Penggunaan K-Means Clustering untuk Meningkatkan Hasil Process Mining pada Data Road Traffic Fine Management

Tugas Besar

Oleh:

Ridho Chan-1301223432

Mochammad Khoirullutfansyah-1301220126

Muhammad Zaidan Dhiyaulhaq-1301223255

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2025

1. PENDAHULUAN

1.1. Latar Belakang

Dalam pengelolaan proses bisnis modern, pemahaman terhadap alur proses yang kompleks sangat penting untuk meningkatkan efisiensi dan kualitas layanan. Kompleksitas proses dapat menyebabkan berbagai masalah seperti keterlambatan, pekerjaan ulang, dan pemborosan sumber daya. Oleh karena itu, pendekatan seperti Business Process Management (BPM) telah berkembang untuk memastikan bahwa strategi bisnis selaras dengan kebutuhan pelanggan dan pemangku kepentingan [1].

Salah satu pendekatan yang banyak digunakan dalam BPM adalah process mining, yaitu teknik analisis data yang memanfaatkan event log dari sistem informasi untuk menemukan, memverifikasi, dan meningkatkan model proses aktual [1]. Dengan process mining, organisasi dapat mengidentifikasi pola perilaku, variasi jalur proses, serta mengevaluasi kesesuaian antara proses yang dijalankan dengan model yang diharapkan. Teknik ini mencakup tiga aktivitas utama: process discovery, conformance checking, dan process enhancement [1].

Namun, tantangan besar muncul ketika event log yang digunakan sangat beragam dan mencerminkan berbagai variasi proses. Hal ini menyebabkan model proses yang dihasilkan menjadi terlalu kompleks (spaghetti model) dan sulit dianalisis [2]. Salah satu solusi untuk mengatasi hal ini adalah dengan menerapkan teknik clustering, yang bertujuan mengelompokkan kasus-kasus serupa agar analisis process mining menjadi lebih terfokus dan representatif [3].

Algoritma K-Means merupakan salah satu metode clustering yang banyak digunakan dalam segmentasi data proses. Penelitian oleh Cirne et al. menunjukkan bahwa penerapan K-Means sebelum tahap discovery dapat menyederhanakan model awal dan meningkatkan nilai fitness dalam evaluasi conformance [4]. Pramudia et al. juga membuktikan bahwa pendekatan ini dapat memperbaiki hasil model proses secara signifikan dibandingkan pendekatan non-segmentasi [5]. Selain itu, Kurniati et al. menunjukkan bahwa penerapan clustering dalam konteks kesehatan dapat memberikan model yang lebih tajam dan mendalam, meskipun terdapat trade-off terhadap aspek generalisasi dan kesederhanaan [3].

Dalam konteks pengelolaan denda lalu lintas, proses administratif melibatkan banyak jalur berbeda, seperti pengajuan banding, pemberitahuan, hingga pembayaran. Hal ini menjadikan dataset Road Traffic Fine Management sebagai objek studi yang ideal untuk

mengevaluasi efektivitas segmentasi berbasis clustering terhadap kualitas model process mining. Penelitian ini berfokus pada penerapan K-Means clustering sebelum tahap discovery dengan algoritma Inductive Miner, untuk mengukur dampaknya terhadap metrik evaluasi model proses seperti fitness, precision, generalization, dan simplicity.

1.2. Tujuan

1. Menganalisis efektivitas K-Means dalam meningkatkan hasil dari proses model.
2. Mengevaluasi hasil evaluasi model process mining (fitness, precision, generalization, simplicity) pasca-clustering.

2. PENELITIAN TERDAHULU

Penelitian-penelitian sebelumnya telah banyak membahas kombinasi antara teknik process mining dan clustering untuk meningkatkan kualitas model proses. Penelitian terhadap data pengaduan pelanggan dari UWV dengan menggunakan dua algoritma discovery, yaitu Alpha Miner dan Inductive Miner. Hasil evaluasi menunjukkan bahwa Inductive Miner memberikan skor fitness yang lebih tinggi (0.93) dibandingkan Alpha Miner (0.48), namun dari sisi precision dan generalization, Alpha Miner justru lebih unggul masing-masing dengan skor 0.79 dan 0.83 [5]. Hal ini menunjukkan adanya trade-off antara kelengkapan dan ketepatan dalam pemodelan proses.

Selain itu terdapat penelitian yang mengusulkan pendekatan gabungan antara algoritma α-Algorithm dengan teknik clustering k-means. Penelitian ini menunjukkan bahwa penerapan k-means sebelum tahap discovery dapat mengurangi kompleksitas model awal dan meningkatkan nilai fitness hingga 6% sampai 30% tergantung pada jenis proses yang diteliti [4]. Clustering membantu memisahkan jalur proses yang beragam dan mengurangi pengaruh noise terhadap hasil model, sehingga menghasilkan representasi proses yang lebih baik.

Sementara itu, pada penelitian yang menggunakan data pasien menggunakan SimpleKMeans dalam konteks process mining untuk analisis lintasan penyakit. Penelitian ini berhasil meningkatkan nilai fitness meskipun hanya sebesar 0.0006, namun peningkatan precision cukup signifikan sebesar 0.2600. Penurunan nilai generalization dan simplicity masing-masing sebesar 0.0754 dan 0.007 mengindikasikan bahwa model yang dihasilkan menjadi lebih spesifik namun kurang general dan sederhana [3].

Berdasarkan ketiga penelitian tersebut, dapat disimpulkan bahwa integrasi teknik clustering sebelum tahap process discovery memberikan dampak positif terhadap performa model, khususnya dalam konteks data yang kompleks dan heterogen. Oleh karena itu, studi ini mengadopsi pendekatan serupa dengan memanfaatkan algoritma Inductive Miner serta metode clustering K-Means untuk mengevaluasi peningkatan performa model proses pada dataset pengelolaan denda lalu lintas.

3. BAHASAN UTAMA

3.1. Dataset

Dataset *Road Traffic Fine Management Process* merupakan kumpulan data log peristiwa (*event log*) nyata yang berasal dari sistem informasi manajemen denda pelanggaran lalu lintas di Italia. Dataset ini dipublikasikan oleh Eindhoven University of Technology melalui repositori 4TU.ResearchData dan telah banyak digunakan dalam penelitian proses bisnis [2].

