JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 8
No. 1
April 2022

Data Mining dengan Segmentasi Pengguna pada Keamanan Sistem *File*

Agus Pamuji[#1]

[#]*Bimbingan Konseling Islam, IAIN Syekh Nurjati Cirebon*
Jl Perjuangan By Pass Sunyaragi 45132 Kota Cirebon Jawa Barat
[1]agus.pamuji@syekhnurjati.ac.id

***Abstrak*— Salah satu sumber daya yang menjadi pertimbangan kritis adalah sistem file. Hampir semuanya terlibat dalam menghubungkan pengguna dengan sistem file. Manajemen pengguna, *file* dan konfigurasi akan menjadi fokus permasalahan jika dikaitkan dengan keamanan. Pengguna pada sistem *file* dianggap memiliki identitas ketika terhubung dengan sistem. Disamping itu, atribut izin dan hak yang ada pada pengguna sebagai pelengkap identitas. Saat ini terjadi peningkatan aktiftas dalam sistem *file* sehingga menjadi lebih kompleks . Sistem yang kompleks dan pengguna yang belum terkelola dengan baik maka berpotensi ancaman keamanan file. Dalam studi ini, telah dilakukan penelusuran dan investigasi pada aktivitas dengan log riwayat aktivitas pengguna dalam sistem *file* khususnya pendekatan *data mining* . Metode klustering ditujukan untuk menganalisis dengan menghasilkan luaran pengetahuan berupa kluster. Pembentukan kluster ditunjang dengan teknik K-Means. Hasil pengelompokan menjadi segmentasi terhadap pengguna pada sistem file. Hasil akhir merepresentasikan adanya 5 kluster pada teknik K-Means. Model dengan teknik K-Means terbukti menjadi model yang efektif dibuktikan dengan nilai akurasi pada metode Davies Bouldin Index (DBI). Tambahan pengukuran lain adalah dengan F- Measures untuk meninjau hasil akurasi penempatan kluster pada kasus dengan teknik K-Means. Dengan demikian, metode klustering dengan teknik K-Means merupakan metode yang dianggap handal ketika mensegmentasikan data pengguna terkait dengan aktivitas pada sistem file.**

Kata kunci*— *Data Mining* , *Klustering, K-Means*, Sistem *File*, Keamanan *File

I. PENDAHULUAN

Setiap hari akan ada pengguna yang berinteraksi dengan teknologi informasi [1]. Pengguna memiliki berbagai macam aktivitas. Sistem *file* yang ada semakin kompleks dan melibatkan banyak pengguna. Hampir semua aktfititas di dalam sistem *file* bisa memberikan peluang pada pengguna [2] melakukan membuat, perubahan, memindahkan data atau *file* yang ada pada perangkat komputer. Aktivitas dalam manajemen *file* sangat sensitif dan juga penuh dengan resiko apabila ditinjau dari aspek keamanan.

Monitoring sebagai tindakan preventif dalam upaya menjaga selain melindungi adanya tindakan yang berhubungan dengan penyalahgunaan. Dengan demikian, sistem *file* sangat penting memerlukan monitoring ketika berhadapan dengan sistem yang kompleks dan pengguna yang beresiko kacau. Setiap pengguna akan selalu diberi identitas tentang informasinya dan juga akses. Selebihnya, penerapan monitoring terhadap akun pengguna sistem *file* tetap diupayakan. Kendali terhadap keterlibatan pengguna dalam sistem *file* akan menjadi tujuan dari monitoring dalam sistem *file* oleh administrator [3].

Akses dan izin merupakan dua hal penting dalam sistem *file*. Pengguna pada umumnya diberikan dua hal tersebut. Tujuannya adalah untuk bisa berinteraksi dengan sistem atau dengan pengguna lain seperti berbagi informasi [4]. Pembatasan terhadap aktivitas pengguna harus dibatasi. Alasan utamanya adalah sebagai upaya perlindungan, pencegahan pada sumber daya data ata file. Pembatasan pada sisi akses dan izin pada pengguna dimaksudkan dalam upaya antisipasi terhadap data yang dianggap sensitif. Dengan demikian, pengguna diberi izin dalam mengakses sistem *file* dan atribut wewenangnya.

Sudah dipastikan, hampir semua sistem *file* memiliki sistem yang tidak hanya mulai menjadi rumit namun penuh dengan resiko penyalahgunaan. Tambahannya, dengan melibatkan pengguna yang banyak dan tersebar menjadi peluang menurunkan kinerja pengamanan jika ditinjau dari sisi internal. Kenyataannya adalah hampir pengamanan eksternal secara penuh dilakukan namun tidak perhatian pada aspek internal [5]. Sistem file, sebagai wadah berisi data dan aktivitas yang ada didalamnya menunjukan adanya aktivitas secara drastis meningkat [6].

