KLUSTERISASI LOG AKTIFITAS PADA SISTEM KEAMANAN FILE DENGAN PENDEKATAN DATA MINING

Agus Pamuji[1], Heri Satria Setiawan[2]

Fakultas Ushuluddin, Adab dan Dakwah, IAIN Syekh Nurjati Cirebon[1]
Fakultas Teknik dan Ilmu Komputer, Universitas Indraprasta PGRI[2]

email: [1)]agus.pamuji@syekhnurjati.ac.id, herisatria@gmail.com[2]

Abstrak

Sistem file merupakan area atau sumber daya yang bisa dianggap kritis disampping memiliki sensivitas yang tinggi. Sistem file berkaitan dengan manajemen pengguna dan dimonitor melalui log riwayat aktiftas didalamnya. Pengguna dapat berinteraksi penuh dalam sistem file yang dilengkapi dengan izin dan haknya. Aktiftas pada sistem file akan menjadi lebih kompleks dan berkaitan dengan masalah keamanan dalam sistem file. Dengan demikian, perlindungan terhadap sistem file menjadi perhatian utama. Dalam studi ini, akan dilakukan analisis tterhadap log riwayat aktifitas pengguna dalam sistem file dengan pendekatan data mining. Adapun metode yang diterapkan dalam analisisnyaa menggunakan metode klustering. Tujuan padaa klustering adalah mengelompokan atau segmentasi pengguna terkait aktifitas pada sistem file. Teknik K-Means diterapkan pada metode klustering dan disajikan hasil kinerja dengan mengidentifikasi adanya temuan 5 kluster dalam log aktifitas. Dalam upaya memastikan model atau metode klustering dengan teknik K-Means itu akurat, maka metode Davies Boulding dan F-Measure digunakan untuk memastikan kualitas akurasi pada metode klustering. Oleh sebab itu, metode klustering dengan teknik K-Means merupakan metode yang dianggap cukup baik dalam mengelompokan data pengguna terkait dengan aktiftas pada sistem file.

Kata Kunci: Data Mining, Klustering, K-Means, Sstem File, Keamanan File

Abstract

The file system is an area or resource that can be considered critical in addition to having high sensitivity. The file system is related to user management and is monitored through the activity history log in it. Users can interact fully in the file system equipped with their permissions and rights. Activities on the file system will become more complex and related to security issues in the file system. Thus, the protection of the file system is a major concern. In this study, an analysis of user activity history logs in the file system will be carried out using a data mining approach. The method applied in the analysis uses the clustering method. The purpose of clustering is to group or segment users regarding activities on the file system. The K-Means technique is applied to the clustering method and the performance results are presented by identifying the findings of 5 clusters in the activity log. In an effort to ensure that the model or clustering method using the K-Means technique is accurate, the Davies Boulding and F-Measure methods are used to ensure the quality of accuracy in the clustering method. Therefore, the clustering method with the K-Means technique is a method that is considered quite good in grouping user data related to activities on the file system.

Keywords: *Data Mining, Clustering, K-Means, File System, File Security*

1. PENDAHULUAN

Hampir setiap saat, sesorang atau pengguna berinteraksi dengan teknologi terutama pada teknologi informasi [1] . Dengan interaksi pada informasi, pengguna akan selalu punya peluang membuat, memodifikasi, menghilangkan data atau informasi pada perangkat teknologi informasi [2].

Pada sistem file pada umumnya dilengkapi dengan fitur monitoring [3]. Pemberlakukan

monitoring dilakukan terhadap akun atau pengguna ketika terhubung dengan sistem file. Tujuan monitoring adalah sebagai upaya kendali terhadap pengguna dalam trafik manajemen pengguna pada sistem file.

Sehubungan dengan aktifitas didalam sistem file, maka perlu ada pembatasan. Hal ini dilakukan sebagai upaya perlindungan, pencegahan pada data. Pembatasan ini penting seabagai upaya antisipasi terhadap data yang dianggap sensitif. Oleh sebab itu, setiap pengguna diberi izin dalam mengakses sistem file dan atribut wewenangnya [4].

Sistem file yang ada saat ini menjadi rumit dan banyak melibatkan pengguna yang terhubung didalamnya. Apabila melihat dataset sementara pada log aktifitas pengguna terjadi peningkatan aktifitas. Peningkatan aktifitas ditelusuri dan diobservasi ada kecendrungan aktifitas dan izin diluar batas. Penyebabnya adalah permintaan dari pengguna kepada admin dalam pengaksesan [5].

