Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

ANALISIS TIPE DATA CAMPURAN: SEBUAH TINJAUAN LITERATUR SISTEMATIS

Hasta Pratama[1*], Fetty Fitriyanti Lubis[2], Jaka Sembiring[3]
[1]Program Doktor Teknik Elektro dan Informatika, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung
[2,3]Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung
Email: [1*]33220015@std.stei.itb.ac.id, [2]fettyfitriyanti@itb.ac.id, [3]jaka@itb.ac.id
(* : coressponding author)

Abstrak- Penelitian ini bertujuan untuk mengetahui arah penelitian pada analisis tipe data campuran. Saat ini, dunia tengah dibanjiri oleh data yang semakin beragam, tidak hanya terbatas pada tipe data numerik atau kategorikal saja, tetapi juga dapat berupa kombinasi dari keduanya. Dalam *Data Mining*, analisis data campuran sangat menantang karena data numerik dan data kategorik memiliki sifat yang berbeda. Metodologi penelitian ini menggunakan kerangka PICOC (*Population, Intervention, Comparation, Outcome, Context*) untuk mengumpulkan dan meneliti literatur terkait. Temuan utama dari survei literatur komprehensif ini mengungkapkan bahwa sebagian besar penelitian terkait data campuran disebarluaskan di jurnal bereputasi Q1 yang artinya topik analisis data campuran masih menarik. Model *clustering* adalah model yang paling sering digunakan dalam bidang analisis data campuran, meskipun metrik akurasi masih menjadi tolok ukur evaluasi yang berlaku dengan membandingkan dengan *data cluster* yang seharusnya. Pengelolaan data campuran umumnya menggunakan teknik *normalization* yaitu dengan melakukan normalisasi ukuran agar dapat menyatukan kedua tipe data. Kesimpulan dari hasil tinjauan literatur yaitu perlunya pengembangan pada data campuran yang tidak berlabel, tidak hanya modelnya tetapi juga metrik untuk mengukur hasilnya. Selain itu, penelitian ini menggarisbawahi pentingnya pengembangan model yang komprehensif mulai dari pemilihan fitur hingga evaluasi model. Untuk itu, penelitian mengenai analisis tipe data campuran masih merupakan bidang yang memiliki banyak ruang untuk eksplorasi dan potensi inovasi. Potensi ini terutama terlihat dalam bidang pengembangan model dinamis karena tipe data campuran memiliki keragaman penanganan mengikuti jenis tipe data campuran yang dianalisis.

Kata Kunci: analisis data campuran, analisis data, data mining, tinjauan literatur sistematis, picoc.

Abstract- *This research aims to determine the direction of research in the analysis of mixed data types. The world is currently filled with increasingly diverse data, especially in terms of data types, which are not only numerical or categorical but can be both (mixed). In Data Mining, the analysis of mixed data poses significant challenges because numerical and categorical data exhibit different properties. The research methodology employed in this study utilizes the PICOC framework (Population, Intervention, Comparison, Outcome, Context) to collect and review relevant literature. The primary findings from this comprehensive literature survey reveal that a majority of the research related to mixed data is published in reputable journals Q1, indicating sustained interest in the topic of mixed data analysis. Clustering models emerge as the most frequently used models in the field of mixed data analysis. However, it's noteworthy that accuracy metrics remain the predominant evaluation benchmark, often leading to comparisons with the ideal clustered data. The management of mixed data typically involves normalization techniques, specifically normalizing the scale to amalgamate the two types of data. The conclusion drawn from the results of the literature review is the necessity to develop unlabeled mixed data, encompassing both the model and metrics required to assess the outcomes. Additionally, this research emphasizes the significance of a comprehensive development model, ranging from feature selection to evaluation models. Therefore, the analysis of mixed data types remains a field with ample opportunities for exploration and potential innovation. This potential is primarily evident in the field of dynamic model development because mixed data types exhibit diverse handling, depending on the specific types of mixed data being analyzed.*

Keywords*: mixed data analysis, data analysis, data mining, systematic literature review, picoc.*

1. PENDAHULUAN

Perkembangan data yang semakin beragam dan terus berkembang baik tipe, sumber maupun jenisnya sehingga membutuhkan analisis yang lebih spesifik. Banyak data riil yang berupa data campuran, yaitu data bertipe numerik dan bertipe kategorik [1]. Banyak analisis yang dilakukan mulai dari *feature selection*, *clustering* maupun *classification*, hingga metrik untuk tipe data campuran. Meskipun penelitian tentang analisis pada data campuran telah ada lebih dari satu dekade tetapi kebutuhannya semakin terasa.

Tuntutan dan tantangan terkait tipe data campuran semakin besar. Dalam bidang analisis dan pemrosesan data, keberadaan tipe data campuran menimbulkan tantangan yang berat. Alasannya terletak pada fakta bahwa tipe data yang beragam menuntut operasi dan fungsi yang berbeda untuk analisis dan manipulasi. Sifat tipe data yang beragam ini tidak hanya mempersulit proses analisis, tetapi juga meningkatkan waktu yang diperlukan untuk melakukan persiapan sebelum dilakukan analisis. Selain itu pemrosesan data melalui konversi dapat menimbulkan

Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

potensi kehilangan atau pengurangan informasi [1]. Misalnya, selama operasi seperti pemetaan data dan konversi jenis data.

