PENGGUNAAN MODEL KLUSTERISASI DENGAN METODE K-MEANS UNTUK MENDETEKSI AKTIVITAS PENGGUNA WEB MENGGUNAKAN RAPIDMINER BERDASARKAN USER-AGENT-BASED STUDI KASUS PADA APLIKASI-APLIKASI PADA STMIK JAKARTA STI&K

Febianto Arifien
STMIK Jakarta STI&K
Jl. BRI No.17, Radio Dalam, Kebayoran Baru, Jakarta Selatan 12140
febianto@jak-stik.ac.id

ABSTRAK

Permasalahan interoperabilitas mendeteksi ketidakcocokan antar hubungan yang tidak standar.Hypertext transfer protocol (HTTP) telah menjadi standar de facto untuk protokol lapisan aplikasi transport seperti JSON, SOAP, gaming, voice over IP, video streaming dan perangkat lunak lainnya. Penelitian ini difokuskan pada HTTP user-agent (UA), yang dikirim oleh browser web ke server web untuk menyampaikan sistem operasi, jenis dan versi browser client, mesin render dan nama aplikasi dalam hal lalu lintas dari perangkat seluler. Tantangan utamanya adalah desain fitur ekstensi yang dapat membantu membedakan ekstensi berbahaya. Pembelajaran mesin teknik seperti K-means dan FCM clustering algoritma menggunakan alat RapidMiner digunakan untuk melakukan analisis, data log akses web berdasarkan UA (User-Agent) yang telah dikumpulkan melalui internet.Dengan menganalisis kode sumber dan aktivitas ekstensi selama runtime,Server web menggunakan informasi tersebut untuk menyesuaikan respons mereka terhadap browser web untuk rendering yang tepat. Sehingga dapat dibedakan kelompok atau kluster yang memiliki anggota yang sedikit untuk dapat diproses lebih lanjut apakah kelompok tersebut memiliki standar format yang sesuai atau tidak.

Kata Kunci :*Interoperabilitas, User-Agent, HTTP, K-Means, RapidMiner*

PENDAHULUAN

Browser web seperti Internet Explorer, Opera, Firefox, Safari, dan Chrome, digunakan untuk berinteraksi dengan semua informasi di Internet. Untuk meningkatkan fungsi dan personalisasi pengalaman penjelajahan user, ekstensi diizinkan untuk menggunakan komponen browser (misalnya riwayat, bookmark, sumber daya laman web, dan sosial favorit fitur). Pengguna dapat menginstal berbagai ekstensi dari toko resmi vendor browser seperti Add-Ons Opera, Galeri Ekstensi Safari, dan Toko Web Chrome atau dari situs web pihak ketiga.[1]

Hypertext transfer protocol (HTTP) telah menjadi standar de facto untuk protokol lapisan aplikasi transport seperti JSON, SOAP, gaming, voice over IP, video streaming dan perangkat lunak lainnya. Dari perspektif pengukuran jaringan dan analisis lalu lintas, sangat penting untuk dapat mengklasifikasikan dan memisahkan berbagai aplikasi yang diangkut melalui HTTP dalam membantu tugas-tugas keamanan jaringan seperti deteksi malware, terutama karena HTTP telah menjadi media utama untuk aktivitas terlarang di Internet seperti unduhan drive-by, phishing ,perintah-dan-kontrol botnet (C&C), klik penipuan dan sebagainya.[2]

Masalah interoperabilitas muncul ketika dua atau lebih sistem yang tidak kompatibel dimasukkan dalam hubungan. Secara umum, ketidakcocokan ini terkait dengan lapisan interoperabilitas. Hal ini akan memunculkan hambatan dengan kurangnya seperangkat standar yang kompatibel untuk memungkinkan penggunaan teknik komputasi heterogen untuk berbagi dan bertukar data antara dua sistem atau lebih.[3]

