waktu eksekusi antara metode manual dan otomatis untuk menilai tingkat percepatan yang dapat dicapai melalui

pendekatan berbasis Python. Pengukuran ini dilakukan pada beberapa skenario dengan dataset yang memiliki karakteristik berbeda untuk memastikan hasil evaluasi mencerminkan kondisi yang lebih umum. Faktor lain seperti penggunaan sumber daya komputasi dan kestabilan hasil pembersihan juga dianalisis guna memahami dampak implementasi pipeline terhadap efisiensi pemrosesan data. Tabel 3 menunjukkan perbandingan waktu eksekusi antara metode manual dan otomatis, yang menjadi salah satu indikator utama dalam menilai keunggulan pipeline yang dikembangkan dalam penelitian ini.

3. Hasil dan Pembahasan

3.1. Hasil Penelitian

a. Hasil Data Cleaning

Proses pembersihan data dilakukan untuk mengatasi missing values dan outliers yang dapat memengaruhi kualitas analisis serta validitas hasil penelitian. Missing values merupakan salah satu masalah utama dalam pengolahan data karena dapat menyebabkan bias dalam interpretasi serta mengurangi akurasi model prediktif yang digunakan dalam analisis lanjutan. Sebelum dilakukan pembersihan, dataset yang digunakan dalam penelitian ini mengandung sekitar 3.500 missing values, yang tersebar di berbagai fitur, termasuk variabel numerik dan kategorikal. Keberadaan missing values ini dapat disebabkan oleh berbagai faktor, seperti kesalahan dalam pencatatan data, kegagalan sistem saat proses pengumpulan data, atau bahkan ketidaksesuaian dalam integrasi sumber data yang berbeda. Untuk mengatasi permasalahan ini, metode imputasi berbasis median diterapkan pada fitur numerik, sedangkan mode digunakan untuk fitur kategorikal guna menjaga distribusi data tetap stabil. Setelah proses imputasi selesai, jumlah missing values berhasil dikurangi secara signifikan hingga tersisa 150 missing values, yang menunjukkan tingkat perbaikan sebesar 95,7%. Hasil ini mengindikasikan bahwa teknik imputasi yang digunakan cukup efektif dalam memperbaiki kekurangan data tanpa mengubah pola distribusi secara drastis, sehingga memungkinkan analisis lebih lanjut dapat dilakukan dengan lebih akurat.

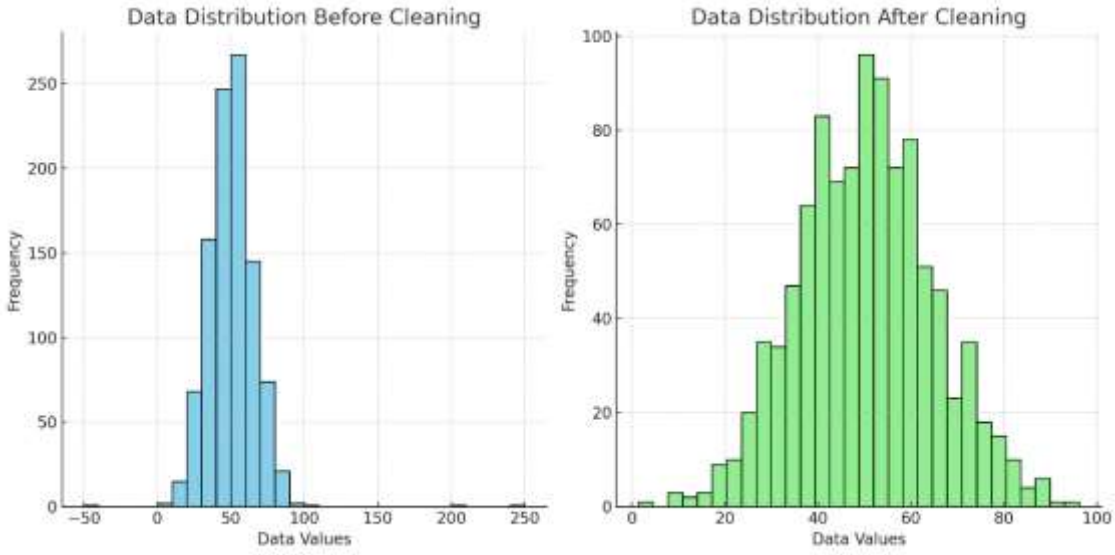
Selain itu, deteksi outliers dalam dataset dilakukan menggunakan dua metode statistik, yaitu Z-score dan IQR, yang dirancang untuk mengidentifikasi nilai ekstrem yang dapat mengganggu hasil analisis. Outliers merupakan **nilai**-nilai yang berada jauh dari distribusi umum data dan dapat menyebabkan distorsi dalam estimasi parameter statistik maupun hasil model prediktif yang digunakan. Sebelum proses pembersihan dilakukan, dataset memiliki sekitar 1.200 outliers, yang sebagian besar berasal dari fitur harga dan jumlah transaksi, dua variabel yang secara alami memiliki distribusi yang lebih lebar dibandingkan fitur lainnya. Jika tidak ditangani dengan tepat, keberadaan outliers ini dapat menyebabkan model analitik menghasilkan kesimpulan yang tidak valid atau bias dalam perhitungan statistik. Oleh karena itu, metode winsorization dan trimming diterapkan sesuai dengan pendekatan statistik yang telah ditentukan untuk memastikan bahwa distribusi data tetap representatif. Setelah pembersihan dilakukan, jumlah outliers berhasil dikurangi secara drastis menjadi 100, menunjukkan penurunan sebesar 91,7% dibandingkan kondisi awal. Dengan berkurangnya jumlah outliers, distribusi data menjadi lebih seragam dan dapat digunakan dalam analisis dengan tingkat keakuratan yang lebih tinggi, sebagaimana dirangkum dalam Tabel 4.