Dalam makalah penelitian ini, kami membahas mengenai peluang dan masalah tipe data campuran dengan menggunakan metode Tinjauan Literatur Sistematis (SLR). SLR dalam penelitian ini menggunakan metode PICOC (*Population, Intervention, Comparation, Outcome, Context*) agar tinjauan literatur lebih runut dan memiliki tujuan yang lebih jelas [2]. Makalah ini terdiri dari 4 bagian. Bagian pertama memberikan pembahasan mengenai latar belakang pemilihan topik ini. Bagian kedua menjelaskan SLR sebagai metodologi yang digunakan pada makalah ini. Bagian ketiga menjelaskan hasil yang diperoleh berupa tinjauan dan jawaban pada pertanyaan penelitian. Bagian keempat adalah kesimpulan penelitian dan saran untuk penelitian berikutnya.

2. METODE PENELITIAN

2.1 Tinjauan Literatur Sistematis

SLR adalah studi sekunder yang dirancang untuk memetakan secara sistematis, mengidentifikasi, mengevaluasi secara kritis, mengkonsolidasikan, dan mensintesis temuan-temuan dari studi primer yang berkaitan dengan topik penelitian tertentu [2]. Metodologi SLR telah menjadi standar yang diakui untuk memperoleh wawasan yang komprehensif dengan melakukan tinjauan yang ketat terhadap penelitian terdahulu yang relevan. Tujuan utama dari pelaksanaan SLR adalah untuk meringkas penelitian terdahulu secara ringkas, menunjukkan kesenjangan yang ada yang perlu diatasi antara penelitian terdahulu dan penelitian saat ini, menghasilkan laporan yang kohesif atau sintesis, dan menetapkan kerangka kerja penelitian.

Tujuan dari tinjauan literatur dalam penelitian ini adalah untuk memeriksa secara komprehensif pendekatan untuk mengelola data campuran, yang mencakup kerangka kerja, metodologi, dan platform yang digunakan. Selain itu, penelitian ini juga berusaha untuk mempelajari aspek-aspek tambahan termasuk peneliti yang produktif dan tempat publikasi terkemuka yang berkaitan dengan domain data mining dalam konteks data campuran. Untuk memastikan hasil yang kuat, penelitian ini akan memanfaatkan database jurnal terkemuka seperti IEEE Xplore, Scopus, SpringerLink, ScienceDirect, dan ACM, yang mencakup tahun 2019 hingga 2023.

2.2 Pertanyaan Penelitian

Tujuan utama dari menyusun pertanyaan penelitian adalah untuk menjadi panduan bagi seluruh proses tinjauan literatur. Elemen penting ini memastikan bahwa tinjauan tersebut mempertahankan lintasan yang jelas dan terfokus, sehingga meningkatkan efisiensi pencarian data. Dengan merumuskan pertanyaan penelitian yang terdefinisi dengan baik, para peneliti dapat menyederhanakan upaya mereka untuk menggali informasi yang relevan dari sumber-sumber literatur yang tersedia. Pada Tabel 1 merupakan atribut PICOC yang digunakan untuk membangun pertanyaan-pertanyaan penelitian.

Tabel 1. Atribut PICOC.

Indikator	Terminasi
Population	Pencarian *Mixed Data Types Analysis*, *Mixed Data types Clustering* dari 2019-2023
Intervention	Penggunaan analisis data untuk tipe data campuran
Context	-
Outcome	Model dengan performa terbaik
Comparison	Perbedaan penggunaan Teknik data analisis pada tipe data campuran

Pada Tabel 2 di bawah ini, kami menyajikan gambaran umum yang komprehensif tentang pertanyaan penelitian spesifik yang telah diuraikan untuk penelitian ini. Pertanyaan-pertanyaan ini berfungsi sebagai fondasi yang mendasari seluruh tinjauan literatur sistematis yang dibangun, untuk memastikan eksplorasi yang cermat dan terarah pada area penelitian yang dipilih. Empat pertanyaan penelitian pada makalah ini tidak hanya berfokus pada konten penelitian saja tetapi juga meliputi pada aspek pendukung yaitu pada pertanyaan pertama yaitu tempat penelitian dipublikasi. Hal tersebut dapat menjadi acuan bagi peneliti berikutnya sehingga dapat menargetkan tempat publikasi penelitiannya. Sementara itu, tiga pertanyaan penelitian lainnya berfokus pada konten penelitian terutama pada analisis data campuran.

Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Tabel 2. Pertanyaan penelitian.

ID	Pertanyaan Penelitian	Motivasi
RQ1	Jurnal mana yang paling menonjol dalam bidang analisis data campuran dalam ilmu data?	Mengidentifikasi target tempat publikasi untuk makalah tentang analisis data campuran
RQ2	Algoritme atau model mana yang paling sering digunakan?	Mengidentifikasi algoritme atau model yang didukung untuk analisis data campuran.
RQ3	Apa saja metrik yang digunakan untuk pengukuran?	Mengidentifikasi metrik penting untuk mengukur analisis data campuran.
RQ4	Bagaimana penanganan tipe data campuran?	Mengidentifikasi teknik yang diperlukan untuk mengelola tipe data campuran

2.3 Temuan Penelitian

Untuk menjawab pertanyaan penelitian yang telah diuraikan sebelumnya, penelitian ini memulai pencarian komprehensif untuk makalah penelitian dalam basis data jurnal yang sudah mapan, dengan menggunakan istilah pencarian yang tepat selama bulan September. Sistematika pencarian mencakup kata kunci tertentu seperti ("*data mining*" OR "*data analysis*" OR "*data analytics*" OR *clustering* OR *cluster*) AND ("*mixed data*" OR "*qualitative quantitative data*"). Hasil dari pencarian yang sangat teliti ini disajikan pada Tabel 3.