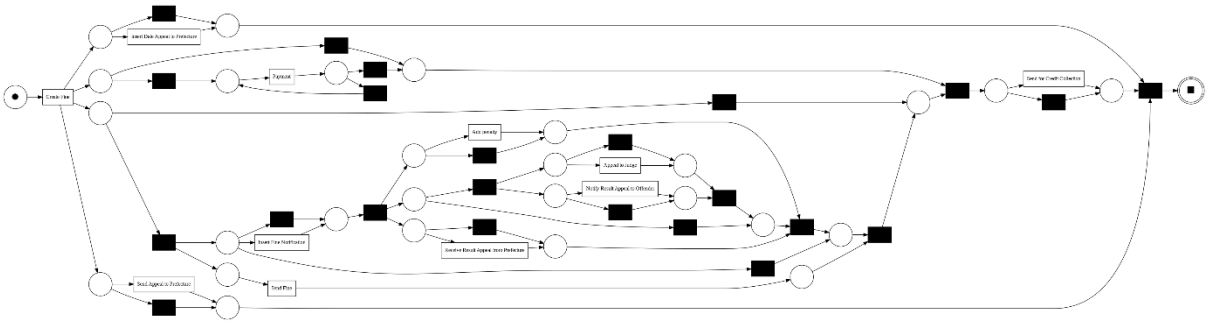
Dataset ini merekam alur proses penanganan denda lalu lintas, mulai dari saat denda dikeluarkan hingga penyelesaiannya, termasuk aktivitas-aktivitas seperti pengiriman surat pemberitahuan, pembayaran, atau eskalasi ke pengadilan. Periode data mencakup dari 1 Januari 2000 hingga 18 Juni 2013, dengan jumlah total sekitar 150.370 kasus (cases) yang masing-masing mewakili satu proses penanganan denda.

3.2. Data Preprocessing

Data processing dilakukan dengan memfilter atribut-atribut yang relevan untuk process mining yaitu activity, resource, dan timestamp.

3.3. Process Discovery

Setelah menentukan atribut yang akan digunakan, proses model akan dibuat dengan menggunakan inductive miner yang ditampilkan dengan petri net, berikut hasilnya.



3.4. Conformance Checking

Analisis konformansi menghubungkan peristiwa dalam *event log* dengan aktivitas-aktivitas dalam *process model* untuk mengidentifikasi kesesuaian dan perbedaan di antara keduanya. Teknik ini merupakan tahapan akhir dari penelitian dan dievaluasi berdasarkan empat dimensi kualitas sebagai berikut [6]:

1. **Fitness**

   Fitness mengukur sejauh mana perilaku dalam *event log* dapat direpresentasikan oleh *process model*. Nilai fitness berada dalam rentang 0 hingga 1, di mana nilai 1 menunjukkan bahwa model proses dapat sepenuhnya mereproduksi setiap trace dalam event log melalui model yang dihasilkan, dan sebaliknya.

2. **Precission**
   Precision mengukur sejauh mana model dapat merepresentasikan perilaku proses yang dijelaskan dalam event log tanpa memberikan kemungkinan perilaku yang terlalu luas (oversimplification). Nilai precision berada antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa model memiliki tingkat presisi yang baik.
3. **Generalisasi**
   Generalisasi mengukur kemampuan model dalam menangani variasi yang mungkin tidak muncul dalam event log. Nilai yang mendekati 1 menunjukkan bahwa model mampu menangkap semua kemungkinan variasi. Formula untuk menghitung generalisasi adalah sebagai berikut:
4. **Simplicity**
   Simplicity mengukur seberapa mudah model dapat dipahami oleh manusia, dan secara langsung berkaitan dengan kompleksitas model. Nilainya berkisar antara 0 hingga 1, di mana nilai 1 menunjukkan bahwa model bersifat sederhana, sementara nilai lebih rendah menunjukkan model yang kompleks dengan banyak elemen.

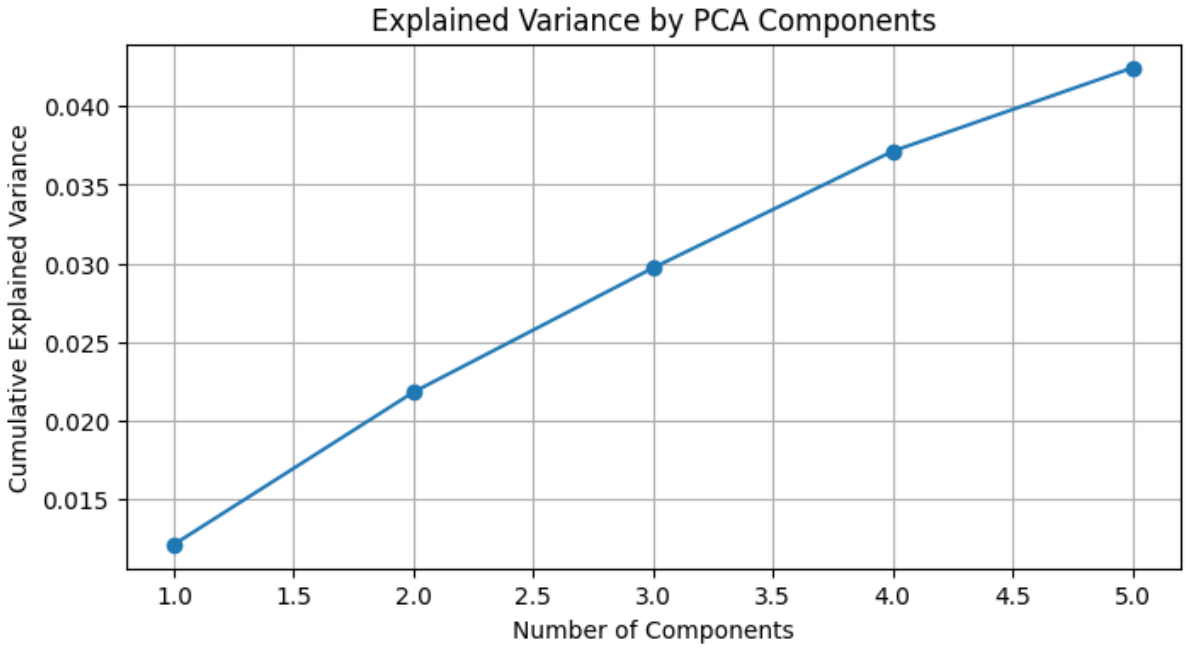
3.5. Data Preprocessing for Clustering

Clustering dilakukan untuk mengelompokan data dengan akurat dengan salah satu pendekatannya ialah dengan metode k-means[7]. Setelah itu preprocessing dilakukan untuk mengoptimalkan hasil clustering [8] salah satunya seperti memilih kolom akan dipakai dalam hal ini dibuat atribut duration yaitu durasi dari suatu trace, lalu activity count yaitu jumlah activity dari suatu trace dan lain-lain. Lalu, didapatkan 8 atribut yang dipakai yaitu amount, resource, dismissal, concept:name, vehicleClass, points, duration, activityCount, berikut gambarannya:

	case:concept:name	amount	org:resource	dismissal	concept:name	vehicleClass	points	duration	activityCount
0	A1	35.0	561	NIL	Create Fine > Send Fine	A	0.0	134	2
1	A100	106.5	561	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	971	5
2	A10000	110.0	561	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	550	5
3	A10001	110.0	537	#	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	189	6
4	A10004	110.0	537	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	741	5
...	...	...	...	...	...	...	...	...	...
150365	V9995	393.0	25	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	490	5
150366	V9996	131.0	25	NIL	Create Fine > Send Fine > Payment	A	0.0	60	3
150367	V9997	393.0	25	NIL	Create Fine > Send Fine > Insert Fine Notifica...	M	0.0	490	5
150368	V9998	393.0	25	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	490	5
150369	V9999	393.0	25	NIL	Create Fine > Send Fine > Insert Fine Notifica...	A	0.0	490	5

150370 rows × 9 columns

Setelah itu, dilakukan encoding untuk data kategorikal dan standarisasi untuk data numerikal. Lalu, analisis PCA dilakukan untuk menentukan atribut yang paling berpengaruh dalam data [9], hasil PCA bisa dilihat gambar.



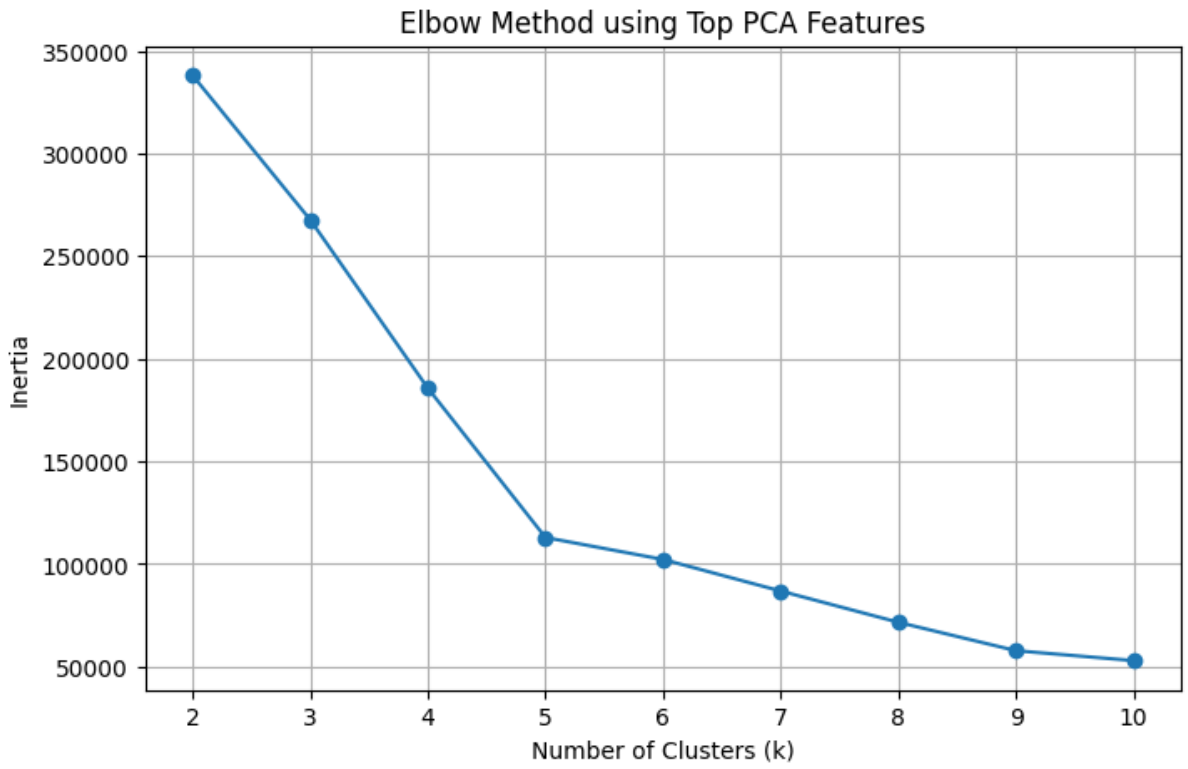
Dengan hasil dari masing-masing fitur sebagai berikut:

```
Top contributing features to PC1:
                                          PC1 Weight
activityCount                               0.461652
duration                                    0.461390
last_activity_Send for Credit Collection    0.447051
last_activity_Payment                       0.367965
amount                                      0.279067
last_activity_Send Fine                     0.171190
last_activity_Send Appeal to Prefecture     0.147546
points                                      0.135959
dismissal_#                                 0.133552
dismissal_NIL                               0.125738
```