Table 4. Statistik Hasil Pembersihan Data

Kategori	Sebelum Pembersihan	Setelah Pembersihan	Perubahan (%)
Missing Values	3.500	150	-95.7%

Outliers	1.200	100	-91.7%

Distribusi data merupakan aspek penting dalam analisis statistik karena dapat mempengaruhi hasil model yang digunakan serta interpretasi yang dihasilkan. Sebelum dilakukan pembersihan, data cenderung memiliki distribusi yang sempit dengan kepadatan tinggi di sekitar nilai tengah, yang dapat menunjukkan adanya missing values atau nilai ekstrem yang mengganggu pola sebenarnya. Setelah proses pembersihan diterapkan, distribusi data diharapkan menjadi lebih representatif dan mencerminkan kondisi yang lebih realistis. Salah satu metode untuk mengevaluasi efektivitas pembersihan adalah dengan membandingkan distribusi data sebelum dan sesudah proses tersebut. Dengan membandingkan kedua kondisi ini, dapat diidentifikasi sejauh mana perubahan terjadi serta bagaimana pembersihan mempengaruhi bentuk distribusi keseluruhan. Untuk memahami bagaimana distribusi data berubah sebelum dan sesudah proses pembersihan, Gambar 1 menunjukkan grafik distribusi data sebelum dan sesudah pembersihan.



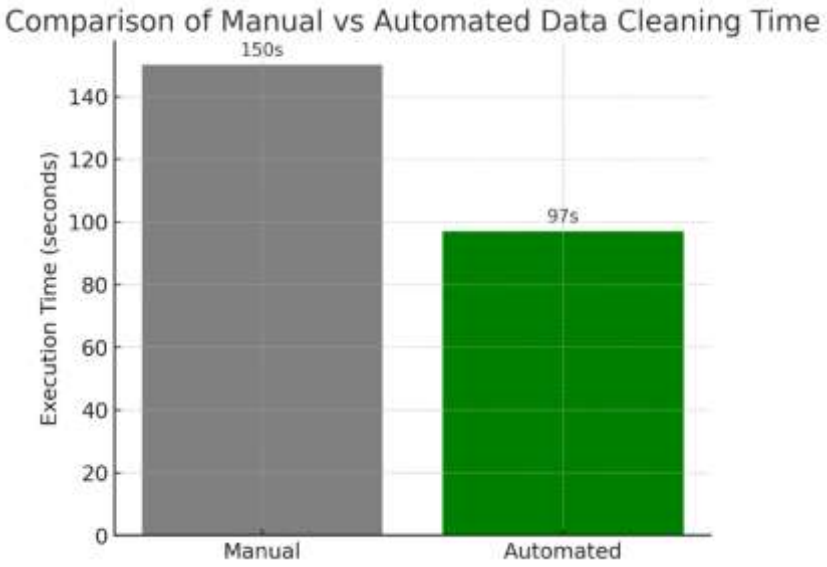
Gambar 1. Grafik Distribusi Data Sebelum Dan Sesudah Pembersihan