Kecenderungan aktivitas meningkat akan diobservasi dan diinvestigasi. Berdasarkan hasil temuan dan identifikasi, hampir penyebabnya adalah karena ada permintaan mengenai akses dan izin dari beberapa pengguna. Kondisi ini menjadi tidak terkendali ditambah dengan konfigurasi manajemen pengguna dengan file.

Berdasarkan kondisi pada kasus dalam studi ini adalah dengan melakukan pendekatan dengan *data mining* . *Data mining* , *Knowledge Discovering Database* (KDD) merupakan proses menyeleksi, menggali data atau informasi untuk menemukan pengetahuan. Hal yang

mendasari kemampuan dengan kekuatan data pada *data mining* , memiliki dua kategori utama yaitu descriptive mining, *data mining* bisa menelusuri karakteristik, deskripsi dari data dan informasi secara rinci. Selanjutnya, bagaimana *data mining* bisa mengenal, mengidentifikasi pola berdasarkan data sebagai temuan yang melibatkan variabel lain, dinamakan predictive mining. Kasus pada studi ini mengacu pada penemuan informasi dan data dengan metode klustering yang termasuk pada descriptive mining [7].

Data mining , kemampuan menelusuri data, konsep ini memiliki tiga tujuan jika dikaitkan dengan kasus analisis log riwayat aktivitas pengguna. Segala aktivitas yang dilakukan pengguna dalam sistem *file* bisa dideskripsikan secara rinci. *Explanatory*, ditujukan pada tujuan ini dengan memberikan nilai atribut data dan informasi. Berdasarkan data yang dianalisis pada *data mining* , selanjutnya dikonfirmasi dengan hipotesis yang diajukan, *Confirmatory*. *Exploratory*, menganalisis, menginvestigasi dengan data yang baru pada domain relasi. Hasil pada investigasi ada kecenderungan anomali terhadap data log riwayat pengguna sistem file, *Exploratory*.

Menganalisis data pada bidang *data mining* membutuhkan usaha tinggi. Beberapa teknik dan metode *data mining* sudah dipersiapkan menghadapi berbagai tipe data. Sistem *file* yang ada memiliki sistem yang kompleks dan pengguna serta aktivitas beragam. Hampir semua aktivitas direkam dan hasil rekaman dapat ditelusuri dan menganalisis didalamnya. Dengan demikian, analisis pada log aktiftas pengguna terkait dengan sistem *file* akan menemukan pola. Jika pada kasus ini, maka akan menghasilkan luaran analisis dalam bentuk kluster. Luaran bentuk kluster mengggunakan metode klustering dan K-Means sebagai teknik yang diterapkan analisis data. Bentuk klustering mengelompokan berdasarkan atribut dengan nilai yang sama selain dianggap sebagai kelompok Unsupervised Learning. Selanjutnya, log data riwayat aktivitas pengguna dipartisi [8].

Dua jenis dataset yang mendasari analisis *data mining* bersamaan dengan kasus investigasi log riwayat aktivitas pengguna sistem file. Pertama, dataset bersifat publik. Kasus penyelesaian dan eksperimen pada dataset publik hampir banyak diterapkan. Komparable adalah salah satu keunggulan dari dataset publik. Walaupun dengan kelebihan dapat dibandingkan dengan yang lain, namun ada dataset yang tidak tersedia. Ada dataset berkaitan dengan kebutuhan dan masalah tertentu pada bidangnya. Dengan demikian, kedua Dataset private sebagaimana dengan fungsinya sebagai alat dan objek ditujukan pada kasus tertentu. Dalam kasus ini, dilakukan investigasi terhadap dataset yang berisi log aktivitas pengguna termasuk dalam kategori dataset private [9].