Salah satu fitur utama yang menjadi fokus penelitian adalah pada bagian header HTTP user-agent (UA), yang dikirim oleh browser web ke server web untuk menyampaikan sistem operasi, jenis dan versi browser client, mesin render dan nama

aplikasi dalam hal lalu lintas dari perangkat seluler . Server web menggunakan informasi tersebut untuk menyesuaikan respons mereka terhadap browser web untuk rendering yang tepat. Namun, string UA juga digunakan oleh malware untuk kegiatan terlarang, misalnya, sebagai cara untuk menipu browser yang digunakan oleh client dengan cara on-click-fraud-events atau pengaksesan terlarang ketika mengakses event, sebagai cara untuk membocorkan informasi pribadi dari host yang terinfeksi atau untuk berkomunikasi dengan server C&C. Baru-baru ini, string UA telah digunakan sebagai cara untuk mengeksploitasi server yang rentan terhadap serangan Shellshock.[2]

Penelitian yang berkaitan dengan deteksi aktivitas menggunakan profil user-agent seperti pengembangan mesin CFG (Context-Free-Grammar) untuk mengidentifikasi string UA (User-Agent) yang terdiri dari dua fase: (1) Mengekstrak string UA dari tiga sumber: lingkungan sand-box dan jejak jaringan. (2) Menulis CFG secara bertahap [1] Yao Wang , dan Wei Shao memperkenalkan pendekatan berbasis machine-learning untuk mendeteksi ekstensi Chrome yang berbahaya. Tantangan utamanya adalah desain fitur ekstensi yang dapat membantu membedakan ekstensi berbahaya. Dengan menganalisis kode sumber dan aktivitas ekstensi selama runtime, kami membuat serangkaian fitur yang secara manual dapat mewakili ekstensi. Penelitian yang dilakukan berkaitan dengan menganalisis data penggunaan web seperti pada penelitian M.Santhanakumar dan C.Christopher Colombus menerapkan pembelajaran mesin teknik seperti K-means dan FCM clustering algoritma menggunakan alat RapidMiner. Untuk melakukan analisis, data log akses web berdasarkan UA (User-Agent) yang telah dikumpulkan melalui internet. File tersebut berisi urutan detail akses pengguna. Di kedua algoritma pengelompokan, awalnya pusat cluster dipilih secara acak didasarkan pada cluster centroid dan clustering data dapat dievaluasi.[4]. Penelitian ini bertujuan untuk membuat pengelompokan berdasarkan informasi UA (User-Agent) untuk mendeteksi aktivitas yang tidak normal agar mencegah masuknya malware atau virus. Penggunaan metode clustering dimaksudkan untuk memisahkan informasi log yang sesuai dengan yang tidak sesuai berdasarkan format UA (User-Agent) yang telah dikumpulkan melalui internet.

TINJAUAN PUSTAKA

A. USER-AGENT DARI JENIS BROWSER POPULER

Browser web mungkin adalah aplikasi paling populer yang digunakan pada mesin desktop dan laptop. Tidak mengherankan bahwa mereka terdiri dari sejumlah besar string UA yang dilihat dalam dataset Tabel 1 menunjukkan beberapa contoh string UA untuk browser umum yang ditemukan dalam dataset Dapat terlihat bahwa string UA dihasilkan oleh browser ini mengandung kata kunci yang mirip dan berbagi komponen struktural tertentu. Sebagai contoh, string UA yang dihasilkan oleh browser IE dimulai dengan kata kunci ‘Mozilla / 4.0’ atau ‘Mozilla / 5.0’, diikuti hanya dengan serangkaian kata kunci yang dilampirkan oleh tanda kurung, termasuk istilah dan versi 'MSIE'.[2]

Tabel 1.*User-Agent yang Dihasilkan oleh Browser Web*

MSIE	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 6.1; en-US; .NET CLR 1.1.22315)
	Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.2; Trident/6.0)
Firefox	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:29.0) Gecko/20120101 Firefox/29.0
	Mozilla/5.0 (X11; U; Linux i686; de-DE; rv:1.7.6) Gecko/20050306 Firefox/1.0.1
Chrome	Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/37.0.2049.0 Safari/537.36
	Mozilla/5.0 (Linux; Android 4.0.3; GT-I9100 Build/IML74K) AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.133 Mobile Safari/535.19
Safari	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_7; en-us) AppleWebKit/534.16+ (KHTML, like Gecko) Version/5.0.3 Safari/533.19.4
	Mozilla/5.0 (Windows; U; Windows NT 5.1; it) AppleWebKit/522.13.1 (KHTML, like Gecko) Version/3.0.2 Safari/522.13.1
Opera	Opera/9.80 (X11; Linux x86_64; U; en) Presto/2.9.168 Version/11.50
	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.0) Opera 12.14