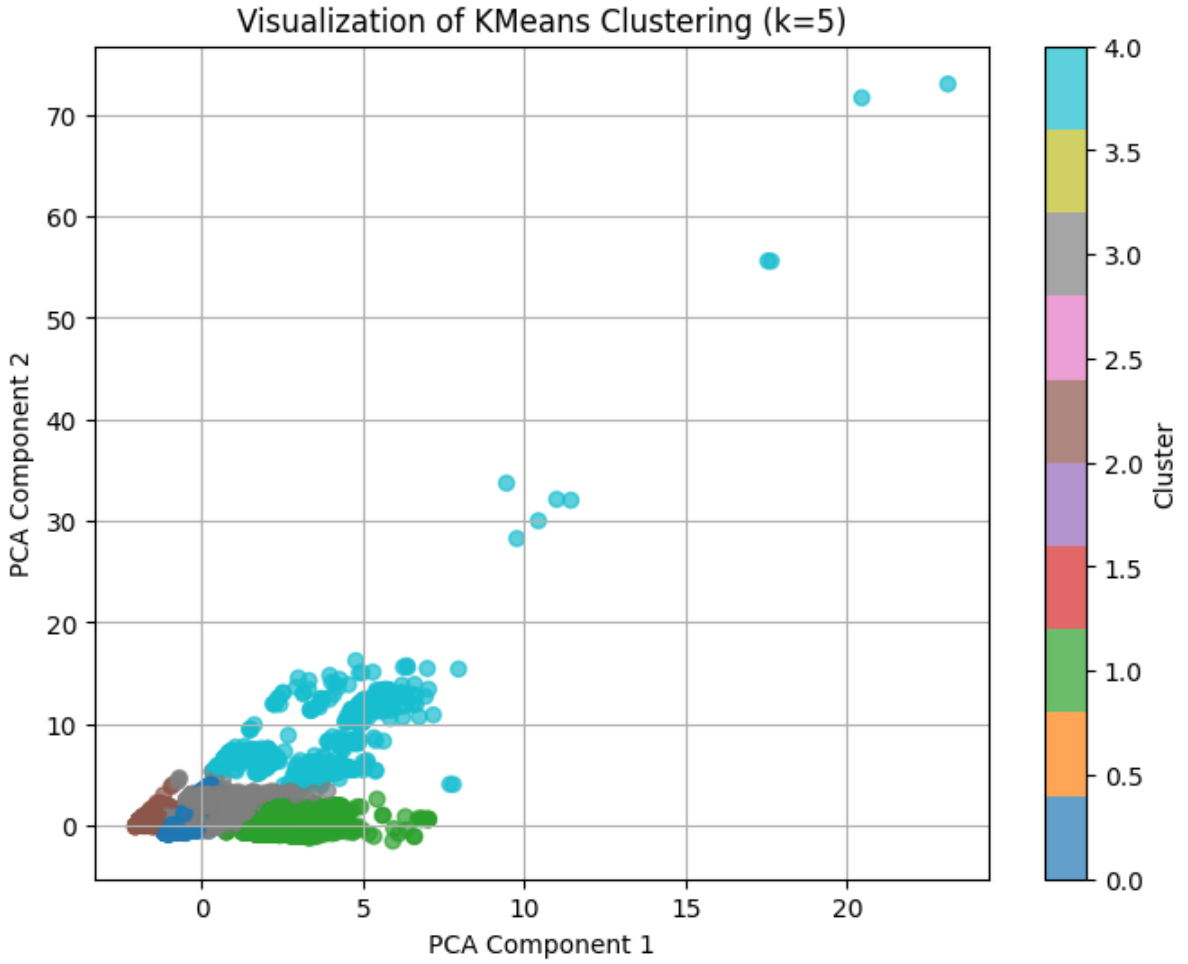
Dipilih fitur dengan PC1 Weight > 0.2.

3.6. K-Means

Setelah mendapatkan atribut untuk melakukan metode elbow dan silhoutte score dapat dilakukan untuk menentukan nilai k terbaik [10] dengan hasil sebagai berikut:



Dapat dilihat berdasarkan Elbow Method bahwa k=5 adalah nilai k terbaik.

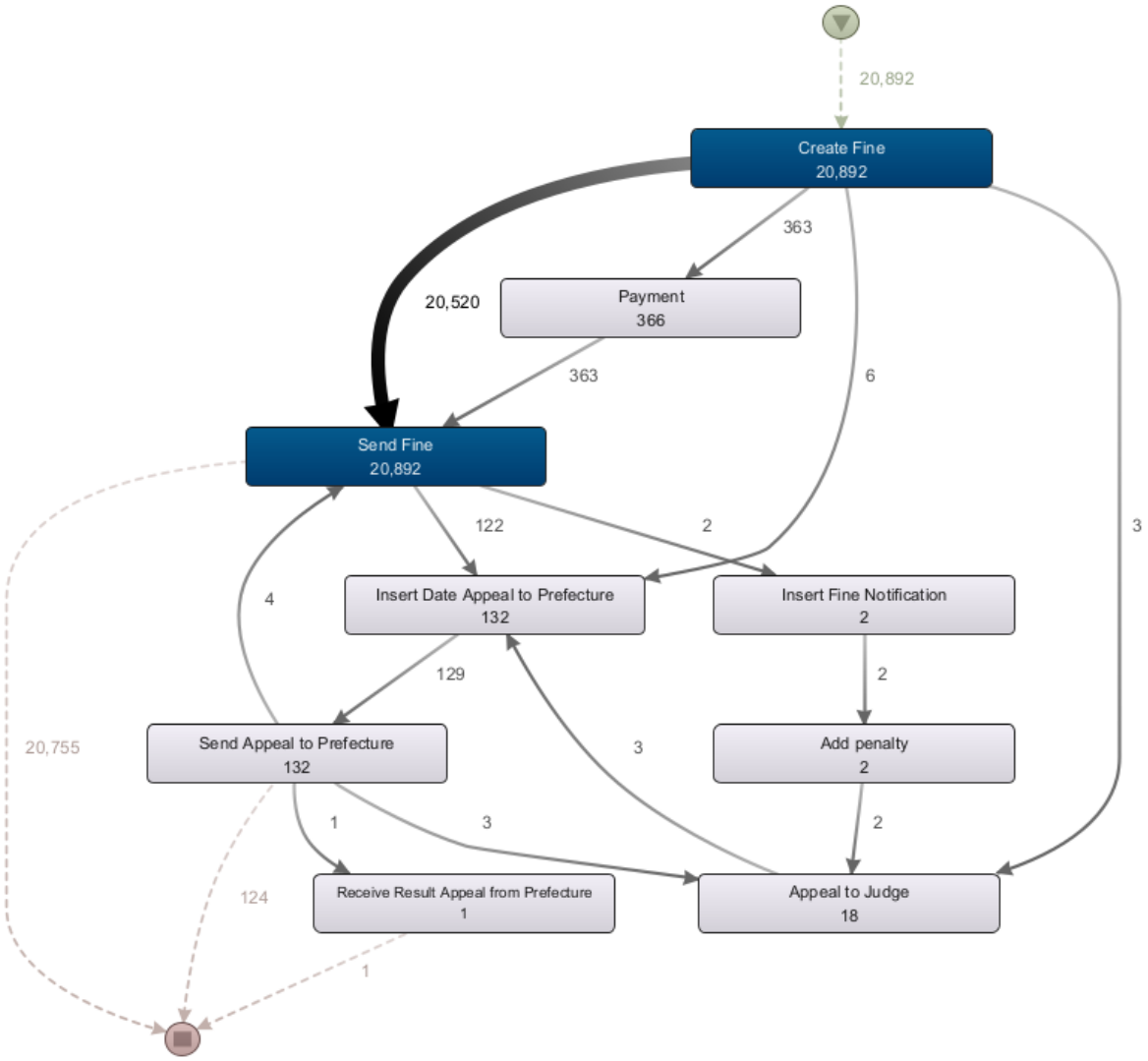


Dapat terlihat walaupun untuk beberapa bagian K-Means bisa memisahkan data, namun untuk terdapat beberapa data noise yang tidak bisa ditangani dengan baik oleh K-Means.