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

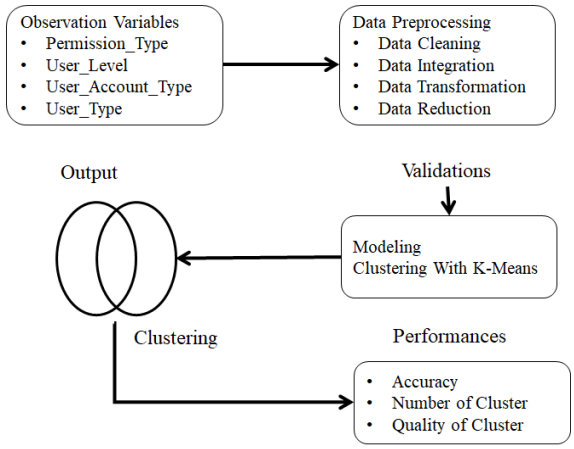
Kontribusi utama dari metode dan pendekatan penelitian yang diusulkan adalah upaya untuk mengimprovisasi teknik dan metode keamanan sistem file, memprediksi ancaman log aktivitas berlebihan, mengidentifikasi potensi aktivitas berlebihan melalui teknik clustering, dan meningkatkan efisiensi dan efektivitas pada skala keamanan file.

Dalam makalah ini akan diorganisasikan dimana terdiri dari pendahuluan mengenai sekilas keamanan *file* dan keterkaitan dengan *data mining*. Kedua, metode penelitian yaitu dengan memuat metode atau kerangka kerja yang diusulkan melalui teknik K-Means. Ketiga, pembahasan tentang kinerja model dan pengukuran model. Keempat, menyimpulkan sebagai inti dari pembahasan terkait dengan keamann *file* pada *data mining* .

II. METODE PENELITIAN

Metode *data mining* dalam kasus log aktivitas pengguna terhadap sistem *file* akan dijelaskan. Adapun teknik yang digunakan adalah dengan K-Means. K Means termasuk dalam metode klustering dalam *data mining* . Dengan demikian, konsep *data mining* akan membuat kluster proses yang terjadi ketika menganalisis log data aktivitas sistem file. Analisis *data mining* dapat dilakukan setelah adanya tahap persiapan data. tahapan persiapan data memakan waktu yang cukup lama. Persiapan data adalah tahap awal melakukan proses mining dengan metode kluster.

Studi mengenai analisis segmentasi log riwayat aktivitas pengguna dalam sistem *file* dapat disajikan dalam bentuk kerangka kerja. Kerangka kerja pada studi dapat disajikan pada gambar dibawah ini.



Gambar 1. Kerangka kerja metode *clustering – K-means*

A. *Variabel Observasi (Variables of Observation)*

Faktor – faktor penentuan analisis, ada empat variabel yang diobservasi dalam analisis log aktivitas . Pertama, Jenis Izin (*type permission*) mendeskripsikan varian izin terkait pada koneksi pada *file* system [11]. Jenis izin terdiri dari full control dimana pengguna memiliki akses penuh

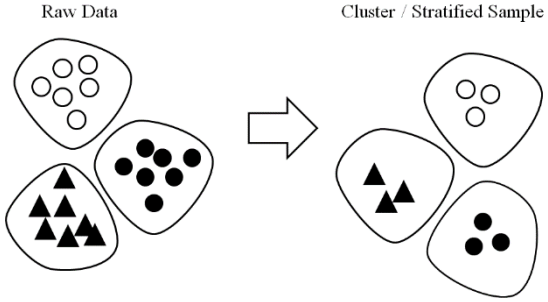
terhadap *file* sistem, pengguna memiliki semua akses kecuali dapatmenghapus *file* yaitu Grant, pengguna yang memiliki *file* sekaligus pembuatnya yaitu *Owner*,pengguna hanya dapat membukaa *file* tertentu yaitu Read, dan pengguna tidak hanya membuka namun melakukan modifikasi yaitu write [12].

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*) [13].

B. Dataset

Log riwayat aktivitas disesuaikan dengan jenis studi yang berkaitan dengan data. Profil dataset termasuk jenis private. Sejumlah data diobservasi dan menelusuri dari data log riwayat aktivitas . Ada terdapat 1560 log data riwayat aktivitas [14]. 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 [15].

Data mentah, dataset log riwayat aktivitas 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 teknik *Cluster / Stratified Sampling* [16].



Gambar 2. *Stratified sampling*

C. Data Preprocessing

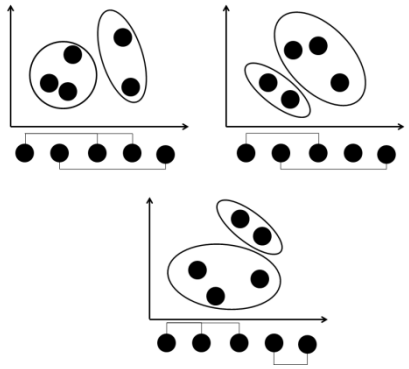
Tahap persiapan data merupakan tahap awal dan memakan waktu lama. Adapun beberapa langkah dan proses adalah pembersihan data, menghilangkan data yang dianggap noise dalam dataset log aktivitas . Selain itu mengisi data secara manual apabila ditemukan kosong. Merapihkan data apabila belum memiliki format benar, dan memperbaiki data tidak konsisten. Semua ini disebut *data cleaning* [17].