B. K-MEANS CLUSTERING

Clustering berkaitan dengan pengelompokan objek bersama yang mirip satu sama lain dan berbeda ke benda-benda milik kelompok lain. Ada banyak algoritma pengelompokan seperti k-means, kmedian, DBSCAN, Hierarchical clustering dan Xmeans tersedia untuk melakukan proses pengelompokan. Untuk menganalisis User_Agent dengan algoritma apa pun yang diharuskannya diproses terlebih dahulu. Setelah menemukan kesamaan dari data maka Clustering data dilakukan berdasarkan pada pencarian serupa yang

dilakukan oleh Pengguna. Di sini menggunakan algoritma k-means untuk mengelompokkan data berdasarkan pencarian kesamaan dilakukan oleh pengguna. k-means clustering adalah algoritma pengelompokan eksklusif yaitu setiap objek ditugaskan tepat satu dari satu set gugus. Objek dalam satu cluster mirip dengan masing-masing lain. Kesamaan antara objek didasarkan pada ukuran jarak di antara mereka. Clustering adalah berkaitan dengan pengelompokan bersama benda-benda yang ada mirip satu sama lain dan berbeda dengan objek milik kelompok lain. Clustering adalah teknik untuk mengekstraksi informasi dari data yang tidak berlabel. Clustering sangat berguna dalam banyak skenario yang berbeda misalnya dalam aplikasi pemasaran, pengguna mungkin tertarik untuk menemukan kelompok pelanggan dengan perilaku pembelian serupa. Dalam k-means clustering, centroid (pusat cluster) harus mencari tahu dulu. Centroid dihitung menggunakan jarak Euclidean dan dihitung oleh persamaan 1. Kadang-kadang centroid adalah salah satu poin dalam cluster. Kemudian, tetapkan nilai k untuk cluster diperlukan untuk diproses. Umumnya kecil nilai integer dapat ditetapkan untuk nilai k. Kemudian pilih k objek secara acak dan gunakan ini sebagai set awal k centroid. Alokasikan setiap objek ke cluster, yang terdekat dengan pusat massa dan hitung ulang centroid dari k cluster. Ulang menghitung centroid hingga centroid mungkin menjadi optimal.[4] Jarak Euclidean dapat dihitung oleh persamaan berikut :

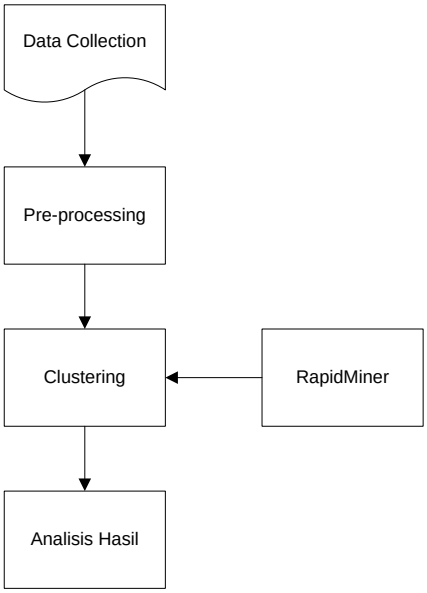
$(n)(n-1)/2$ (1)