Secara umum, dataset diklasifikasi terdiri dari dua bagian yaitu dataset publik dan dataset private. Dataset publik banyak dipakai dalam penelitian khususnya data mining dikarenakan skala komparable. Sedangkan dataset private adalah dataset yang ada pada kasus tertentu dan tidak dipublikasi. Dataset private biasanya tersedia pada instansi atau objek terkait secara khusus. Dalam kasus ini, dilakukan investigasi terhadap dataset yang berisi log aktifitas pengguna termasuk dalam kategori dataset private [6]. Data mining, *Knowledge Discovering Database* (KDD) merupakan proses menyaring, menambang data atau informasi untuk mendapatkan pengetahuan. Kemampuan dasar dengan kekuatan data pada data mining, memiliki dua kategori utama yaitu bagaimana data mining bisa menelusuri karakteristik, deskripsi dari data dan informasi secara rinci. Dengan demikian kemampuan seperti ini adalah *descriptive mining* [7]. Kedua, bagaimana data mining bisa mengenal, mengidentifikasi pola berdasarkan data sebagai temuan yang melibatkan variabel lain, dinamakan *predictive mining*. Dengan demikian, kasus pada studi ini mengacu pada penemuan informasi dan data dengan metode klustering yang termasuk pada *descriptive mining*[8].

Dengan dua basis kemampuan diatas, maka tujuan dari data mining yang diinterpretasikan pada kasus ini ada tiga. Pertama, mampu menjelaskan beberapa aktfitas observasi pada log riwayat aktifitas pengguna pada sistem file, dikenal *Explonatory*. Kedua, membuat konfirmasi suatu hipotesis yang telah ada dan jika tersedia, namun kasus ini tidak mengakomodasi, disebut *Confirmatory*. Ketiga, menganalisis datangnya data baru pada suatu domain relasi yang dianggap anomali, data log riwayat pengguna sistem file sudah teridentifikasi, dinamakan *Exploratory* [9].

Konsep data mining dengan metode clustering diterapkan untuk menganalisis log data aktifias pengguna. Penganalisisan pada log aktifitas ini akan ditemukan pola dan pengetahuan terhadap data yang besar dinamakan konsep data mining. Data mining yang diterapkan pada kasus ini akan menggunaka algoritma Clustering. Algoritma ini termasuk dalam kelompok Unspervised Learning yaitu K-Means. Kinerja K-Means mengelompokan data kedalam beberapa kelompok data lain pada sistem partisi.

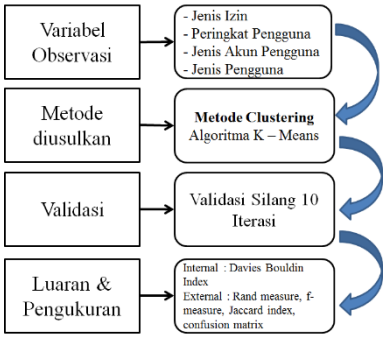
Motivasi utama pada kasus ini menghasilkan analisis yang digunakan untuk menganalisis dan mengelompokan pengguna terhadap aktifitas pada sistem file. Tidak hanya mengelompokan namun mengukur kinerja model untuk memastikan apakah model atau metode klustering dengan K-Means itu akurat. Sistem file saat ini terkonfimasi memuat informasi melekat pada akun pengguna sistem file dan pemberian hak akses ke data dan riwayat aktivitas pengguna [10].

Berikut ini adalah beberapa kontribusi utama dari metode dan pendekatan penelitian yang diusulkan. Kontribusi yang dilakukan antara lain upaya untuk mengimprovisasi teknik dan metode keamanan sistem file, memprediksi ancaman log aktifitas berlebihan, mengidentifikasi potensi aktifitas berlebihan melalui teknik clustering, dan meningkatkan efisiensi dan efektivitas pada skala keamanan file

2. LANDASAN TEORI

Dalam kasus ini dilakukan analisis terhadap data log aktitiftas pengguna dalam sistem file. Adapun kami akan mengadopsi dengan pendekatan data mining. Algoritma Clustering menggunakan k-means sebagai salah satu teknik atau metode didalamnya. Dengan demikian, konsep data mining akan mendeskripsikan proses yang terjadi ketika menganalisis log data aktifitas sistem file.