Tahap berikutnya yang dilakukan terhadap dataset log aktivitas adalah mereduksi dimensionalitas apabila terlalu kompleks . Adapun dalam mereduksi digunakan teknik *Feature Extraction* dan *Feature Selection* [18]. Data yang sudah direduksi dan cleaning, masih tetap harus diolah selanjutnya. Data yang ada pada dataset ada potensi tidak normal, sehingga harus dinormalkan yang disebut dengan data transformation. Langkah terakhir adalah menggabungkan data dari beberapa sumber untuk memudahkan analisis [19].

D. Pemodelan dan Klustering

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 clustering [20].

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 *cluster* [21]. K-Means melakukan klustering secara berjangka. Kondisi ini merujuk pada sistem kerja *Partitioned Clustering*. Dengan Demikian K-Means merupakan pengklusteran secara sederhana.



Gambar 3. Urutan kerja teknik *K-means*

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 (*Euclidean Distance*). Keempat, mengelompokan setiap data dengan mengacu pada jarak paling dekat antara data dengan centroid [22]. 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 [23].

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. Sebuah "n" objek pengamatan dengan "p" variabel, maka terlebih dahulu menentukan ukuran kedekatan sifat antar data. 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].

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

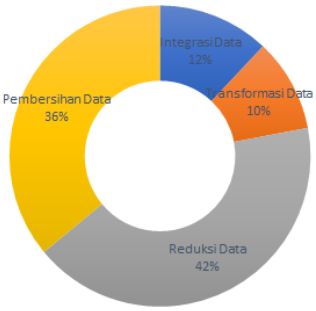
Keterangan informasi persamaan diatas adalah "d" merupakan jarak antara titik pada data x dan titik data posisi y, dimana x = x1,x2,.....xi dan y = y1,y2,.....yi dan "j" mendeskripsikan nilai atribut serta "p" dianggap sebagai dimensi atribut [22].

III. 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 aktivitas 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 [24].

A. *Persiapan Data*

Persiapan data merupakan usaha keras pada data masuk dalam tahap *preprocessing*, merubah, mengisi, menggabungkan dan mengkonversi data [25]. dapat disajikan dalam gambar di bawah ini hampir 40% dataset yang ada harus direduksi. Skala dimensionalitas adalah faktor penyebab utama. Data memiliki ukuran cenderung melebar dan varian yang tinggi sehingga harus disederhanakan melalui teknik reduksi. Berdasarkan analisis didapatkan 36% sebagai urutan kedua yaitu data cleaning dimana beberapa data bernilai kosong dan sebagian lain data tidak sesuai format atau *noise*.

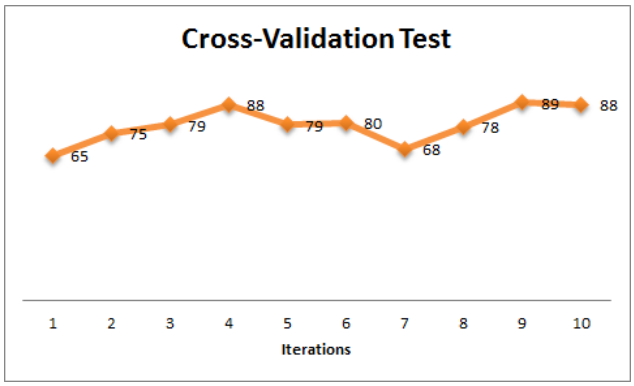


Gambar 4. Representasi *data preprocessing*

Data integration, sebagai penggabungan data dari berbagai macam sumber daya [26]. Pengumpulan data ini cukup sulit karena masih ada beberapa data terpisah belum bisa ditemukan secara cepat. Tambahannya, banyak melibatkan orang atau pengguna lain. Paling terkecil adalah bagaimana merubah bentuk data yang memiliki skala terkecil. Dengan data transformation, data yang tidak normal dilakukan normalisasi sehingga memudahkan dalam analisis *data mining* .