C. RAPIDMINER

Merupakan alat komputasi statistik dikembangkan dan berhasil diterapkan pada berbagai data untuk dianalisis dan memantau prosesnya . RapidMiner proyek dimulai pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Kelompok Intelijen dari Katharina Morik di Universitas Teknologi Dortmund. RapidMiner adalah salah satu alat penambangan data yang digunakan untuk menganalisis informasi yang diakses web. Ini digunakan untuk penelitian, pendidikan, pembuatan prototipe cepat, aplikasi pengembangan, dan aplikasi industri . Aplikasi ini berlisensi Open Source, yang termasuk pembersihan data, transformasi data, optimasi, validasi dan visualisasi. Itu visualisasi berisi melihat data yang dianalisis dalam bentuk sebar plot, Bar, Pie chart, dll ... Itu juga termasuk berbagai pengelompokan dan klasifikasi algoritma untuk melakukan proses analitik. Salah satunya Fitur utama alat ini yaitu, ia akan menganalisis data tanpa coding program, namun jika ada orang ingin menganalisis data dengan pengkodean mereka sendiri itu juga bisa dimasukkan dalam apliaksi. Berbagai jenis dataset dapat diimpor oleh RapidMiner seperti, excel, csv, xml, arff, akses dll . Sejak 2007, RapidMiner telah sangat diperluas dan menjadi salah satu yang paling penambangan data penting dan alat analisis data . Untuk proses analisis ini file UA (User-Agent) adalah dikumpulkan dari Internet dalam periode waktu 10-09-2019 hingga 21-01-2020.[4]

METODE PENELITIAN

Gambaran penelitian yang difokuskan pada analisa pengelompokan menggunakan metode K-Means yang dilakukan menggunakan aplikasi, dapat dijelaskan pada gambar 1.



Gambar 1.*Gambaran Penelitian*

- **Data Collection**

Data Collection atau pengumpulan data didefinisikan sebagai pengumpulan informasi User Agent log akses web yang dilakukan dari sisi Klien melalui beberapa aplikasi yang digunakan di STMIK Jakarta STI&K yaitu Aplikasi Mahasiswa dan Aplikasi Dosen. Contoh informasi yang didapatkan dari UA, contohnya adalah sebagai berikut:

Mozilla/5.0 (Windows NT 10.0; WOW64; rv:68.0) Gecko/20100101 Firefox/68.0

1. Mozilla/5.0 Identifikasi dasar engine blog(Firefox masih merupakan keturunan Mozilla)
2. Windows NT 10.0; WOW64 Identifikasi mesin dan arsitektur komputer(saya pakai Windows NT 10.0 dengan mesin 64 bit)
3. rv:68.0 Revisi dari browser engine
4. Gecko/20100101 Layout Browser Engine yang digunakan
5. Firefox/68.0 Versi dari Firefox-nya sendiri.

- **Pre-Processing**

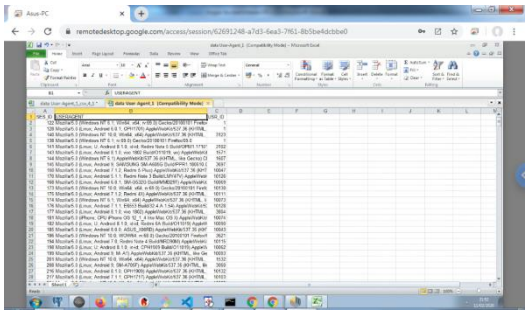
Setelah pengumpulan dataset penggunaan web, lakukan Pre-prosessing pada dataset, data yang dikumpulkan dari web biasanya beragam, heterogen dan tidak terstruktur diperlukan untuk melakukan pre-prosessing seperti pemfilteran data yang tidak perlu dan tidak relevan, prediksi dan mengisi nilai-nilai yang hilang, menghilangkan noise, menyelesaikan inkonsistensi sebelum menerapkan algoritma. Pre-prosessing adalah tugas yang sulit karena keragaman data yang tersedia. Proses pre-prosessing dilakukan manual dengan mengambil informasi-informasi yang tesedia dari data User-Agent lalu dipisahkan berdasarkan kategori yang sama. Hasil data collection seperti yang terlihat pada tabel 2.

Tabel 2.*Hasil Data Collection*

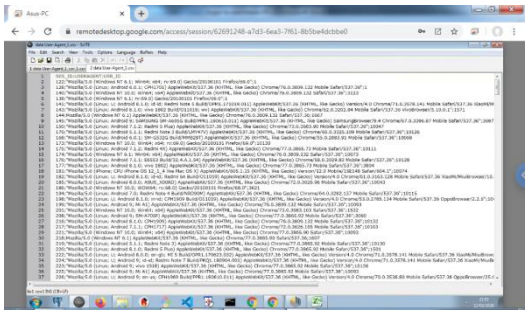
SES_ID	UA	USR_ID
122	Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:69.0) Gecko/20100101 Firefox/69.0	1
...	...	...