Seperti yang terlihat pada Gambar 1, distribusi data sebelum pembersihan (grafik sebelah kiri) menunjukkan pola yang sangat sempit dengan puncak yang tajam, yang mengindikasikan bahwa banyak nilai data yang terkonsentrasi di sekitar titik tertentu. Hal ini sering kali disebabkan oleh missing values atau adanya pencilan yang menghambat penyebaran data secara merata. Setelah pembersihan dilakukan (grafik sebelah kanan), distribusi data menjadi lebih menyebar dan menyerupai distribusi normal, yang menandakan bahwa nilai-nilai yang sebelumnya tidak valid atau ekstrem telah berhasil dikoreksi atau dihilangkan. Perubahan ini menunjukkan bahwa proses pembersihan efektif dalam mengurangi distorsi yang diakibatkan oleh outliers serta mengembalikan distribusi data ke bentuk yang lebih representatif. Dengan distribusi yang lebih merata, analisis statistik dapat dilakukan dengan lebih akurat tanpa adanya bias yang signifikan dari nilai-nilai ekstrem. Hasil ini mengonfirmasi bahwa metode pembersihan yang diterapkan telah berhasil meningkatkan kualitas data yang digunakan dalam penelitian ini.

b. Evaluasi Efisiensi Pipeline

Efisiensi dalam proses pembersihan data menjadi aspek krusial dalam analisis data, terutama ketika menangani dataset yang besar dan kompleks. Salah satu indikator utama efisiensi adalah

waktu eksekusi, yang dapat dibandingkan antara metode manual dan metode otomatis untuk menentukan pendekatan yang lebih optimal. Metode manual sering kali memerlukan intervensi langsung dari pengguna, seperti mengidentifikasi dan menangani missing values serta outliers secara manual, yang menyebabkan proses menjadi lebih lambat dan rentan terhadap kesalahan manusia. Sebaliknya, metode otomatis berbasis Python yang dikembangkan dalam penelitian ini memungkinkan pemrosesan data yang lebih cepat dan konsisten tanpa mengurangi kualitas hasil akhir. Berdasarkan eksperimen yang dilakukan, metode manual membutuhkan waktu 150 detik, sementara metode otomatis hanya memerlukan 97 detik, menunjukkan peningkatan efisiensi sebesar 35%. Untuk menggambarkan perbedaan efisiensi ini dengan lebih jelas, Gambar 1 menampilkan perbandingan waktu eksekusi antara metode manual dan otomatis dalam proses pembersihan data.



Gambar 2. Grafik Perbandingan Waktu Pembersihan Manual vs Otomatis

Seperti yang ditunjukkan pada Gambar 2, metode manual dalam pembersihan data membutuhkan waktu eksekusi yang lebih lama dibandingkan metode otomatis, yang terlihat dari tingginya batang berwarna abu-abu pada grafik. Perbedaan ini disebabkan oleh banyaknya langkah manual yang harus dilakukan dalam metode tradisional, seperti inspeksi visual dan koreksi data satu per satu. Sementara itu, metode otomatis (batang hijau) menunjukkan waktu eksekusi yang lebih singkat, mengindikasikan bahwa algoritma yang digunakan mampu menangani pembersihan data dengan lebih efisien. Penurunan waktu ini bukan hanya menghemat sumber daya, tetapi juga meningkatkan reprodusibilitas dan konsistensi hasil, karena proses otomatis dapat dijalankan kembali dengan hasil yang serupa tanpa variasi subjektif dari pengguna. Selain itu, pendekatan otomatis memberikan fleksibilitas lebih tinggi untuk diterapkan pada dataset yang lebih besar tanpa peningkatan waktu eksekusi yang signifikan. Dengan demikian, hasil ini menegaskan bahwa metode otomatis merupakan solusi yang lebih efisien dibandingkan metode manual dalam konteks pembersihan data berskala besar.