Studi pembahasan analisis log riwayat aktifitas pengguna dalam sistem file dapat disajikan dalam bentuk kerangka kerja [11]. Kerangka kerja pada studi dapat disajikan pada gambar dibawah ini.



Gambar 1. Kerangka kerja diusulkan

Faktor – faktor penentuan analisis, ada empat variabel yang diobservasi dalam analisis log aktifitas. Pertama, Jenis Izin (type permission) mendeskripsikan varian izin terkait pada koneksi pada file system. Jenis izin terdiri dari full control dimana pengguna memiliki akses penuh terhadap file sistem, pengguna memiliki semua akses kecuali dapatmenghapus file yaitu Grant, pengguna yang memiliki file sekaligus pembuatnya yaitu Owner,pengguna hanya dapat membukaa file tertentu yaitu Read, dan pengguna tidak hanya membuka namun melakukan modifikasi yaitu write.

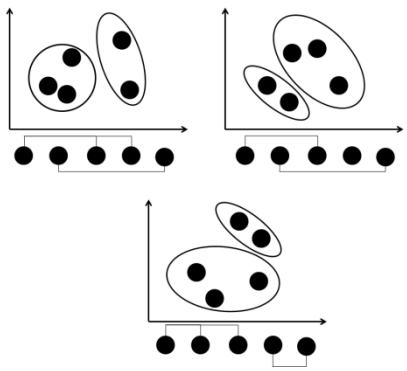
Kedua, peringkat pengguna dalam sistem file (yaitu 1, 2, dan seterusnya). Ketiga, jenis akun pengguna meliputi pengguna hanya sebagai tamu atau pendatang (Guest), pengguna dengan akses regular, pengguna yang memiliki akses penuh (super user account, dan agen dari system secara otomatis (System). Keempat, jenis pengguna dalam sistem file diantaranya adalah pengguna pemula (beginner user), pengguna yang senior (Intermediate User) dan pengguna yang sudah cukup ahli (expert user).

Log riwayat aktifitas disesuaikan dengan jenis studi yang berkaitan dengan data. Profil dataset termasuk jenis private. Sejumlah data diobservasi dan menelusuri dari data log riwayat aktifitas. Ada terdapat 2300 log data riwayat aktifitas. Perolehan dataset yang sudah dikumpulkan dari berbagai sumber masih dalam lingkup sistem file. Selain jumlah data yang cukup besar, terdapat 4 atribut atau sebagai variabel observasi. Adapun data yang ditelusuri didapat pada periode dari tahun 2020 sampai tahun 2021. Dengan demikian, pengumpulan data ini termasuk menghabiskan waktu yang cukup lama.

Data mentah, dataset log riwayat aktifitas yang diunduh dan didapatkan berbagai sumber sistem file. Data tersebut belum secara langsung diproses dalam keperluan analisis. Selanjutnya, dataset ini akan diproses dalam tahap persiapan data. waktu yang dibutuhkan lebih lama dibandingkan saat pengumpulan dataset. Data mentah (*raw data*) sekaligus termasuk sampel dan dataset. Kasus ini menggunakan *simple random sampling.*

Metode clustering, dilakukan mempartisi dataset. Bentuk cluster dibuat berdasarkan pada kemiripan. Proses penyimpanan dalam bentuk representasi cluster juga dilaksanakan seperti centroid, dan diameter. Oleh sebab itu, data akan menjadi lebih efektif jika data pada kluster bukan hanya pada pengukuran. Metode kluster berpotensi memiliki hirarki kluster. Selain itu, dapat disimpan pada struktur pohon indeks multi demensi. Dengan demikian, terdapat beberapa pilihan klustering dan algoritma klustering.

Teknik K-Means dalam metode klustering memiliki kinerja mengelompokkan dengan data bervolume besar dan waktu relatif cepat. Penentuan awal kluster merupakan tahap awal menjadi kendala. Inisiasi nilai centroid diawal menjadi penyebab dalam pembentuk kluster [12]. K-Means melakukan klustering secara berjangka. Kondisi ini merujuk pada sistem kerja Partitioned Clustering. Dengan Demikian K-Means merupakan pengklusteran secara sederhana.