B. *Validasi Silang (Cross-Validation)*

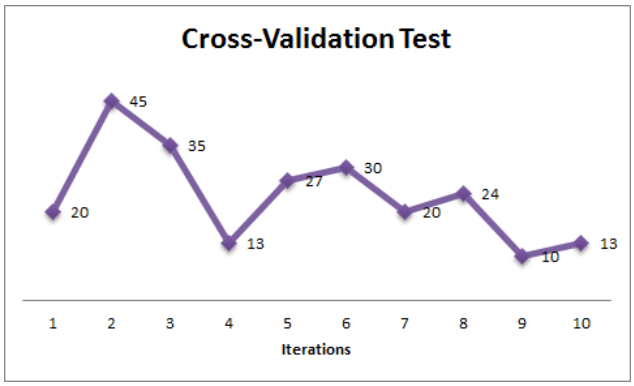
Pada metode validasi silang merupakan metode dalam pengolahan data statistik yang dapat diterapkan dalam mengevaluasi kinerja model atau algoritma [27]. Tidak hanya pada bidang statistik namun bidang eksak lain menggunakan validasi silang [28]. Kemampuan dalam mengevaluasi model atau algoritma menjadi 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 [29]. Ada sekelompok data yang dilatih dan ada yang di uji dengan sebaran perbandingan 75% data latih dan 25% data uji. Dalam upaya mengurangi waktu komputasi dengan tetap menjafa keakuratan maka validasi silang menjadi pilihan dan rekomendasi [20]. Teknik validasi silang akan dilakukan sebanyak k sebagai iterasi.



Gambar 5. Hasil pengukuran akurasi benar dalam %

Dengan melihat hasil analisis validasi silang pada akurasi benar, indikasi menyatakan cukup baik. Alur grafik cenderung naik jika ditinjau dari pergerakan data. iterasi pertama dapat dianggap cukup baik dengan nilai 65%. Kondisi yang sama mengalami kenaikan yang cukup

signifikan. Angka kenaikan terus melaju sampai pada puncak. Penurunan terjadi pada iterasi 7 dan distabilkan dengan peningkatan angka akurasi diiterasi 9. Dengan demikian, validasi silang sangat baik untuk memastikan data apakah valid atau tidak.



Gambar 6. Hasil pengukuran akurasi salah dalam %

Berdasarkan gambar hasil uji validasi silang, akurasi nilai salah menunjukan fase atau iterasi pertama dinilai cukup rendah. Fase berikut mengalami kenaikan yang signifikan. Iterasi ke 4 menjadi penurunan yang baik namun ketika iterasi selanjutnya mengalami kenaikan dengan puncak hampir 30% pada iterasi ke 6. Posisi ke 6 berperan sebagai penanda penurunan. Dengan demikian rata – rata kesalahan sebesar 23,7 %.

Berikut ini akan disajikan hasil validasi data latih dan data uji. Dalam penelitian ini terdapat 238 data training dan 68 data testing. Dalam pemrosesannya terdapat 2 klasifikasi yaitu klasifikasi benar dan klasifikasi salah. Dengan demikian dari keduanya akan ditunjukan nilai akurasi seperti dibawah ini.

TABEL I
HASIL UJI VALIDASI SILANG

Pengujian	Data Training	Data Testing	Klasifikasi Benar	Klasifikasi Salah
Iterasi 1	238	68	36	32
Iterasi 2	238	68	21	47
Iterasi 3	238	68	22	46
Iterasi 4	238	68	33	35
Iterasi 5	238	68	58	10
Iterasi 6	238	68	64	4
Iterasi 7	238	68	24	44
Iterasi 8	238	68	61	7
Iterasi 9	238	68	42	26
Iterasi 10	238	68	46	22

Hasil validasi yang menerapkan 10 iterasi pengujian disajikan pada tabel diatas memberikan hasil bahwa terdapat 256 data training dan 25 data testing. Rata – rata klasifikasi benar menunjukan 41% lebih besar dari data testing dengan nilai 27,3%. Dengan demikian dapat diberikan dua nilai akurasi yaitu akurasi benar dan akurasi salah.

C. Kluster dan Penentuan Jarak Tiap Kluster

Penentuan jarak menggunakan persamaan Euclidean Distance untuk diimplementasikan dalam perhitunga jarak antara centroid dengan beberapa data. Tabel dibawah ini adalah representasi dari hasil pengukuran. Ada 25 pengguna yang dilibatkan sehingga kalkulasi dilakukan sebanyak n = 25. Dengan demikian, 5 kluster telah terbentuk memberikan informasi jarak dari centroid pada pengguna.