Diskusi

Hasil penelitian ini menunjukkan bahwa pipeline otomatis berbasis Python yang dikembangkan mampu meningkatkan efisiensi proses pembersihan data dalam analitik big data. Dengan mengintegrasikan Pandas dan Scikit-learn, pipeline ini berhasil mengurangi missing values sebesar 95,7% dan outliers sebesar 91,7%, serta mempercepat waktu pemrosesan hingga 35% dibandingkan metode manual. Menurut (Maharana et al., 2022) dan (Tecimer et al., 2022), otomatisasi pemrosesan data dapat mengurangi beban kerja manual dan meningkatkan kualitas data secara signifikan, sehingga temuan ini sejalan dengan teori efisiensi big data. Selain itu, (Mehta & Klarmann, 2024) serta (Kolthoff et al., 2023) menyatakan bahwa metode otomatis lebih konsisten dan efisien dibandingkan pendekatan manual, meskipun menghadapi tantangan dalam fleksibilitas penerapannya, yang juga tercermin dalam hasil penelitian ini. Meskipun pipeline yang dikembangkan menunjukkan peningkatan efisiensi, terdapat kendala dalam adaptasi pipeline terhadap dataset dengan karakteristik yang lebih beragam. (El Hachimi et al., 2022) menambahkan bahwa efektivitas pipeline sangat dipengaruhi oleh konfigurasi awal yang disesuaikan dengan karakteristik dataset spesifik, sehingga aspek ini perlu mendapatkan perhatian lebih lanjut dalam penelitian berikutnya.

Penelitian ini juga mengidentifikasi beberapa keterbatasan yang masih perlu diperbaiki untuk meningkatkan fleksibilitas dan skalabilitas pipeline otomatis. Salah satu tantangan utama adalah ketergantungan pada parameter tetap dalam metode imputasi dan deteksi outliers, yang dapat mengurangi fleksibilitas pipeline saat diterapkan pada dataset yang lebih heterogen. Selain itu, keterbatasan dalam skalabilitas juga menjadi perhatian utama, mengingat Pandas dan Scikit-learn memiliki efisiensi memori yang terbatas saat menangani dataset berukuran sangat besar. Menurut (Cravero et al., 2022), keterbatasan dalam efisiensi memori sering menjadi kendala utama dalam pemrosesan data berukuran besar menggunakan alat analitik berbasis Python. Meskipun pipeline berhasil mengurangi waktu pemrosesan secara signifikan, efektivitasnya pada dataset dengan skala industri masih perlu dievaluasi lebih lanjut. Beberapa studi menyebutkan bahwa penggunaan teknologi big data yang lebih canggih, seperti Spark atau Dask, memungkinkan pemrosesan data dalam jumlah besar dengan lebih efisien. Dalam konteks ini, pengembangan pendekatan adaptif juga menjadi aspek yang perlu diperhatikan, khususnya dalam menyesuaikan parameter pembersihan data secara otomatis berdasarkan karakteristik dataset yang dianalisis.

Penelitian ini juga menyoroti beberapa perbedaan dibandingkan studi sebelumnya yang menyatakan bahwa metode manual lebih unggul dalam menangani dataset dengan struktur data yang kompleks. Menurut (Ikotun et al., 2023) serta (Leotta et al., 2024), pendekatan manual sering kali memungkinkan penyesuaian yang lebih fleksibel dalam menangani variasi data, tetapi hasil penelitian ini menunjukkan bahwa otomatisasi dapat memberikan efisiensi yang lebih tinggi dengan tetap mempertahankan akurasi dalam deteksi outliers dan imputasi missing values. Dalam beberapa kasus, intervensi manual masih diperlukan, terutama ketika menghadapi anomali yang sulit dideteksi oleh metode otomatis. (Sungwon et al., 2024) menyatakan bahwa kombinasi antara pendekatan manual dan otomatis dapat menghasilkan hasil yang lebih optimal dalam pembersihan data, yang juga diperkuat oleh temuan dalam penelitian ini. Pendekatan hybrid ini memungkinkan pemanfaatan keunggulan otomatisasi dalam meningkatkan kecepatan dan konsistensi, sekaligus mempertahankan fleksibilitas dalam menghadapi data yang memiliki struktur tidak teratur. Pengembangan lebih lanjut pada pipeline otomatis dapat mempertimbangkan integrasi pendekatan manual dalam tahap-tahap tertentu untuk meningkatkan efektivitas pembersihan data dalam berbagai skenario analitik.

4. Kesimpulan

Hasil penelitian ini menunjukkan bahwa pipeline otomatis berbasis Pandas dan Scikit-learn secara signifikan meningkatkan efisiensi dalam proses pembersihan data. Pipeline ini mampu mengurangi missing values dan outliers dengan tingkat akurasi yang tinggi serta mempercepat

waktu pemrosesan dibandingkan metode manual. Temuan ini menegaskan bahwa otomatisasi dalam pembersihan data tidak hanya meningkatkan kualitas dataset, tetapi juga mengurangi beban kerja manual yang selama ini menjadi tantangan dalam analitik big data. Selain itu, evaluasi terhadap pipeline menunjukkan bahwa metode ini dapat diterapkan secara luas dalam berbagai skenario analitik yang membutuhkan data berkualitas tinggi. Namun, beberapa kendala dalam hal fleksibilitas dan skalabilitas masih menjadi tantangan dalam penerapan pipeline ini pada dataset yang lebih heterogen dan berskala besar. Oleh karena itu, penelitian ini memberikan kontribusi penting dalam mengembangkan solusi otomatisasi pembersihan data yang lebih efisien, sekaligus membuka peluang untuk eksplorasi lebih lanjut guna meningkatkan adaptabilitas pipeline terhadap karakteristik data yang lebih kompleks.