Gambar 2. Urutan kerja teknik K-Means

3. METODE PENELITIAN

Tahapan kerja dengan metode kluster melalui teknik K-Means ini memiliki 6 fase. Pertama, jumlah kluster yang terjadi dapat

ditentukan diawal pada variabel k. Kedua, pembentukan nilai centroid (k) secara random. Ketiga, jarak setiap data terhadap masing centroid dihitung. Adapun proses kalkulasi dengan dengan menggunakan persamaan korelasi antar dua objek (*Eucledia Distance*). Keempat, mengelompokan setiap data dengan mengacu pada jarak paling dekat antara data dengan centroid. Kelima, centroid baru (k) sudah bisa ditentukan melalui kalkulasi rata – rata dari data pada centroid yang sama. Keenam, langkah ketiga dapat dijalankan kembali apabila posisi centroid baru memiliki ketidaksamaan dengan centroid lama [13].

Dikonfirmasi, kluster merupakan kumpulan data dimana apabila ada objek data yang terletak didalam kluster harus memiliki kemiripan. Bagi data yang tidak berada dalam satu kluster tidak mimiliki kemiripan [14]. Sebuah "n" objek pengamatan dengan "p" variabel, maka terlebih dahulu menentukan ukuran kedekatan sifat antar data [15]. Standar data yang bisa di gunakan adalah analisis jarak dengan *Euclidean distance*, antar dua objek dari *P* dimensi pengamatan. Sebagai objek pertama yang diamati adalah X = [x1,x2,.....xp] dan Y =[y1,y2,.....yp].

$$D_{(x,y)} = \sqrt{\sum_{j=1}^{p}(xj - yj)^2} \qquad (1)$$

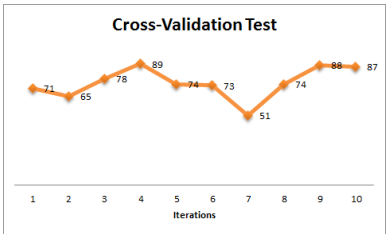
Keterangan informasi persamaan diatas adalah "d" merupakan jarak antara titik pada data x dan titik data posisi y, dimana x = x1,x2,.....xi dan y = y1,y2,.....yi dan "j" mendeskripsikan nilai atribut serta "p" dianggap sebagai dimensi atribut [16].

4. HASIL DAN PEMBAHASAN

Data mining, dengan metode klustering pada kasus analisis log riwayat sistem berkas membuktikan kinerjanya. Proses data mining diawali dengan validasi terhadap dataset. Teknik validasi silang adalah metode yang direkomendasikan dengan kinerja terbaik. Selanjutnya, proses kluster pada dataset akan diberikan pada setiap data pengguna terkait dengan aktifitas pada sistem file. Luaran dari metode kluster adalah bentuk kluster yang sudah terbentuk. Bagian akhir akan menjelaskan evaluasi kinerja model atau metode k-means dengan teknik Davis Bouldin dan F Measure.

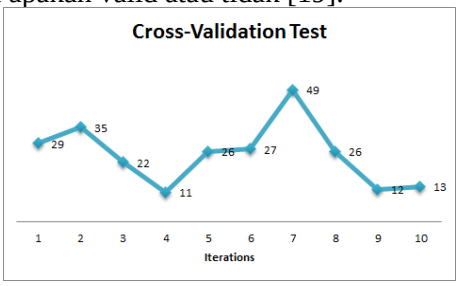
4.1. Validasi Silang

Pada metode validasi silang merupakan metode dalam pengolahan data statistik yang dapat diterapkan dalam mengevaluasi kinerja model atau algoritma [17]. Tidak hanya pada bidang statistik namun bidang eksak lain menggunakan validasi silang. Kemampuan dalam mengevaluasi model atau algoritma menjadi karakteristik validasi silang (cross validation). Terkait dengan kasus data mining keamanan, penggunaan validasi silang, data akan dibagi menjadi dua bagian. Pertama, data latih dan data uji. Ada sekelompok data yang dilatih dan ada yang di uji dengan sebaran perbandingan 80% data latih dan 20% data uji. Dalam upaya mengurangi waktu komputasi dengan tetap menjafa keakuratan maka validasi silang menjadi pilihan dan rekomendasi. Teknik validasi silang akan dilakukan sebanyak k sebagai iterasi [18].



Gambar 3. Hasil Pengukuran Akurasi Benar dalam %

Dengan melihat hasil analisis validasi silang pada akurasi benar, indikasi menyatakan cukup baik. Alur grafik cenderung stabil jika ditinjau dari pergerakan data. iterasi pertama dapat dianggap cukup baik dengan nilai 71%. Kondisi yang sama mengalami kenaikan meskipun diawali penurunan yang tidak signifikan. Angka kenaikan terus melaju sampai pada puncak. Penurunan terjadi pada iterasi 7 dan distabilkan denganpeningkatan angka akurasi diiterasi 10. Dengan demikian, validasi silang sangat baik untuk memastikan data apakah valid atau tidak [19].