TABEL II
JARAK DATA TIAP KLUSTER

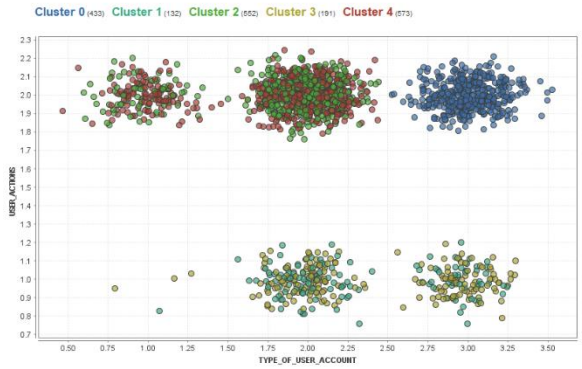
Nama Pengguna	K1	K2	K3	K4	K5
PLN_GM_01	0,384	0,327	1,713	1,696	0,255
PLN_GM_02	1,238	0,825	1,689	0,593	1,846
PLN_GM_03	1,589	1,175	1,792	1,471	0,738
SALES_04	1,727	0,268	0,697	1,968	0,587
SALES_05	0,792	0,586	1,237	1,624	1,925
CUST_ADMIN_01	1,642	1,712	1,982	1,168	0,681
CUST_ADMIN_02	1,897	0,651	1,489	1,936	1,297
NOC_ADM_01	0,973	1,511	0,569	1,953	1,798
NOC_ADM_02	0,641	0,578	0,488	1,911	0,111
SLS_RTL_01	1,546	1,678	1,294	0,593	1,872
SLS_RTL_02	1,175	0,788	1,683	0,937	1,479
MARK_01	0,985	0,674	1,718	1,819	0,279
BSN_01	0,252	1,846	1,373	0,814	0,549
BSN_02	1,994	0,428	1,335	0,183	1,238
MARK_04	0,189	1,331	0,818	1,136	0,364
MARK_05	1,592	1,524	1,458	1,211	1,972
TMK_01	0,387	0,225	0,329	0,444	0,734
TMK_02	1,996	0,125	1,539	1,755	1,944
ACC_ADM_01	0,754	0,465	0,422	1,272	1,194
ACC_ADM_02	1,767	1,358	1,683	1,979	1,579
GA_01	1,755	0,894	0,626	1,956	0,598
GA_02	0,712	0,783	0,594	1,676	0,846
BILL_01	0,171	1,658	1,526	0,584	0,931
BILL_02	0,238	1,441	1,266	1,411	0,295
HRD_02	1,321	1,757	0,475	1,499	1,958

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 0,384 pada kluster 1. 0,327 dimiliki kluster 2. Kluster 3 memiliki 1,713. Kluster 4 memiliki 1,696 dan 0,255 pada kluster 5. Berdasarkan pada kelima kluster tersebut, data pengguna 1 atau pertama memiliki jarak terpendek dengan kluster 5. Dengan demikian, data pengguna 1 masuk dalam kluster 5. Proses penentuan jarak terpendek ini juga dilakukan terhadap 24 data lain. Hasil pengklusteran dapat disajikan pada tabel dibawah ini.

TABEL III
PENEMPATAN DATA PADA KLUSTER JARAK TERDEKAT

Nama Pengguna	K1	K2	K3	K4	K5
PLN_GM_01					x
PLN_GM_02				x	
PLN_GM_03					x
SALES_04		x			
SALES_05		x			
CUST_ADMIN_01					x
CUST_ADMIN_02		x			
NOC_ADM_01			x		
NOC_ADM_02					x
SLS_RTL_01				x	
SLS_RTL_02		x			
MARK_01					x
BSN_01	x				
BSN_02				x	
MARK_04	x				
MARK_05				x	
TMK_01		x			
TMK_02		x			
ACC_ADM_01			x		
ACC_ADM_02		x			
GA_01					x
GA_02			x		
BILL_01	x				
BILL_02	x				
HRD_02			x		

Selanjutnya, disajikan ringkasan hasil analisis dengan k-means terkait dengan log riwayat *file* sistem. Kluster pertama, Permission_type menempati urutan pertama dibanding dengan 3 atribut lain mencapai 40%, user_account_type mencapai 29% dan yang terkecil adalah user_level.



Gambar 7. Hasil *klustering* dengan *K-means*

Kluster 2, User_type hampir dominan mencapai 54% dibanding dengan atribut lain hampir sama polanya dengan kluster 5. Sedangkan cluster 3, user_level yang paling unggul. Kluster 4, user_account_type dan user_level hampir memiliki kesamaan dengan 40% . berikut ini adalah gambar penyajian sebaran kluster dari masing – masing atribut atau variabel observasi.