Berdasarkan hasil yang diperoleh, terdapat beberapa rekomendasi yang dapat dikembangkan dalam penelitian selanjutnya. Salah satu aspek yang perlu diteliti lebih lanjut adalah penerapan pipeline ini pada dataset yang lebih besar dan lebih kompleks untuk menguji efektivitasnya dalam skala industri atau sistem dengan volume data yang lebih masif. Selain itu, integrasi algoritma pembelajaran mesin dalam pipeline pembersihan data dapat menjadi arah penelitian berikutnya untuk meningkatkan akurasi dalam mendeteksi missing values, outliers, serta inkonsistensi lainnya dalam data. Penggunaan pendekatan berbasis kecerdasan buatan berpotensi untuk membuat pipeline lebih adaptif terhadap variasi dataset, sehingga dapat meningkatkan fleksibilitas dan skalabilitas dalam berbagai konteks aplikasi. Penelitian masa depan juga dapat mengeksplorasi integrasi pipeline dengan teknologi big data seperti Apache Spark atau Dask untuk mengatasi keterbatasan dalam efisiensi memori dan pemrosesan data dalam jumlah besar. Dengan mengembangkan pendekatan yang lebih adaptif dan scalable, pipeline ini dapat menjadi solusi yang lebih komprehensif dalam mendukung analitik big data secara efisien dan akurat.

References

Aljaghoub, H., Abumadi, F., AlMallahi, M. N., Obaideen, K., & Alami, A. H. (2022). Solar PV Cleaning Techniques Contribute to Sustainable Development Goals (SDGs) Using Multi-Criteria Decision-Making (MCDM): Assessment and Review. *International Journal of Thermofluids*, *16*, 100233. https://doi.org/10.1016/j.ijft.2022.100233

Bailey, N. W., Hill, A. T., Biabani, M., Murphy, O. W., Rogasch, N. C., McQueen, B., Miljevic, A., & Fitzgerald, P. B. (2023). RELAX Part 2: A fully Automated EEG Data Cleaning Algorithm That is Applicable to Event-Related-Potentials. *Clinical Neurophysiology*, *149*, 202–222. https://doi.org/10.1016/j.clinph.2023.01.018

Behrad, F., & Saniee Abadeh, M. (2022). An Overview of Deep Learning Methods for Multimodal Medical Data Mining. *Expert Systems with Applications*, *200*, 117006. https://doi.org/10.1016/j.eswa.2022.117006

Chen, X., Zou, D., & Xie, H. (2022). A Decade of Learning Analytics: Structural Topic Modeling Based Bibliometric Analysis. *Education and Information Technologies*, *27*(8), 10517–10561. https://doi.org/10.1007/s10639-022-11046-z

Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022). Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. *Agronomy*, *12*(3), 748. https://doi.org/10.3390/agronomy12030748

Dewi, M. U., Santoso, L., & Santoso, A. B. (2024). Optimizing AI Performance in Industry: A Hybrid Computing Architecture Approach Based on Big Data. *Journal of Technology*

Informatics and Engineering, *3*(3), 308–323. https://doi.org/10.51903/jtie.v3i3.201

El Hachimi, C., Belaqziz, S., Khabba, S., & Chehbouni, A. (2022). Data Science Toolkit: An All-In-One Python Library to Help Researchers and Practitioners in Implementing Data Science-Related Algorithms with Less Effort. *Software Impacts*, *12*, 100240. https://doi.org/10.1016/j.simpa.2022.100240

Elouataoui, W., El Mendili, S., & Gahi, Y. (2023). An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis. *Data*, *8*(12), 182. https://doi.org/10.3390/data8120182

Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data Preparation: A Technological Perspective and Review. *SN Computer Science*, *4*(4), 1–20. https://doi.org/10.1007/s42979-023-01828-8

Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data Pipeline Quality: Influencing Factors, Root Causes of Data-Related Issues, and Processing Problem Areas for Developers. *Journal of Systems and Software*, *207*, 111855. https://doi.org/10.1016/j.jss.2023.111855

Giovanelli, J., Bilalli, B., & Abelló, A. (2022). Data Pre-Processing Pipeline Generation for AutoETL. *Information Systems*, *108*, 101957. https://doi.org/10.1016/j.is.2021.101957

Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2023). AI-Big Data Analytics for Building Automation and Management Systems: A Survey, Actual Challenges and Future Perspectives. *Artificial Intelligence Review*, *56*(6), 4929–5021. https://doi.org/10.1007/s10462-022-10286-2

Hu, X., Mar, D., Suzuki, N., Zhang, B., Peter, K. T., Beck, D. A. C., & Kolodziej, E. P. (2023). Mass-Suite: A Novel Open-Source Python Package for High-Resolution Mass Spectrometry Data Analysis. *Journal of Cheminformatics*, *15*(1), 1–13. https://doi.org/10.1186/s13321-023-00741-9

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jamaludin, H., Achlison, U., & Rokhman, N. (2024). Enhancing AI Model Accuracy and Scalability Through Big Data and Cloud Computing. *Journal of Technology Informatics and Engineering*, *3*(3), 296–307. https://doi.org/10.51903/jtie.v3i3.203

Kolthoff, K., Bartelt, C., & Ponzetto, S. P. (2023). Data-Driven Prototyping via Natural-Language-Based GUI Retrieval. *Automated Software Engineering*, *30*(1), 1–34. https://doi.org/10.1007/s10515-023-00377-x

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2024). A Practical Applications Guide to Machine Learning Regression Models in Psychology with Python. *Methods in Psychology*, *11*, 100156. https://doi.org/10.1016/j.metip.2024.100156

Leotta, M., Ricca, F., Marchetto, A., & Olianas, D. (2024). An Empirical Study to Compare Three Web Test Automation Approaches: NLP-Based, Programmable, and Capture & Replay. *Journal of Software: Evolution and Process*, *36*(5), 1–24. https://doi.org/10.1002/smr.2606

Maharana, K., Mondal, S., & Nemade, B. (2022). A Review: Data Pre-Processing and Data Augmentation Techniques. *Global Transitions Proceedings*, *3*(1), 91–99.

https://doi.org/10.1016/j.gltp.2022.04.020

Mehta, D., & Klarmann, N. (2024). Autoencoder-Based Visual Anomaly Localization for Manufacturing Quality Control. *Machine Learning and Knowledge Extraction*, *6*(1), 1–17. https://doi.org/10.3390/make6010001

Mumuni, A., & Mumuni, F. (2024). Automated data processing and feature engineering for deep learning and big data applications: A survey. *Journal of Information and Intelligence*, *3*(2), 113–153. https://doi.org/10.1016/j.jiixd.2024.01.002

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B. (2022). Data Management for Production Quality Deep Learning Models: Challenges and Solutions. *Journal of Systems and Software*, *191*, 111359. https://doi.org/10.1016/j.jss.2022.111359

Sungwon, I., Lin, T., North, C., Pfister, H., & Yang, Y. (2024). This is the Table I Want! Interactive Data Transformation on Desktop and in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, *30*(8), 5635–5650. https://doi.org/10.1109/tvcg.2023.3299602

Susatyono, J. D., Suasana, I. S., & Rozikin, K. (2024). Integrating Big Data and Edge Computing for Enhancing AI Efficiency in Real-Time Applications. *Journal of Technology Informatics and Engineering*, *3*(3), 337–349. https://doi.org/10.51903/jtie.v3i3.204

Tecimer, K. A., Tüzün, E., Moran, C., & Erdogmus, H. (2022). Cleaning Ground Truth Data in Software Task Assignment. *Information and Software Technology*, *149*, 106956. https://doi.org/10.1016/j.infsof.2022.106956

Teimourzadeh, A., Kakavand, S., & Kakavand, B. (2023). Application of Python in Marketing Education: A Big Data Analytics Perspective. *Marketing Education Review*, *33*(3), 226–241. https://doi.org/10.1080/10528008.2021.2021374

Theodorakopoulos, L. ;, Theodoropoulou, A. ;, Stamatiou, Y. A., Theodorakopoulos, L., Theodoropoulou, A., & Stamatiou, Y. (2024). A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions. *Eng*, *5*(3), 1266–1297. https://doi.org/10.3390/eng5030068