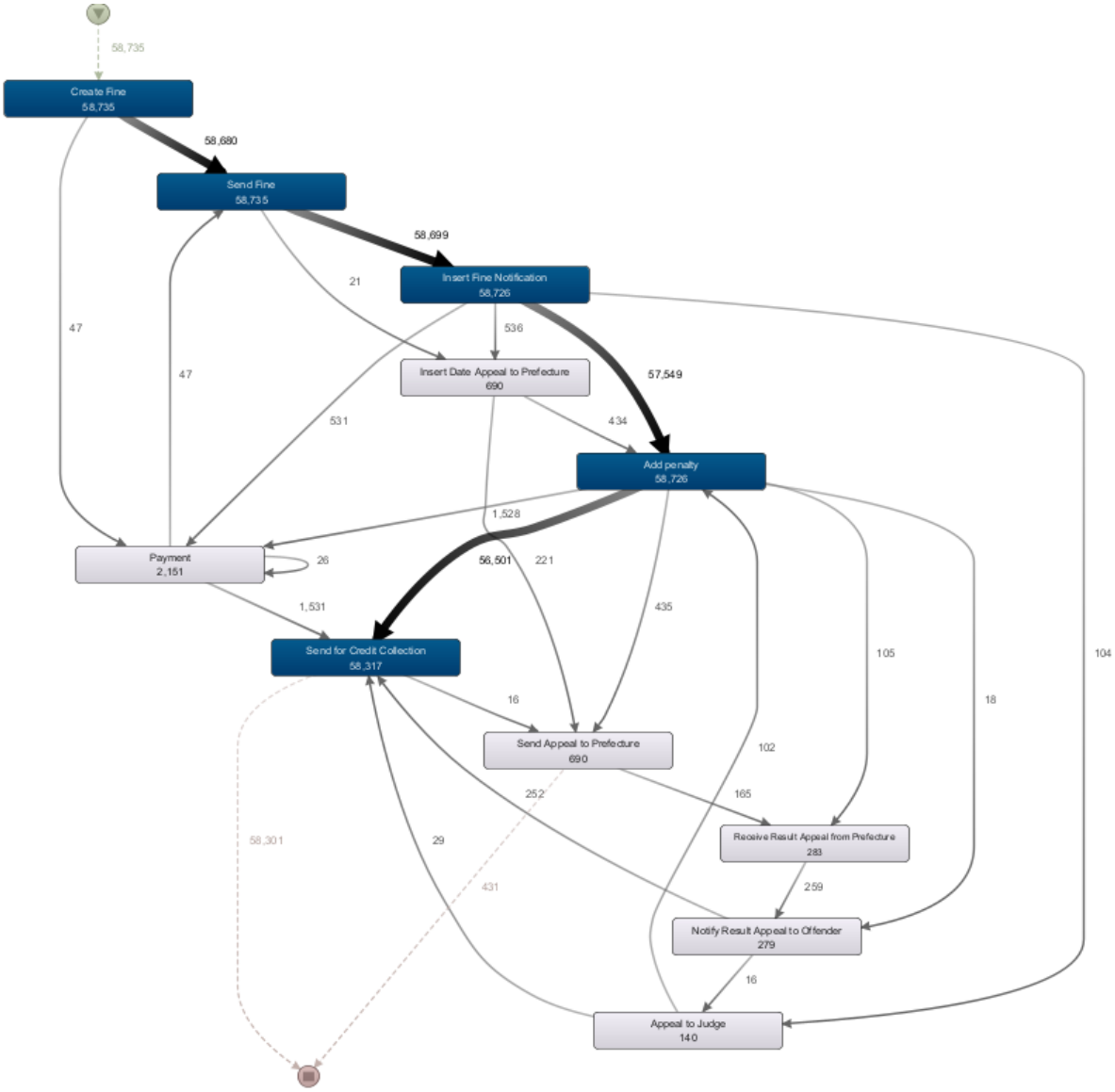
Tabel 3.6.1 Karakteristik Log dengan Cluster

Nama	Aktivitas Awal	Aktivitas Terakhir	Jumlah Trace	Rata-rata Durasi
0	Create Fine	Send Fine, Send Appeal to Prefecture, Appeal to Judge, dan Receive Result Appeal from Prefecture.	20892	13,8 Minggu
1	Create Fine	Send Fine, Payment, Send Appeal to Prefecture, dan Appeal to Judge	59735	23 Bulan
2	Create Fine	Payment, Create Fine, dan Send Fine.	49577	5 Hari
3	Create Fine	Payment, Send Appeal to Prefecture, Receive Result Appeal from Prefecture, Notify Result Appeal to Offender, dan Appeal to Judge.	20277	48,7 Minggu
4	Create Fine	Send for Kredit Collection, Payment, Send Appeal to Prefecture, Appeal to Judge, Receive Result Appeal from Prefecture, dan Notify Result Appeal to Offender.	889	19,8 Bulan