Gambar 8. Sebaran kluster variabel observasi

D. Evaluasi dan 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 Ci dan Cj.

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

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

Berikut ini disajikan persamaan R dan Davies Bouldin Index.

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

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

Davies Bouldin index (DBI) mendeskripsikan sebagai skema evaluasi internal. Dengan DBI, dataset log riwayat aktivitas 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. Adapun hasil pengukuran dapat disajikan pada tabel dibawah ini.

TABEL IV
HASIL PENGUKURAN MODEL K-MEANS

Pengukuran	Hasil Observasi
Davis Bouldin	0,113
F Measure	88%

Berdasarkan tabel di atas, kinerja model dengan ukuran Davies Bouldin menunjukan keterangan analisis data dalam kluster. Hasil dari DBI, memberikan keterangan bahwa model atau metode klustering dengan K-Means dengan kategori baik. Indikasinya adalah nilai hasil observasi mencapai 0,113. Semakin kecil semakin baik dengan rentang nilai terendah adalah 0 dan tertinggi adalah 1. Berbeda dengan F measure masih sama dalam konsep pengukuran akurasi. Nilai Measure dianggap baik disebabkan hasil perolehan mencapai 88%.

F-Measure merupakan kombinasi antara recall dan precision. Recall berperan dapat mengkomparasi pada data antara positif benar dengan banyak data yang yang secara aktual posisitif. Precission merupakan komparasi antara positif benar dengan banyak data yang diprediksi positif. Berdasarkan tabel diatas, dapat disimpulkan model dengan metode K-Means dapat dianggap baik dan akurat pada kasus terkait.

IV. KESIMPULAN

Sistem basis data yang menghimpun data dalam bentuk log riwayat aktivitas pengguna telah dianalisis dan dibuat kluster. Data pada log riwayat aktivitas pengguna umumnya cenderung memiliki ketidaksamaan. Karakteristik lainnya adalah berkaitan dengan didalam kelompok yang sama. Dengan demikian, dataset pada log aktivitas 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 aktivitas sudah dilakukan dan diidentifikasi terdapat 5 kelompok sebagai kluster terhadap pengguna pada sistem.

Meninjau pada kualitas metode klustering khususnya pada teknik K-Means pada kasus ini bergantung pada tiga hal. Pertama, kepadatan dan jarak pada tiap kluster diobservasi untuk mencari kemiripan sebagai pengukuran yang ada pada dataset.

Kedua, penemuan data yang dianggap tidak normal atau dengan kata lain anomali bisa membantu dalam implementasi analisis klustering termasuk pada identifikasi pola. Dengan demikian, data yang tidak wajar ini bukan hanya anomali pada klustering tetapi bisa diprediksi bahaya anomali. Ketiga, kelebihan ketika menemukan beberapa atau semua pola pada kasus didalam dataset bisa dianggap sebagai klustering bahkan pada dataset kasus log aktivitas tersembunyi.

Ada tantangan dan kebutuhan dalam metode klustering terutama pada kasus ini. Pertama, aspek skalabilitas membuat pengertian bahwa pada kasus log aktivitas pengguna masih berfokus pada pengambilan sampel. Meskipun metode klustering lebih cocok dengan data yang kecil seperti sampel namun bagaimana menangani data bervolume besar dan kecepatan tinggi. Data yang besar menjadi lebih kompleks s dan peluang kluster bisa meningkat. Kedua, kinerja menangani berbagai jenis atribut yang berbeda. Pada kasus log aktivitas 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 aktivitas 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 aktivitas .

REFERENSI

[1] J. Umarani, S. Manikandan, D. Centre, and T. Nadu, "Implementation of *Data mining* Concepts in R Programming 1," *Int. J. Trendy Res. Eng. Technol.*, vol. 4, no. 1, pp. 1–7, 2020.

[2] X. Wang and F. Liu, "Data-driven relay selection for physical-layer security: A decision tree approach," *IEEE Access*, vol. 8, no. 1, pp. 12105–12116, 2020, doi: 10.1109/ACCESS.2020.2965963.

[3] S. Mohammad, "Overview on *Data mining* With Cloud," *J. Emerg. Technol. Innov. Res.*, vol. 4, no. 12, pp. 519–523, 2021.

[4] I. Kholod, A. Shorov, and S. Gorlatch, "Efficient distribution and processing of data for parallelizing *data mining* in mobile clouds," *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 11, no. 1, pp. 2–17, 2020, doi: 10.22667/JOWUA.2020.03.31.002.