- Label SES_ID merupakan id unik dari session yang dibentuk ketika user login ke aplikasi
- Label USERAGENT diperoleh melalui fungsi javascripst navigator.userAgent, yaitu nilai header agen-pengguna yang dikirim oleh browser ke server. Nilai yang dikembalikan, berisi informasi tentang nama, versi, dan platform browser.
- Label USR_ID merupakan id yang diambil dari tabel USER, berdasarkan relasi antara tabel User dengan tabel Session.

Proses pre-prosesing masih dilakukan secata manual, yaitu memisahkan label user-agent yang awalnya satu kolom menjadi empat kolom dengan kriteria yang telah dijelaskan pada bagian data collection. Contohnya:

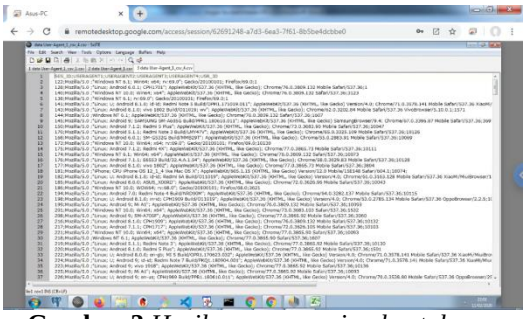


Gambar 1.*Pre-prosessing bentuk excel*

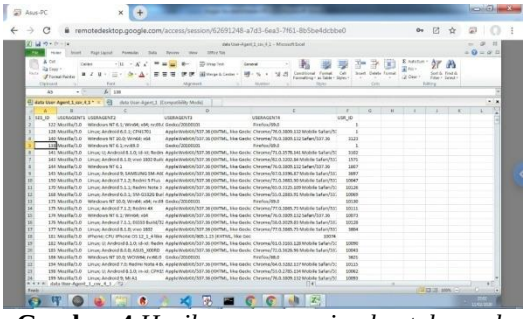


Gambar 2.*Pre-prosessing bentuk csv*

Data awal dalam bentuk excel dikonversi terlebih dahulu kedalam bentuk csv, terlihat pada gambar 1 dan gambar 2. Kemudian dilakukan pemisahan kolom user-agent menjadi 4 kolom dengan mengidentifikasi susunan informasi seperti pada data collection dengan menambahkan string ";" sebagai pemisah kolom. Hasilnya terlihat pada gambar 3 dan 4.



Gambar 3.*Hasil pre-prosessing bentuk csv*



Gambar 4.*Hasil pre-prosessing bentuk excel*

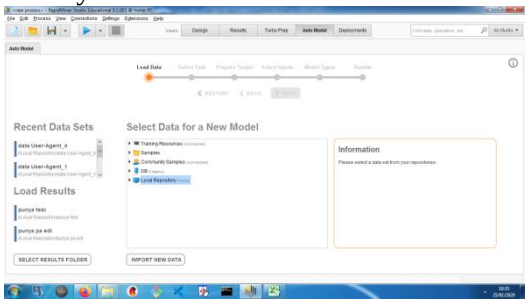
Gambar 4 memperlihatkan hasil pemisahan label User-agent dari satu kolom menjadi 4 kolom berdasarkan informasiIdentifikasi dasar engine blog, Identifikasi mesin dan arsitektur komputer, revisi dari browser, layout Browser Engine yang digunakan dan versi dari Firefox-nya sendiri.

- **Penggunaan RapidMiner dengan Metode Clustering**

Model pengolahan data yang dipilih berdasarkan aplikasi RapidMiner sudah tersedia menggunakan Auto Model, dikarenakan kemudahan dan proses automatisasi yang memudahkan pengguna tanpa banyak melakukan penyesuaian parameter dan lain sebagainya. Langkah-langkah yang disediakan adalah sebagai berikut:

A. Load Data

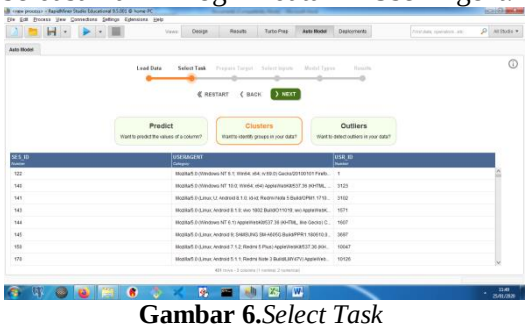
Modul Load Data adalah modul untuk mengimport data ke dalam aplikasi. Data yang digunakan untuk penelitian ini sebanyak 451 data dari tabel Session.