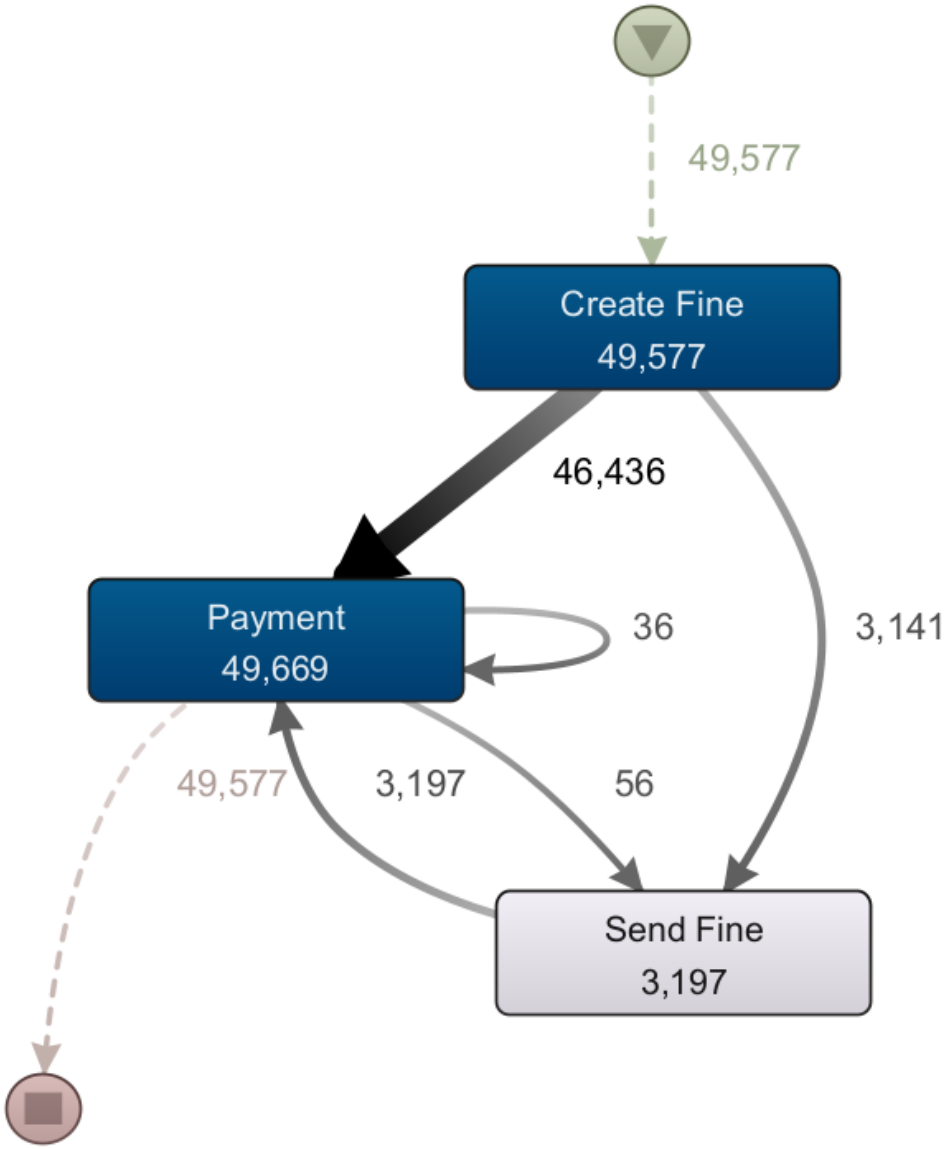
Berikut adalah process model dari masing-masing cluster, dengan 100% activities dan path 50%:
Cluster 0:


20,892
Create Fine
20,892
363
20,520
Payment
366
363
6
Send Fine
20,892
122
2
3
4
Insert Date Appeal to Prefecture
132
Insert Fine Notification
2
129
2
20,755
Send Appeal to Prefecture
132
3
Add penalty
2
1
3
2
124
Receive Result Appeal from Prefecture
1
Appeal to Judge
18
1

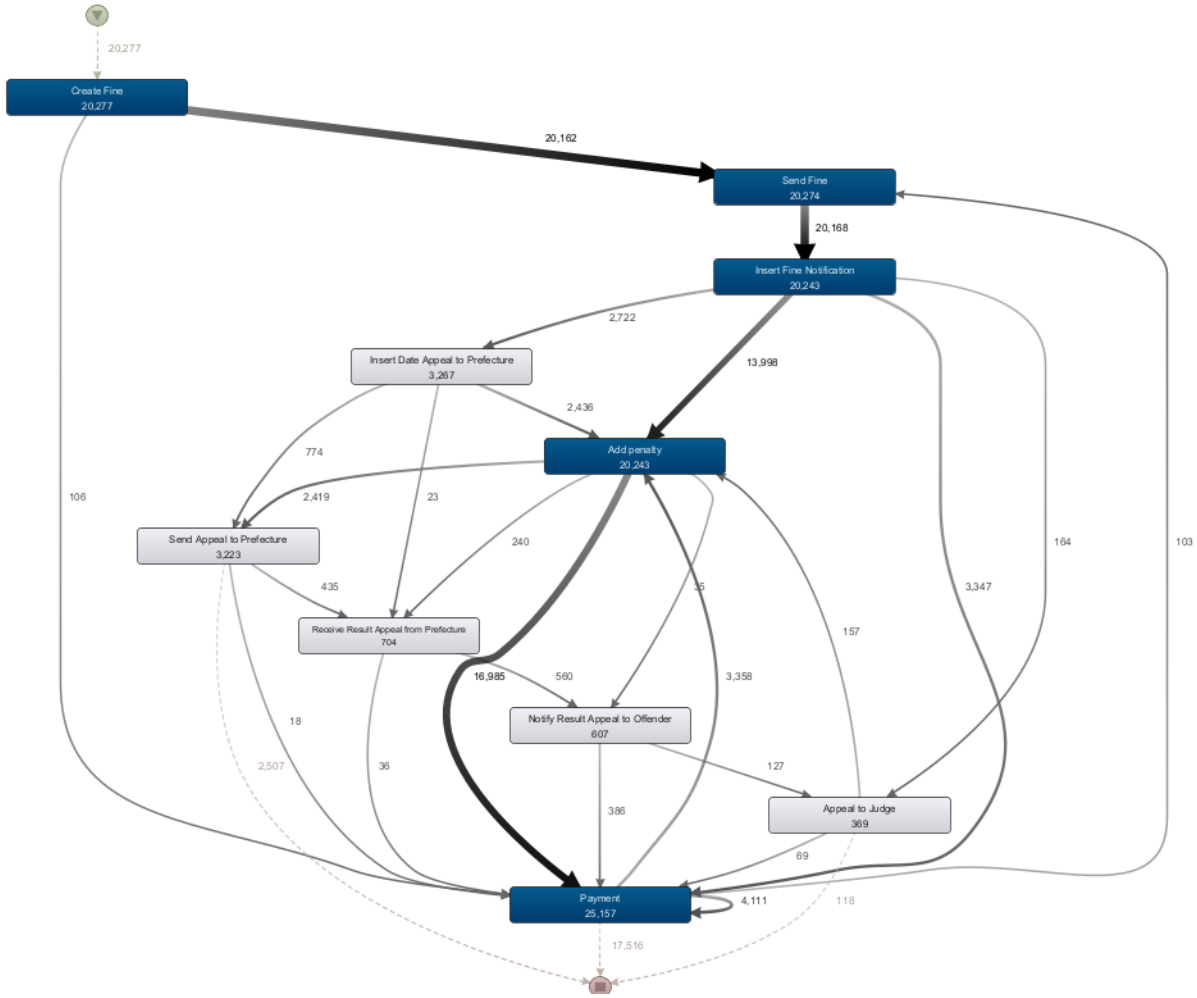
Cluster 1:


58,735
Create Fine
58,735
58,680
Send Fine
58,735
58,699
Insert Fine Notification
58,726
21
47
536
47
Insert Date Appeal to Prefecture
690
57,549
531
434
Add penalty
58,726
1,528
Payment
2,151
26
56,501
221
1,531
435
Send for Credit Collection
58,317
105
104
16
18
Send Appeal to Prefecture
690
102
252
165
58,301
29
Receive Result Appeal from Prefecture
283
431
259
Notify Result Appeal to Offender
279
16
Appeal to Judge
140

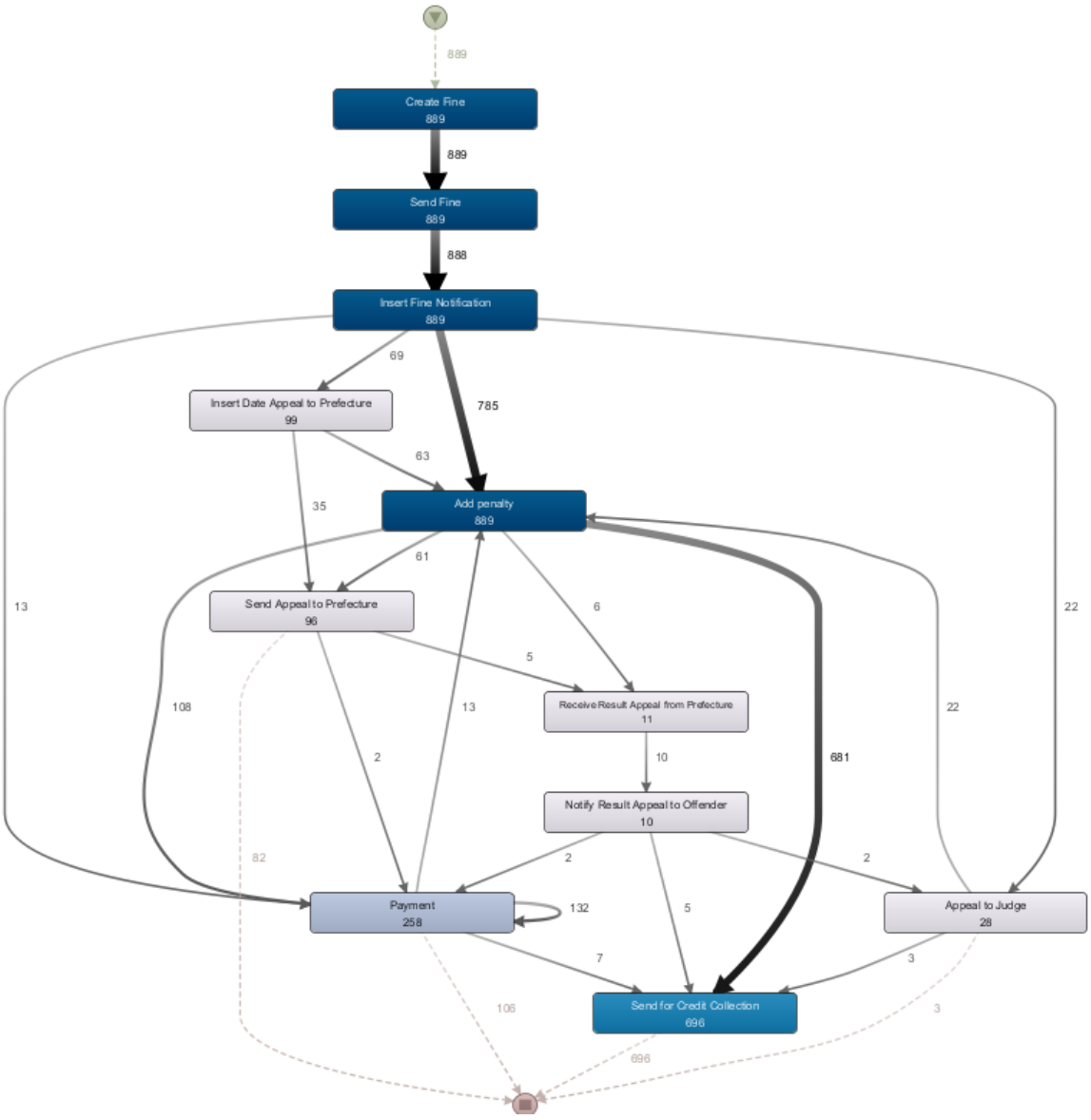
Cluster 2:


49,577
Create Fine
49,577
46,436
Payment
49,669
36
3,141
49,577
3,197
56
Send Fine
3,197

Cluster 3:


20,277
Create Fine
20,277
20,162
Send Fine
20,274
20,168
Insert Fine Notification
20,243
2,722
13,998
Insert Date Appeal to Prefecture
3,267
2,436
774
Add penalty
20,243
106
2,419
23
240
164
103
Send Appeal to Prefecture
3,223
435
35
3,347
Receive Result Appeal from Prefecture
704
157
16,985
560
3,358
Notify Result Appeal to Offender
607
18
2,507
36
127
386
Appeal to Judge
369
69
Payment
25,157
4,111
118
17,516

Cluster 4:



3.7. Conformance Checking for Clustering

Setelah mendapatkan hasil dari clustering dilakukan conformance checking untuk mengevaluasi hasil yang didapatkan, didapatkan hasil sebagai berikut:

	cluster	fitness	precision
0	cluster_0	{'perc_fit_traces': 100.0, 'average_trace_fitn...	0.709269
1	cluster_1	{'perc_fit_traces': 100.0, 'average_trace_fitn...	0.483808
2	cluster_3	{'perc_fit_traces': 100.0, 'average_trace_fitn...	0.713992
3	cluster_2	{'perc_fit_traces': 100.0, 'average_trace_fitn...	0.999842
4	cluster_4	{'perc_fit_traces': 100.0, 'average_trace_fitn...	0.691441