Gambar 4. Hasil Pengukuran Akurasi Salah dalam %

Berdasarkan gambar hasil uji validasi silang, akurasi nilai salah menunjukan fase atau iterasi pertama dinilai cukup tinggi. Fase berikut mengalami penurunan walaupun meningkat tidak signifikan. Iterasi ke 4 menjadi penurunan yang baik namun ketika iterasi selanjutnya mengalami kenaikan dengan puncak hampir 50% pada iterasi ke 7. Posisi ke 8 berperan sebagai penanda penurunan nilai kesalahan hampir mendekati pada iterasi ke 4. Dengan demikian rata – rata kesalahan sebesar 25 % [20].

Berikut ini akan disajikan hasil validasi data latih dan data uji. Dalam penelitian ini terdapat 256 data training dan 25 data testing. Dalam pemrosesannya terdapat 2 klasifikasi yaitu klasifikasi benar dan klasifikasi salah. Dengan demikian dari keduanya akan ditunjukan nilai akurasi seperti dibawah ini [21].

Tabel 1. Hasil Uji Cross Validation

Uji	Data Uji	Data Latih	Benar	Salah
1.	256	25	17	8
2.	256	25	22	3
3.	256	25	23	2
4	256	25	15	10
5	256	25	19	6
6	256	25	24	1
7	256	25	15	10
8	256	25	22	3
9	256	25	17	8
10	256	25	19	6

Hasil validasi yang menerapkan 10 iterasi pengujian disajikan pada tabel diatas memberikan hasil bahwa terdapat 256 data training dan 25 data testing. Dengan demikian dapat diberikan dua nilai akurasi yaitu akurasi benar dan akurasi salah [22].

3.2 Kluster dan penentuan jarak tiap kluster

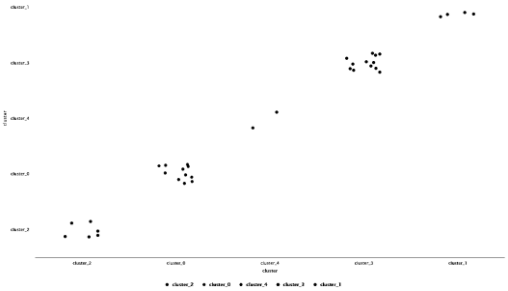
Perhitungan jarak antara data dengan centroid dapat ditentukan dengan persamaan Euclidean Distance. Hasil perrhitungan dapat disajikan pada tabel dibawah ini. Perhitungan jarak dilakukan sebanyak 25 data. Terdapat 5 kluster yang memberikan informasi jarak dari tiap pengguna pada centroid sebagaimana tabel dibawah ini [23].

Tabel 2. Pembentukan Kluster

Atribut	K1	K2	K3	K4	K5
PLN_GM_01	0,195	0,136	0,825	0,921	1,388
PLN_GM_02	1,372	0,277	0,792	0,578	0,592
PLN_GM_03	1,726	0,168	1,388	1,828	1,598
SALES_04	1,253	0,365	0,554	1,383	0,917
SALES_05	0,579	1,727	1,775	0,828	0,224
CUST_ADMIN_01	1,973	1,131	0,412	1,177	1,552
CUST_ADMIN_02	0,956	1,437	0,358	0,563	0,856
NOC_ADM_01	0,758	1,254	1,143	0,847	1,398
NOC_ADM_02	0,494	1,562	0,983	0,742	0,859
SLS_RTL_01	1,226	0,642	1,112	1,226	1,358
SLS_RTL_02	0,689	1,885	0,357	1,228	1,673
MARK_01	0,892	1,541	1,236	0,676	1,339
BSN_01	1,588	1,153	1,344	1,995	1,773
BSN_02	1,846	0,398	0,494	0,371	1,135
MARK_04	0,581	0,373	1,918	1,819	0,618
MARK_05	0,981	1,774	0,817	0,568	1,178

Pengelompokan data terkait dengan kluster dilakukan. Kelompok kluster suatu data dapat ditentukan dari jarak terpendek data pada tabel terhadap kluster. Contoh, pengguna pertama, memiliki jarak 0,195 pada kluster 1. 0,136 dimiliki kluster 2. Kluster 3 memiliki 0,825. Kluster 4 memiliki 0,921 dan 1,388 pada kluster 5. Berdasarkan pada kelima kluster tersebut, data pengguna 1 atau pertama memiliki jarak terpendek dengan kluster 2. Dengan demikian, data pengguna 1 masuk dalam kluster 2.