Tabel 3. Hasil temuan

Index	Basis Data Jurnal	Jumlah Artikel
1	IEEE Xplore	84
2	ScienceDirect	1.496
3	SpringerLink	2.372
4	Scopus	477
5	ACM Digital Library	227
Total		4.656

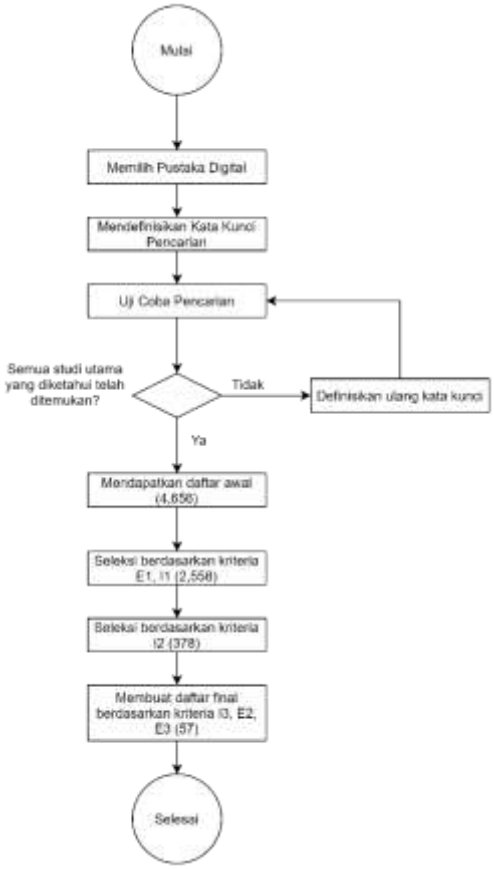
Penelitian ini menerapkan beberapa kriteria inklusi dan eksklusi untuk menyaring artikel-artikel untuk eksplorasi lebih lanjut. Tabel 4 menampilkan kriteria-kriteria ini. Kriteria inklusi adalah kriteria untuk artikel yang akan dimuat dalam SLR. Kriteria inklusi yang penelitian ini gunakan terdiri dari tiga kriteria. Sedangkan kriteria eksklusi yang digunakan pada penelitian ini untuk mengeluarkan artikel yang tidak sesuai terdiri dari tiga kriteria.

Tabel 4. Kriteria inklusi dan eksklusi

Nomor	Jenis Kriteria		Kriteria
1	Kriteria Inklusi	I1:	Artikel ditulis dalam bahasa Inggris.
		I2:	Artikel yang berkaitan dengan topik-topik dalam ilmu komputer.
		I3:	Akses teks lengkap tersedia untuk artikel.
2	Kriteria Eksklusi	E1:	Artikel dalam buku, publikasi, tinjauan konferensi, atau makalah.
		E2:	Artikel yang disajikan dalam bentuk survei atau tinjauan pustaka.
		E3:	Artikel yang tidak memiliki fokus utama pada algoritme atau model.

Gambar 1 mengilustrasikan tahapan penyaringan yang dilakukan sesuai dengan kriteria yang ditentukan, termasuk judul/metadata, abstrak, akses resmi, dan tinjauan teks lengkap. Pada penyaringan awal, kriteria I1 dan E1 yang diterapkan pada bagian judul dan metadata menghasilkan pengurangan jumlah makalah menjadi 2.558. Pada penyaringan kedua, penerapan kriteria I2 semakin mengurangi jumlah makalah menjadi 378 makalah. Penyaringan tahap ketiga, dengan menerapkan kriteria I3, E2, E3 dan melakukan evaluasi (dibahas pada sub bab berikutnya), mempersempit jumlah makalah menjadi 57.

Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 1. Tahapan Tinjauan Literatur Sistematis [2]

2.4 Evaluasi

Kualitas informasi yang dikumpulkan dari survei *state of the art* kemudian dinilai dengan apa yang disebut *Check List Primary Quality Assessment of Research*. Daftar ini akan memberikan metode untuk menilai secara kuantitatif kualitas makalah yang dipilih untuk tinjauan sistematis. Skala tiga tingkat telah ditetapkan sebagai berikut:

- Ya, 1 point
- Tidak, 0 point
- Sebagian, 0.5 points

Tabel 5. Pertanyaan Assesmen

Nomor	Pertanyaan	Pilihan
1	Apakah artikel tersebut dikutip?	Ya / Tidak
2	Apakah tujuan penelitian dinyatakan dengan jelas?	Ya / Tidak / Sebagian
3	Apakah peserta penelitian atau unit observasi telah dijelaskan secara memadai?	Ya / Tidak / Sebagian
4	Apakah pengumpulan data dilakukan dengan sangat baik?	Ya / Tidak / Sebagian
5	Apakah potensi perancu dikontrol secara memadai dalam analisis?	Ya / Tidak / Sebagian
6	Apakah pendekatan dan perumusan analisis disampaikan dengan baik?	Ya / Tidak / Sebagian
7	Apakah temuan-temuannya kredibel?	Ya / Tidak / Sebagian