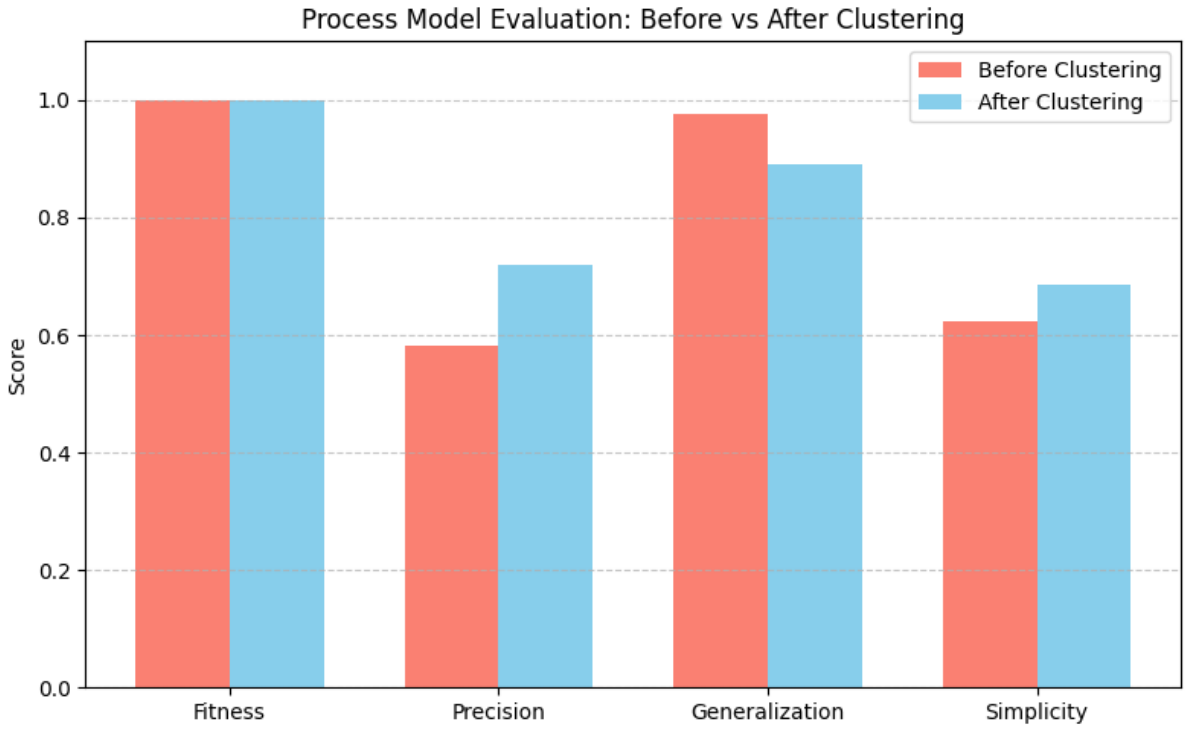
	generalization	simplicity
0	0.809383	0.671642
1	0.931377	0.626374
2	0.917322	0.634409
3	0.976658	0.846154
4	0.813896	0.643836

Dengan rata-rata untuk masing-masing metrik sebagai berikut:

```
Average Precision: 0.7196705452324237
Average Generalization: 0.8897273017090551
Average Simplicity: 0.6844826964322285
```

```
average_trace_fitness': 1.0,
```

Dan jika dibandingkan dengan hasil sebelumnya maka didapatkan visualisasi sebagai berikut:



Berdasarkan hasil ini didapatkan nilai fitnes tetap, dan terdapat kenaikan pada precision sebesar 0,137 (dari 0,582 menjadi 0,719) , dan simplicity sebesar 0,061 (dari 0,623 menjadi 0,684) serta penurunan di generalization sebesar 0,086 (dari 0,975 menjadi 0,889).

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Berdasarkan hasil tersebut, dapat disimpulkan bawah clustering berhasil meningkatkan performa dari model proses terutama dalam precision meskipun terdapat penurunan di generalization hal ini menunjukan dengan clustering model proses yang didapat lebih spesifik sehingga terjadi penurunan di generalization.

4.2. Saran

Saran untuk penelitian selanjutnya ialah, untuk mengeskplor fitur-fitur yang bisa dibuat untuk digunakan

DAFTAR PUSTAKA

[1] W. Van der Aalst, *Process mining: Data science in action*. Springer Berlin Heidelberg, 2016. doi: 10.1007/978-3-662-49851-4.

[2] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Balanced multi-perspective checking of process conformance," *Computing*, vol. 98, no. 4, pp. 407–437, Apr. 2016, doi: 10.1007/S00607-015-0441-1.

[3] A. P. Kurniati *et al.*, "Patient Clustering to Improve Process Mining for Disease Trajectory Analysis Using Indonesia Health Insurance Dataset," *2024 7th International Conference on Artificial Intelligence and Big Data, ICAIBD 2024*, pp. 88–93, 2024, doi: 10.1109/ICAIBD62003.2024.10604436.

[4] R. Cirne, C. Melquiades, R. Leite, E. Leijden, A. MacIel, and F. B. D. L. Neto, "Data Mining for Process Modeling: A Clustered Process Discovery Approach," *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, pp. 587–590, Sep. 2020, doi: 10.15439/2020F95.

[5] R. G. Pramudia, R. Ariandi, F. S. Salma, and R. Andreswari, "Process Mining Analysis and Implementation on Customer Complaints Dataset," *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, pp. 30–34, 2022, doi: 10.1109/ICSINTESA56431.2022.10041581.

[6] A. Marshall, I. Mcchesney, Z. Tariq, A. T. S. Ireddy, and S. V Kovalchuk, "An Experimental Outlook on Quality Metrics for Process Modelling: A Systematic Review and Meta Analysis," *Algorithms 2023, Vol. 16, Page 295*, vol. 16, no. 6, p. 295, Jun. 2023, doi: 10.3390/A16060295.

[7] A. Kumar, "Customer segmentation of shopping mall users using k-means clustering," *Advancing SMEs Toward E-Commerce Policies for Sustainability*, pp. 248–270, Dec. 2022, doi: 10.4018/978-1-6684-5727-6.CH013.

[8] Y. Chen, H. Zhou, J. Chen, Y. Sari, P. B. Prakoso, and A. R. Baskara, "On Some Data Pre-processing Techniques For K-Means Clustering Algorithm," *J Phys Conf Ser*, vol. 1489, no. 1, p. 012029, Mar. 2020, doi: 10.1088/1742-6596/1489/1/012029.

[9] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Apr. 2016, doi: 10.1098/RSTA.2015.0202,.

[10] A. A. Aktas, O. Tunali, and A. T. Bayrak, "Comparative unsupervised clustering approaches for customer segmentation," *Proceedings - 2021 2nd International Conference on Computing and Data Science, CDS 2021*, pp. 530–535, Jan. 2021, doi: 10.1109/CDS52072.2021.00097.

LAMPIRAN

Link Folder Repository Paper Referensi:

https://drive.google.com/drive/folders/1KR32AFU0xNFzuoPya9wTxgwoncVzv4Ru?usp=sharing

Link Folder File Tugas Besar:
https://drive.google.com/drive/folders/1klm7AGGxe3JGtNiLNwl84gTyD_1wG_OF?usp=sharing