Selanjutnya, disajikan ringkasan hasil analisis dengan k-means terkait dengan log riwayat file sistem. pada kluster pertama jenis akun pengguna menempati peringkat pertama. Jenis akun pengguna lebih dominan dibandingkan dengan aktifitas pengguna, dan status. Dikluster 2, aktifitas pengguna hampir mendominasi penempatan kluster dibanding dengan variabel lain. Hal yang sama dengan kluster 4 bersama – sama dengan memiliki aktifitas pengguna terbanyak. Sedangkan kluster 3 dan 5 memiliki kesamaan yaitu tipe pengguna hampir diatas 50% pada posisi yang sama.



Gambar 5. Hasil klustering dengan K-Means

3.3 Evaluasi kinerja model

Ada dua metode yang digunakan pada evaluasi model dengan metode klustering. Pertama yaitu Davies Bouldin dan F-Measure. Davies-Bouldin Index pertma kali diusulkan oleh oleh David L. Davies and Donald W. Bouldin pada tahun 1979. Adapun bentuk parameter adalah Sum-of square within cluster (SSW) sebagai metrik kohesi dalam sebuah cluster. Separasi dengan Sum-of-square-between-cluster (SSWB) dengan mengukur jarak antara centroid C_i dan C_j.

$$SSW = \frac{1}{N}\sum_{I=1}^{N}\left\|x_i - C_{pi}\right\|^2 \quad (2)$$

$$SSB = \frac{2}{M(M-1)}\sum_{I=1}^{M}\sum_{j=1,j\neq 1}^{M}\left\|C_i - C_j\right\|^2 \quad (3)$$

Rumus R dan DBI

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (4)$$

$$DBI = \frac{1}{K}\sum_{I=1}^{K} max\left(R_{i,j}\right) \quad (5)$$

Davies Bouldin index (DBI) dimaksudkan sebagai skema evaluasi internal. Dengan DBI, dataset log riwayat aktifitas dilakukan kluster kemudian diukur. Metode DBI berperan sebagai pengukur kluster dalam dataset log riwayat ketika menentukan kualitas. Penentuan kadar klustering ditentukan berdasarkan rasio kluster pada diagram scatter. Apabila nilai pada kluster dataset tersebut makin kecil maka semakin baik.

Tabel 4. Hasil pengukuran model K-Means

Pengukuran	Hasil Observasi
Davis Bouldin	0,168
F Measure	87%

Berdasarkan tabel diatas, kinerja model dengan ukuran Davies Bouldin menunjukan keterangan analisis tingkat kepadatan dan jarak antara data dalam kluster [24]. Hasil dari DBI, memberikan keterangan bahwa model atau metode klustering dengan K-Means dengan kategori baik. Indikasinya adalah nilai hasil observasi mencapai 0,168. Semakin kecil semakin baik dengan rentang nilai terendah adalah 0 dan tertinggi adalah 1 [25]. Berbeda dengan F measure masih sama dalam konsep pengukuran akurasi. F-Measure merupakan gabungan antara recall dan precision. Recall berfungsi untuk membandingkan pada data antara positif benar dengan banyak data yang yang secara aktual posisitif. Precission merupakan perbandingan antara positif benar dengan banyak nya data yang diprediksi positif. Berdasarkan tabel diatas, dapat disimpulkan model dengan metode K-Means dapat dianggap baik dan akurat pada kasus terkait [26].

5. KESIMPULAN DAN SARAN

Kumpulan data dalam bentuk log riwayat aktifitas pengguna dianggap sebagai klustering. Data pada log riwayat aktifitas penggunas memiliki kemiripan. Karakteristik lainnya adalah berkaitan dengaan didalam kelompok yang sama. Alternatif lain punya ketidakmiripan. Dengan demikian, dataset pada log aktifitas pengguna cenderung memiliki kemiripan. Bentuk kemiripan berupa antara penguna satu dengan pengguna lain. Variabel lain seperti jenis akun memiliki hal yang sama. Analisis kluster pada kasus log aktifitas sudah dilakukan dan ditemukan ada 5 kelompok sebagai kluster.