[5] E. Bertino, M. Kantarcioglu, C. G. Akcora, S. Samtani, S. Mittal, and M. Gupta, "AI for Security and Security for AI," *CODASPY 2021 - Proc. 11th ACM Conf. Data Appl. Secur. Priv.*, pp. 333–334, 2021, doi: 10.1145/3422337.3450357.

[6] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "Efficient privacy preservation of big data for accurate *data mining* ," *Inf. Sci. (Ny).*, vol. 527, no. xxxx, pp. 420–443, 2020, doi: 10.1016/j.ins.2019.05.053.

[7] N. Bhandari and P. Pahwa, *Comparative analysis of privacy-preserving data mining techniques*, vol. 56. Springer Singapore, 2019.

[8] P. Rupprecht *et al.*, "A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging," *Nat. Neurosci.*, vol. 24, no. 9, pp. 1324–1337, 2021, doi: 10.1038/s41593-021-00895-5.

[9] K. N. Durai, R. Subha, and A. Haldorai, "A Novel Method to Detect and Prevent SQLIA Using Ontology to Cloud Web Security," *Wirel. Pers. Commun.*, vol. 117, no. 4, pp. 2995–3014, 2021, doi: 10.1007/s11277-020-07243-z.

[10] M. Jha, *Smart Intelligent Computing and Applications*, vol. 104. Springer Singapore, 2019.

[11] 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.

[12] D. Bhayani, "Identification of Security Breaches in Log Records using *Data mining* Techniques," *Int. J. Pure Appl.*

Math., vol. 119, no. 15, pp. 743–756, 2018.

[13] 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.

[14] 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.

[15] R. T. H. Hasan and S. Y. Ameen, "Security Enhancement of IoT and Fog Computing Via Blockchain Applications," *J. Soft Comput. Data Min.*, vol. 2, no. 2, pp. 26–38, 2021, doi: 10.30880/jscdm.2021.02.02.003.

[16] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," *Proc. 2019 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2019*, pp. 49–54, 2019, doi: 10.1109/ICTS.2019.8850982.

[17] 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.

[18] J. Clark and F. Provost, "Unsupervised dimensionality reduction versus supervised regularization for classification from sparse data," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 871–916, 2019, doi: 10.1007/s10618-019-00616-4.

[19] M. R. Anwar, R. Panjaitan, and R. Supriati, "Implementation Of Database Auditing By Synchronization DBMS," *Int. J. Cyber IT Serv. Manag.*, vol. 1, no. 2 SE-Articles, pp. 197–205, 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/53.

[20] W. Fu and P. O. Perry, "Estimating the Number of Clusters Using Cross-Validation," *J. Comput. Graph. Stat.*, vol. 29, no. 1, pp. 162–173, 2020, doi: 10.1080/10618600.2019.1647846.

[21] T. Javid, M. K. Gupta, and A. Gupta, "A hybrid-security model for privacy-enhanced distributed *data mining* ," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2020.06.010.

[22] 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.

[23] 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.

[24] S. Ramos *et al.*, "*Data mining* techniques for electricity customer characterization," *Procedia Comput. Sci.*, vol. 186, no. 3, pp. 475–488, 2021, doi: 10.1016/j.procs.2021.04.168.

[25] R. C. Sharma, K. Hara, and H. Hirayama, "A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data," *Scientifica (Cairo).*, vol. 2017, 2017, doi: 10.1155/2017/9806479.

[26] A. Khobzaoui, M. Benhamouda, and M. Fahsi, "*Data mining* Contribution to Intrusion Detection Systems Improvement," in *ACM International Conference Proceeding Series*, 2020, pp. 1–8, doi: 10.1145/3447568.3448514.

[27] X. N. Bui, H. Nguyen, Y. Choi, T. Nguyen-Thoi, J. Zhou, and J. Dou, "Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm," *Sci. Rep.*, vol. 10, no. 1, pp. 1–17, 2020, doi: 10.1038/s41598-020-66904-y.

[28] L. Vanfretti and V. S. N. Arava, "Decision tree-based classification of multiple operating conditions for power system voltage stability assessment," *Int. J. Electr. Power Energy Syst.*, vol. 123, no. 3, pp. 1–10, 2020, doi: 10.1016/j.ijepes.2020.106251.

[29] M. Rouzbahman *et al.*, "*Data mining* Methods for Optimizing Feature Extraction and Model Selection," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 2020, pp. 1–8, doi: 10.1145/3406601.3406602.