Setiap penelitian dapat memperoleh skor antara 0 dan 7 poin, yang sesuai dengan tingkat kesesuaiannya dengan pertanyaan-pertanyaan penilaian. Nilai yang lebih tinggi untuk item tertentu menunjukkan bahwa makalah tersebut secara efektif membahas tujuan penelitian. Selanjutnya, setengah dari penelitian, khususnya yang

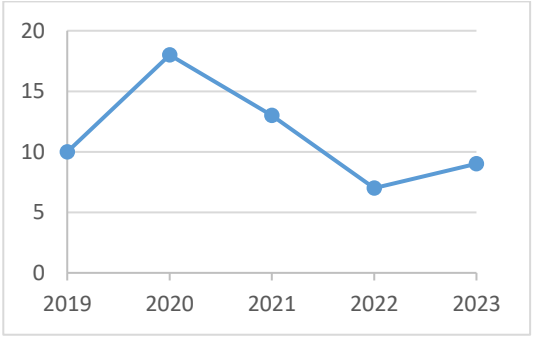
Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

mendapatkan nilai 4 atau lebih tinggi, dipilih untuk membentuk "literatur akhir", sementara yang memiliki nilai lebih rendah tidak diikutsertakan. Dalam penelitian ini, semua makalah dengan nilai lebih dari 4 akan dimasukkan dalam pembahasan. Tabel 5 memberikan daftar rinci pertanyaan penilaian kualitas yang diterapkan pada setiap publikasi.

3. HASIL DAN PEMBAHASAN

3.1 Publikasi Jurnal

Publikasi temuan penelitian memiliki pengaruh yang signifikan terhadap dampak yang akan dihasilkan. Gambar 2 mengilustrasikan tren penelitian analitik pada data campuran dari tahun 2019 hingga 2023. Meskipun terjadi penurunan sejak tahun 2020, tetapi terlihat dari jumlah penelitian menunjukkan bahwa minat ditopik ini masih tinggi. Sehingga subjek analisis data campuran masih dapat dieksplorasi lebih dalam.



Gambar 2. Tren Penelitian Analisis Tipe Data Campuran

Menurut Tabel 6, Jurnal Information Sciences menonjol sebagai penerbit terkemuka di bidang analisis data campuran, dengan 7 makalah atas namanya. Menyusul di bawahnya adalah Jurnal IEEE Access, yang telah menyumbangkan 6 makalah. Kedua jurnal ini memiliki peringkat Q1, yang menunjukkan kredibilitas tinggi dan kutipan ekstensif dari makalah yang mereka terbitkan.

Tabel 6. Daftar Jurnal Teratas dari Artikel Terpilih

Jurnal	Artikel	SJR	Kategori Q
Information Sciences	7	2.29	Q1
IEEE Access	6	0.93	Q1
Chemometrics and Intelligent Laboratory Systems	3	0.65	Q2
Pattern Recognition	3	2.09	Q1
Journal of Intelligent and Fuzzy Systems	3	0.37	Q3
Expert Systems with Applications	2	1.87	Q1
Entropy	2	0.54	Q2

Tabel 7 menunjukkan bahwa makalah tentang analisis data campuran sebagian besar diterbitkan di jurnal Q1. Hal ini menunjukkan bahwa tema analisis data campuran masih mampu untuk menembus jurnal dengan kualitas terbaik.

Tabel 7. Artikel Berdasarkan Kategori Jurnal

Nomor	Media Publikasi	Jumlah
1	Jurnal Q1	37
2	Jurnal Q2	9
3	Jurnal Q3	10
4	Jurnal Q4	1

Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3.2 Algoritma Analisis tipe data campuran

Banyak metode yang digunakan untuk analisis jenis data campuran. Sangatlah penting untuk memahami prevalensi metode-metode ini dalam praktiknya. Sesuai dengan temuan yang disajikan pada Tabel 8, metode *clustering* muncul sebagai fokus utama para peneliti ketika melakukan analisis pada data tipe campuran. Hal ini sesuai dengan fenomena data saat ini yang menunjukkan bahwa sebagian besar data yang tersedia tidak memiliki label yang terkait. Selanjutnya model yang diamati yaitu *Feature Selection.* Meskipun sebagian besar *feature selection* yang diamati digunakan untuk model *classification.* Berdasarkan hasil temuan ini dapat menjadi paduan menarik untuk menyatukan analisis tipe data campuran mulai dari *feature selection* dengan proses *clustering*.

Tabel 8. Model Analisis Tipe Data Campuran.