Jika meninjau pada kualitas metode klustering khususnya pada teknik K-Means pada kasus ini bergantung pada tiga hal. Pertama, pengukuran kemiripan pada dataset log aktitiftas diukur dengan melihat kepadatan dan jarak pada tiap kluster. Kedua, implementasi analisis klustering pada log aktifitas dapat membantu identifikasi pola dan menemukan anomali data atau data yang tidak wajar. Dengan demikian, data yang tidak wajar ini bukan hanya anomali pada klustering tetapi bisa diprediksi bahaya anomali. Ketiga, klustering sebagai metode memiliki kelebihan dalam menemukan beberapa atau semua pola data pada kasus log aktifitas yang tersembunyi.

Tantangan dan kebutuhan metode klustering pada kasus ini bisa dimunculkan. Aspek skalabilitas dimana pada kasus log aktifitas pengguna masih berfokus pada pengambilan sampel. Meskipun metode klustering lebih cocok dengan data yang keccil seperti sampel namun bagaimana menangani data bervolume besar dan kecepatan tinggi. Data yang besar menjadi lebih kompleks dan peluang kluster bisa meningkat. Tantangan lain adalah kemampuan menangani jenis atribut yang berbeda. Pada kasus log aktifitas masih sebagian data dalam bentuk numerik sementara data dengan tipe biner, kategorikal, ordinal, atau bahkan gabungannya menjadi kebutuhan tersendiri. Selain itu, bagaimana menangani bentuk tipe data selain numerik sehingga harus dilakukan transformasi ke bentuk lain.

Dampak studi data mining dalam keamanan komputer atau informasi sangat menentukan. Keamanan komputer atau informasi berusaha melindungi aset yang bernilai tinggi. Sistem yang rumit akan sulit mendeteksi adanya kelemahan dari dalam (internal) dibandingkan dengan eksternal. Data mining hadir untuk menjadi kolaborasi. Peran data mining yaitu dengan menginvestigasi dan menelusuri lebih dalam berbasis pada data. Dengan metode klustering dapat dikelompokan potensi adanya bahaya atau ancaman secara internal jika diliat dari peringkat pengguna atau hasil klustering.

Studi data mining dan keamanan informasi atau komputer dimasa mendatang lebih fokus bagaimana menentukan peringkat bahaya pada aktifitas pengguna pada file sistem. Adapun yang menjadi perhatian adalah dengan peringkat bisa mendeteksi anomali lebih akurat ditunjang dengan teknik – teknik klustering yang lain. Selebihnya melakukan segmentasi terhadap pengguna dan aktifita

6. DAFTAR PUSTAKA

[1] N. Petrovi, V. Roblek, and N. Papachashvili, “Decision Support Based on Data Mining for Post COVID-19 Tourism Industry Decision Support Based on Data Mining for Post COVID-19 Tourism Industry,” in *International SAUM Conference on Systems, Automatic Control and Measurements*, 2021, no. September, pp. 1–5.

[2] M. R. Islam and K. M. Aktheruzzaman, “An Analysis of Cybersecurity Attacks against Internet of Things and Security Solutions,” *J. Comput. Commun.*, vol. 08, no. 04, pp. 11–25, 2020, doi: 10.4236/jcc.2020.84002.

[3] S. Contiu, S. Vaucher, R. Pires, M. Pasin, P. Felber, and L. Reveillere, “Anonymous and confidential file sharing over untrusted clouds,” *Proc. IEEE Symp. Reliab. Distrib. Syst.*, pp. 21–31, 2019, doi: 10.1109/SRDS47363.2019.00013.

[4] D. Iordache, “Database – Web Interface Vulnerabilities,” *Strateg. XXI - Secur. Def. Fac.*, vol. 17, no. 1, pp. 279–287, 2021, doi: 10.53477/2668-2001-21-35.

[5] J. T. McDonald, N. Herron, W. B. Glisson, and R. K. Benton, “Machine learning-based android malware detection using manifest permissions,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2020-Janua, pp. 6976–6985, 2021, doi: 10.24251/hicss.2021.839.

[6] C. Gao, X. Zhang, and H. Liu, “Data and knowledge-driven named entity recognition for cyber security,” *Cybersecurity*, vol. 4, no. 1, 2021, doi: 10.1186/s42400-021-00072-y.

[7] J. I. Zong Chen, “Automatic Vehicle License Plate Detection using K-Means Clustering Algorithm and CNN,” *J. Electr. Eng. Autom.*, vol. 3, no. 1, pp. 15–23, 2021, doi: 10.36548/jeea.2021.1.002.