Nomor	Fokus Model	Literature
1	*Clustering*	[3][4][5][6][7][8][9][10][11][12][13][14][15][16][17] [18][19][20][21][22][23][24][25][26][27][28][29][30] [31][32][33][34] [35][36][37][38]
2	*Feature Selection*	[39][40][41][42][43][44][45] [46][47]
3	*Classification*	[48][49][50][51][52]
4	*Framework*	[53][54][55]
5	*Metrics*	[56][57]
6	*Regression*	[58]
7	*Frequent Pattern*	[59]

3.3 Metrik analisis tipe data campuran

Normalized Mutual Information (NMI) dan *Accuracy* paling banyak digunakan dalam beberapa penelitian ([17], [34], [54]) Namun, sangat penting untuk dicatat bahwa penerapan metrik-metrik ini bergantung pada informasi yang tepat tentang jumlah *cluster*, yang mungkin tidak selalu dapat dilakukan karena karakteristik data tertentu seperti pada data yang tidak memiliki label. Penelitian yang menggunakan *clustering* dan akurasi [36] sebagai metriknya telah memiliki data label sebagai pembanding sehingga tidak dapat diterapkan pada data yang sama sekali tidak memiliki label. Sebagai alternatif, beberapa peneliti telah memilih untuk menilai efektivitas klaster melalui stabilitas dan konsumsi waktu [38]. Dalam konteks pengelompokan, *silhouette coefficient* direkomendasikan untuk mengevaluasi kualitas cluster yang dihasilkan.

3.4 Metode untuk mengatasi tipe data campuran

Data campuran tidak dapat dianalisis secara langsung. Hal ini disebabkan oleh perbedaan ukuran dari masing-masing tipe data. Jika pada tipe numerik kita dapat melakukan operasi aritmatika sehingga dapat melihat rata-rata, varians, dan sebagainya, maka pada tipe kategorik kita hanya dapat melihat modusnya saja, karena jika kita melakukan operasi aritmatika akan terjadi bias karena angka-angka pada data kategorik tidak memiliki arti seperti pada data numerik. Oleh karena itu perlu dilakukan konsolidasi baik berupa konversi, *cleaning*, normalisasi maupun rekayasa fitur, walaupun ada beberapa yang melakukan analisis secara terpisah yaitu dengan melakukan *subsetting*. Pada Tabel 9 terlihat bahwa pada penelitian analisis data campuran masih banyak yang melakukan normalisasi dibandingkan dengan *subsetting*. Normalisasi dilakukan dengan menghitung nilai *entropy* atau menghitung bobot pada setiap fitur kemudian membandingkannya.

Tabel 9. Metode Penanganan Tipe Data Campuran.

Index	Metode	Literatur
1	*Datatype conversion*	[29][47][34][26][22][19][59][15][3]
2	*Subsetting*	[60] [61] [62] [50] [51][45][30][28][11][10][5][4]
3	*Data normalization*	[38][63][44][27][54][18][43][56][16][14][13][12][9][7] [40][6][39]
4	*Feature engineering*	[20]

Idealis : Indonesia Journal Information System
Volume 7, Nomor 2, Juli 2024
ISSN 2684-7280 (*online*)
Halaman 210-218
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

4. KESIMPULAN

Penelitian tinjauan literatur sistematis ini bertujuan untuk mengidentifikasi dan menganalisis metode, metrik, model dan tempat penerbitan makalah ilmiah yang berhubungan dengan analisis data campuran hingga September 2023. Kami melakukan pencarian pada lima penyedia pustaka digital, ScienceDirect, Scopus, SpringerLink, IEEEXplor, dan ACM. Dari 4.656 penelitian yang ditemukan dilakukan seleksi hingga menghasilkan 57 penelitian.

Penelitian ini berfokus pada analisis data bertipe campuran meliputi model yang sering digunakan, metode untuk mengatasi tipe data campuran serta metrik yang tepat untuk mengukurnya. Selain itu, kami menganalisis tempat publikasi yang menerbitkan analisis pada data campuran sebagai referensi bahwa topik ini masih sangat relevan.

Penelitian ini menemukan bahwa penelitan analisis pada data campuran paling banyak dilakukan dengan model *clustering* dan *feature selection*. Selain itu, normalisasi data lebih populer untuk mengatasi perbedaan tipe data. Meskipun, normalisasi data berpotensi untuk menghilangkan informasi dari suatu variabel.

Tantangan penelitian selanjutnya adalah bagaimana mendapatkan performa terbaik untuk mengatasi perbedaan tipe data tanpa menghilangkan informasi. Selain itu penelitian analisis tipe data campuran yang murni tidak berlabel masih minim sehingga dapat menjadi acuan untuk penelitian berikutnya. Untuk mengatasi masalah perbedaan tipe data baik dari segi jumlah variabel yang perlu dipilih sebaiknya dilakukan lebih dinamis agar dapat menyesuaikan dengan kebutuhan penelitian dengan mengacu pada pola data yang digunakan.

DAFTAR PUSTAKA

[1] A. Ahmad dan S. S. Khan, "Survey of State-of-the-Art Mixed Data Clustering Algorithms," *IEEE Access*, vol. 7, hal. 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

[2] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, no. Vol 1, No 1 (2015), hal. 1–16, 2015, [Daring]. Tersedia pada: http://journal.ilmukomputer.org/index.php/jse/article/view/47.

[3] S. Xu, L. Feng, S. Liu, dan H. Qiao, "Self-adaption neighborhood density clustering method for mixed data stream with concept drift," *Eng. Appl. Artif. Intell.*, vol. 89, no. November 2019, hal. 103451, 2020, doi: 10.1016/j.engappai.2019.103451.

[4] A. J. M. S. Arockiam dan E. S. Irudhayaraj, "Reclust: an efficient clustering algorithm for mixed data based on reclustering and cluster validation," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 29, no. 1, hal. 545 – 552, 2023, doi: 10.11591/ijeecs.v29.i1.pp545-552.

[5] R. Bi, D. Guo, Y. Zhang, R. Huang, L. Lin, dan J. Xiong, "Outsourced and Privacy-Preserving Collaborative K-Prototype Clustering for Mixed Data via Additive Secret Sharing," *IEEE Internet Things J.*, vol. 10, no. 18, hal. 15810–15821, 2023, doi: 10.1109/JIOT.2023.3266028.

[6] Y. Xu, X. Gao, dan X. Wang, "Nonparametric Clustering of Mixed Data Using Modified Chi-Squared Tests," *Entropy*, vol. 24, no. 12, 2022, doi: 10.3390/e24121749.

[7] Y.-G. Choi, S. Ahn, dan J. Kim, "Model-Based Clustering of Mixed Data With Sparse Dependence," *IEEE Access*, vol. 11, hal. 75945–75954, 2023, doi: 10.1109/ACCESS.2023.3296790.

[8] H. Rezaei dan N. Daneshpour, "Mixed data clustering based on a number of similar features," *Pattern Recognit.*, vol. 143, 2023, doi: 10.1016/j.patcog.2023.109815.

[9] R. J. Kuo, P. Amornnikun, dan T. P. Q. Nguyen, "Metaheuristic-based possibilistic multivariate fuzzy weighted c-means algorithms for market segmentation," *Appl. Soft Comput. J.*, vol. 96, hal. 106639, 2020, doi: 10.1016/j.asoc.2020.106639.

[10] K. R. Nirmal dan K. V. V. Satyanarayana, "Map reduce based removing dependency on K and initial centroid selection MR-REDIC algorithm for clustering of mixed data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, hal. 733–740, 2020, doi: 10.14569/ijacsa.2020.0110292.

[11] K. Balaji, "Machine learning algorithm for feature space clustering of mixed data with missing information based on molecule similarity," *J. Biomed. Inform.*, vol. 125, 2022, doi: 10.1016/j.jbi.2021.103954.

[12] K. Balaji dan K. Lavanya, "Machine learning algorithm for cluster analysis of mixed dataset based on instance-cluster closeness metric," *Chemom. Intell. Lab. Syst.*, vol. 215, 2021, doi: 10.1016/j.chemolab.2021.104346.

[13] T. P. Q. Nguyen, R. J. Kuo, M. D. Le, T. C. Nguyen, dan T. H. A. Le, "Local search genetic algorithm-based possibilistic weighted fuzzy c-means for clustering mixed numerical and categorical data," *Neural Comput. Appl.*, vol. 34, no. 20, hal. 18059–18074, 2022, doi: 10.1007/s00521-022-07411-1.

[14] M. Li, X. Li, dan J. Li, "High-Dimensional Clustering for Incomplete Mixed Dataset Using Artificial Intelligence," *IEEE Access*, vol. 8, hal. 69629–69638, 2020, doi: 10.1109/ACCESS.2020.2986813.

[15] L. Chen, L. Zeng, Y. Mu, dan L. Chen, "Global Combination and Clustering based Differential Privacy Mixed Data Publishing," *IEEE Trans. Knowl. Data Eng.*, hal. 1–12, 2023, doi: 10.1109/TKDE.2023.3237822.

[16] P. D'Urso dan R. Massari, "Fuzzy clustering of mixed data," *Inf. Sci. (Ny).*, vol. 505, hal. 513–534, 2019, doi: 10.1016/j.ins.2019.07.100.

[17] J. Zhou, K. Chen, dan J. Liu, "A clustering algorithm based on the weighted entropy of conditional attributes for mixed data," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 17, hal. 1–13, 2021, doi: 10.1002/cpe.6293.

[18] F. A. Mazarbhuiya, M. Y. Alzahrani, dan A. K. Mahanta, "Detecting anomaly using partitioning clustering with merging," *ICIC Express Lett.*, vol. 14, no. 10, hal. 951 – 960, 2020, doi: 10.24507/icicel.14.10.951.

[19] O. Koren, carina A. Hallin, nir Perel, dan D. Bendet, "Decision-Making Enhancement in a Big Data Environment: Application of the K-Means Algorithm to Mixed Data," *J. Artif. Intell. Soft Comput. Res.*, vol. 9, no. 4, hal. 293 – 302, 2019, doi: 10.2478/jaiscr-2019-0010.

[20] Y. Li, X. Chu, D. Tian, J. Feng, dan W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm

optimization algorithm," *Appl. Soft Comput.*, vol. 113, hal. 107924, 2021, doi: 10.1016/j.asoc.2021.107924.
[21] S. Behzadi, N. S. Müller, C. Plant, dan C. Böhm, "Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm," *Int. J. Data Sci. Anal.*, vol. 10, no. 3, hal. 233–248, 2020, doi: 10.1007/s41060-020-00216-2.
[22] F. Li, Y. Qian, J. Wang, F. Peng, dan J. Liang, "Clustering mixed type data: a space structure-based approach," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 9, hal. 2799–2812, 2022, doi: 10.1007/s13042-022-01602-x.
[23] J. Ji, W. Pang, Z. Li, F. He, G. Feng, dan X. Zhao, "Clustering Mixed Numeric and Categorical Data with Cuckoo Search," *IEEE Access*, vol. 8, hal. 30988–31003, 2020, doi: 10.1109/ACCESS.2020.2973216.
[24] J. Ji, Y. Chen, G. Feng, X. Zhao, dan F. He, "Clustering mixed numeric and categorical data with artificial bee colony strategy," *J. Intell. Fuzzy Syst.*, vol. 36, no. 2, hal. 1521–1530, 2019, doi: 10.3233/JIFS-18146.
[25] S. B. Kather dan B. K. Tripathy, "Clustering mixed data using neighbourhood rough sets," *Int. J. Adv. Intell. Paradig.*, vol. 15, no. 1, hal. 1 – 16, 2020, doi: 10.1504/IJAIP.2020.104103.
[26] B. Duan, L. Han, Z. Gou, Y. Yang, dan S. Chen, "Clustering mixed data based on density peaks and stacked denoising autoencoders," *Symmetry (Basel).*, vol. 11, no. 2, 2019, doi: 10.3390/sym11020163.
[27] K. Balaji, K. Lavanya, dan A. G. Mary, "Clustering algorithm for mixed datasets using density peaks and Self-Organizing Generative Adversarial Networks," *Chemom. Intell. Lab. Syst.*, vol. 203, no. April, hal. 104070, 2020, doi: 10.1016/j.chemolab.2020.104070.
[28] K. Balaji dan K. Lavanya, "Cluster analysis of mixed data based on Feature Space Instance Cluster Closeness Metric," *Chemom. Intell. Lab. Syst.*, vol. 215, no. May, hal. 104370, 2021, doi: 10.1016/j.chemolab.2021.104370.
[29] E. Mousavi dan M. Sehhati, "A generalized multi-aspect distance metric for mixed-type data clustering," *Pattern Recognit.*, vol. 138, 2023, doi: 10.1016/j.patcog.2023.109353.
[30] Z. Lv *et al.*, "An Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN-Based Smart Grid," *IEEE Access*, vol. 7, hal. 45773–45782, 2019, doi: 10.1109/ACCESS.2019.2909048.
[31] X. Yao, J. Wang, M. Shen, H. Kong, dan H. Ning, "An improved clustering algorithm and its application in IoT data analysis," *Comput. Networks*, vol. 159, hal. 63 – 72, 2019, doi: 10.1016/j.comnet.2019.04.022.
[32] H. Petwal dan R. Rani, "An efficient clustering algorithm for mixed dataset of postoperative surgical records," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, hal. 757–770, 2020, doi: 10.2991/ijcis.d.200601.001.
[33] M. Marbac dan V. Vandewalle, "A tractable multi-partitions clustering," *Comput. Stat. Data Anal.*, vol. 132, hal. 167 – 179, 2019, doi: 10.1016/j.csda.2018.06.013.
[34] S. Liu, H. Zhang, dan X. Liu, "A study on two-stage mixed attribute data clustering based on density peaks," *Int. Arab J. Inf. Technol.*, vol. 18, no. 5, hal. 634–643, 2021, doi: 10.34028/iajit/18/5/2.
[35] M. Salman, "A novel clustering method with consistent data in a three-dimensional graphical format over existing clustering mechanisms," *Inf. Sci. (Ny).*, vol. 649, no. February, hal. 119634, 2023, doi: 10.1016/j.ins.2023.119634.
[36] J. Ji, R. Li, W. Pang, F. He, G. Feng, dan X. Zhao, "A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data," *IEEE Access*, vol. 9, hal. 24913–24924, 2021, doi: 10.1109/ACCESS.2021.3057113.
[37] G. Xu, L. Zhang, C. Ma, dan Y. Liu, "A mixed attributes oriented dynamic SOM fuzzy cluster algorithm for mobile user classification," *Inf. Sci. (Ny).*, vol. 515, hal. 280 – 293, 2020, doi: 10.1016/j.ins.2019.12.019.
[38] X. Li, Z. Wu, Z. Zhao, F. Ding, dan D. He, "A mixed data clustering algorithm with noise-filtered distribution centroid and iterative weight adjustment strategy," *Inf. Sci. (Ny).*, vol. 577, hal. 697 – 721, 2021, doi: 10.1016/j.ins.2021.07.039.
[39] Z. Yuan, H. Chen, T. Li, Z. Yu, B. Sang, dan C. Luo, "Unsupervised attribute reduction for mixed data based on fuzzy rough sets," *Inf. Sci. (Ny).*, vol. 572, hal. 67–87, 2021, doi: 10.1016/j.ins.2021.04.083.
[40] L. Sun, L. Wang, W. Ding, Y. Qian, dan J. Xu, "Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems," *Knowledge-Based Syst.*, vol. 192, hal. 105373, 2020, doi: 10.1016/j.knosys.2019.105373.
[41] J. Matute dan L. Linsen, "Hinted Star Coordinates for Mixed Data," *Comput. Graph. Forum*, vol. 39, no. 1, hal. 117 – 133, 2020, doi: 10.1111/cgf.13666.
[42] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, dan W.-C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Syst.*, vol. 37, no. 5, 2020, doi: 10.1111/exsy.12553.
[43] J. Wan, H. Chen, T. Li, X. Yang, dan B. Sang, "Dynamic interaction feature selection based on fuzzy rough set," *Inf. Sci. (Ny).*, vol. 581, hal. 891 – 911, 2021, doi: 10.1016/j.ins.2021.10.026.
[44] S. Solorio-Fernández, J. F. Martínez-Trinidad, dan J. A. Carrasco-Ochoa, "A Supervised Filter Feature Selection method for mixed data based on Spectral Feature Selection and Information-theory redundancy analysis," *Pattern Recognit. Lett.*, vol. 138, hal. 321 – 328, 2020, doi: 10.1016/j.patrec.2020.07.039.
[45] A. Dutt dan M. A. Ismail, "A partition-based feature selection method for mixed data: A filter approach," *Malaysian J. Comput. Sci.*, vol. 33, no. 2, hal. 152–169, 2020, doi: 10.22452/mjcs.vol33no2.5.
[46] N. N. Thuy dan S. Wongthanavasu, "A Novel Feature Selection Method for High-Dimensional Mixed Decision Tables," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 7, hal. 3024–3037, 2022, doi: 10.1109/TNNLS.2020.3048080.
[47] A. Taha, A. S. Hadi, B. Cosgrave, dan S. McKeever, "A multiple association-based unsupervised feature selection algorithm for mixed data sets," *Expert Syst. Appl.*, vol. 212, 2023, doi: 10.1016/j.eswa.2022.118718.
[48] F. Rodriguez-Sanchez, P. Larrañaga, dan C. Bielza, "Incremental Learning of Latent Forests," *IEEE Access*, vol. 8, hal. 224420–224432, 2020, doi: 10.1109/ACCESS.2020.3027064.
[49] Q. Li, Q. Xiong, S. Ji, Y. Yu, C. Wu, dan M. Gao, "Incremental semi-supervised Extreme Learning Machine for Mixed data stream classification," *Expert Syst. Appl.*, vol. 185, 2021, doi: 10.1016/j.eswa.2021.115591.
[50] K. Baati, T. M. Hamdani, A. M. Alimi, dan A. Abraham, "A new possibilistic classifier for mixed categorical and numerical data based on a bi-module possibilistic estimation and the generalized minimum-based algorithm," *J. Intell. Fuzzy Syst.*, vol. 36, no. 4, hal. 3513–3523, 2019, doi: 10.3233/JIFS-181383.
[51] Q. Li, Q. Xiong, S. Ji, Y. Yu, C. Wu, dan H. Yi, "A method for mixed data classification base on RBF-ELM network," *Neurocomputing*, vol. 431, hal. 7–22, 2021, doi: 10.1016/j.neucom.2020.12.032.
[52] T. Kuo dan K. J. Wang, "A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification," *Comput. Ind. Eng.*, vol. 169, no. February, hal. 108164, 2022, doi: 10.1016/j.cie.2022.108164.
[53] J. Muller *et al.*, "Integrated Dual Analysis of Quantitative and Qualitative High-Dimensional Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 6, hal. 2953 – 2966, 2021, doi: 10.1109/TVCG.2021.3056424.
[54] Y. Lee, C. Park, dan S. Kang, "Deep Embedded Clustering Framework for Mixed Data," *IEEE Access*, vol. 11, hal. 33–40, 2023, doi: 10.1109/ACCESS.2022.3232372.
[55] D. T. Dinh, V. N. Huynh, dan S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Inf. Sci.*