[8] A. Supriyatna and W. P. Mustika, “Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 2, no. 2, p. 152, 2018, doi: 10.30645/j-sakti.v2i2.78.

[9] S. M. Toapanta, O. A. Escalante Quimis, L. E. Mafla Gallegos, and M. R. Maciel Arellano, “Analysis for the evaluation and security management of a database in a public organization to mitigate cyber attacks,” *IEEE Access*, vol. 8, no. 2, pp. 169367–169384, 2020, doi: 10.1109/ACCESS.2020.3022746.

[10] S. Pei, F. Nie, R. Wang, and X. Li, “Efficient clustering based on a unified view of k-means and ratio-cut,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, no. c, pp. 1–12, 2020.

[11] M. Syukri Mustafa, M. Rizky Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Citec J.*, vol. 4, no. 2, pp. 151–162, 2017.

[12] Z. kai Feng, W. jing Niu, R. Zhang, S. Wang, and C. tian Cheng, "Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization," *J. Hydrol.*, vol. 576, no. 12, pp. 229–238, 2019, doi: 10.1016/j.jhydrol.2019.06.045.

[13] B. Al Kindhi, T. A. Sardjono, M. H. Purnomo, and G. J. Verkerke, "Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis," *Expert Syst. Appl.*, vol. 121, no. 3, pp. 373–381, 2019, doi: 10.1016/j.eswa.2018.12.019.

[14] W. Yang, H. Long, L. Ma, and H. Sun, "Research on clustering method based on weighted distance density and k-means," in *Procedia Computer Science*, 2020, vol. 166, pp. 507–511, doi: 10.1016/j.procs.2020.02.056.

[15] K. Chowdhury, D. Chaudhuri, and A. K. Pal, "An entropy-based initialization method of K-means clustering on the optimal number of clusters," *Neural Comput. Appl.*, vol. 33, no. 12, pp. 6965–6982, 2021, doi: 10.1007/s00521-020-05471-9.

[16] Y. Feng, S. Zhao, and H. Liu, "Analysis of Network Coverage Optimization Based on Feedback K-Means Clustering and Artificial Fish Swarm Algorithm," in *IEEE Access*, 2020, vol. 8, pp. 42864–42876, doi: 10.1109/ACCESS.2020.2970208.

[17] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[18] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Deep android malware detection and classification," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1677–1683, 2017, doi: 10.1109/ICACCI.2017.8126084.

[19] J. Yu, L. Zhu, R. Qin, Z. Zhang, L. Li, and T. Huang, "Combining k-means clustering and random forest to evaluate the gas content of coalbed bed methane reservoirs," *Geofluids*, vol. 2021, no. 9, pp. 1–8, 2021, doi: 10.1155/2021/9321565.

[20] S. Wahyuningsih and D. R. Utari, "Perbandingan Metode K-Nearest Neighbor , Naive Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit," *Konf. Nas. Sist. Inf. 2018 STMIK Atma Luhur Pangkalpinang, 8 – 9 Maret 2018*, pp. 619–623, 2018.

[21] M. T. Rouabah, A. Tounsi, and N. E. Belaloui, "A mathematical epidemic model using genetic fitting algorithm with cross-validation and application to early dynamics of COVID-19 in Algeria," no. August, 2020, doi: 10.4314/jfas.v12i3.17.

[22] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee attrition prediction using deep neural networks," *Computers*, vol. 10, no. 11, pp. 1–11, 2021, doi: 10.3390/computers10110141.

[23] C. Oktarina, K. A. Notodiputro, and I. Indahwati, "Comparison of K-Means Clustering Method and K-Medoids on Twitter Data," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 189–202, 2020, doi: 10.29244/ijsa.v4i1.599.

[24] F. Tempola and A. F. Assagaf, "Clustering of Potency of Shrimp In Indonesia With K-Means Algorithm And Validation of Davies-Bouldin Index," vol. 1, no. Icst, pp. 730–733, 2018, doi: 10.2991/icst-18.2018.148.

[25] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1235, no. 1, pp. 1–7, 2019, doi: 10.1088/1742-6596/1235/1/012015.

[26] A. A. Vergani and E. Binaghi, "A soft davies-bouldin separation measure," *IEEE Int. Conf. Fuzzy Syst.*, vol. 2018-July, pp. 1–8, 2018, doi: 10.1109/FUZZ-IEEE.2018.8491581.