(Ny)., vol. 571, hal. 418–442, 2021, doi: 10.1016/j.ins.2021.04.076.
[56] A. Grané, G. Manzi, dan S. Salini, "Dynamic Mixed Data Analysis and Visualization," *Entropy*, vol. 24, no. 10, hal. 1–12, 2022, doi: 10.3390/e24101399.
[57] L. Cheng, Y. Wang, dan X. Ma, "An end-to-end distance measuring for mixed data based on deep relevance learning," *Intell. Data Anal.*, vol. 24, no. 1, hal. 83–99, 2020, doi: 10.3233/IDA-184399.
[58] S.-K. Ng, R. Tawiah, dan G. J. McLachlan, "Unsupervised pattern recognition of mixed data structures with numerical and categorical features using a mixture regression modelling framework," *Pattern Recognit.*, vol. 88, hal. 261 – 271, 2019, doi: 10.1016/j.patcog.2018.11.022.
[59] A. Y. Rodriguez-Gonzalez, J. F. Martinez-Trinidad, J. A. Carrasco-Ochoa, J. Ruiz-Shulcloper, dan M. Alvarado-Mentado, "Frequent similar pattern mining using non boolean similarity functions," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, hal. 4931 – 4944, 2019, doi: 10.3233/JIFS-179040.
[60] T. Kuo dan K.-J. Wang, "A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification," *Comput. Ind. Eng.*, vol. 169, 2022, doi: 10.1016/j.cie.2022.108164.
[61] G. Xu, L. Zhang, C. Ma, dan Y. Liu, "A mixed attributes oriented dynamic SOM fuzzy cluster algorithm for mobile user classification," *Inf. Sci. (Ny).*, vol. 515, hal. 280–293, 2020, doi: 10.1016/j.ins.2019.12.019.
[62] J. Ji, R. Li, W. Pang, F. He, G. Feng, dan X. Zhao, "A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data," *IEEE Access*, vol. 9, hal. 24913–24924, 2021, doi: 10.1109/ACCESS.2021.3057113.
[63] J. Zhou, K. Chen, dan J. Liu, "A clustering algorithm based on the weighted entropy of conditional attributes for mixed data," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 17, 2021, doi: 10.1002/cpe.6293.