Gambar 5.*Load Data*

B. Select Task

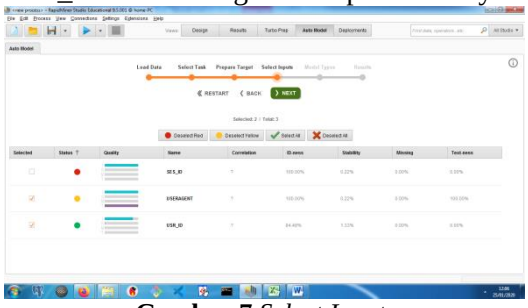
Modul Select Task adalah modul pemilihan metode data mining yang akan dipilih. Untuk penelitian ini dipilih Clusters, karena yang akan diteliti adalah pengelompokan berdasarkan log data User-Agent.



Gambar 6.*Select Task*

C. Select Inputs

Modul Select Input digunakan untuk memilih variabel x dan y yang digunakan untuk melakukan pengelompokkan. Dalam penelitian ini digunakan variabel USERAGENT dan USR_ID sebagai parameternya.

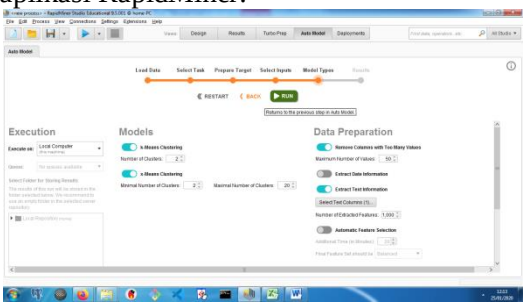


Gambar 7.*Select Input*

D. Model Types

Modul Model Types adalah pemilihan model yang akan digunakan

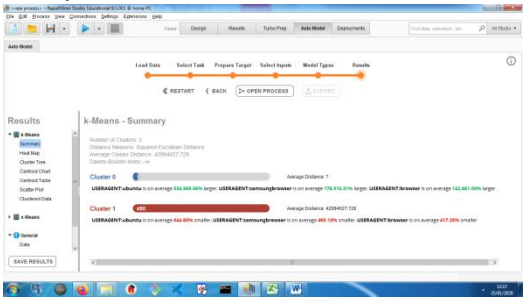
dalam pemrosesannya. Penelitian ini difokuskan penggunaan K-Means sebagai alat untuk membuat pengelompokan, maka dipilih Model k-means Clustering pada aplikasi RapidMiner.



Gambar 8.*Model Types*

E. Result

Pada Modul Result ditampilkan hasil pemrosesan berdasarkan langkah-langkah sebelumnya. Hasil pemrosesan berupa Ringkasan (Summary), Heat Map, Cluster Tree, Centeroid Chart, Centeroid Table, Scatter Plot dan Clustered Data.



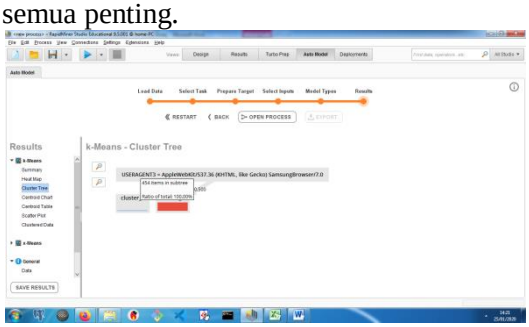
Gambar 9.*Result*

HASIL DAN PEMBAHASAN

Setelah diperoleh hasil pada implementasi model menggunakan aplikasi RapidMiner, dapat dilihat bahwa banyaknya cluster ada 2 dengan jumlah anggota yang sangat signifikan, yaitu cluster 0 yang beranggotakan 1 record dan cluster 1 yang beranggotakan 451 record dari 452 data.

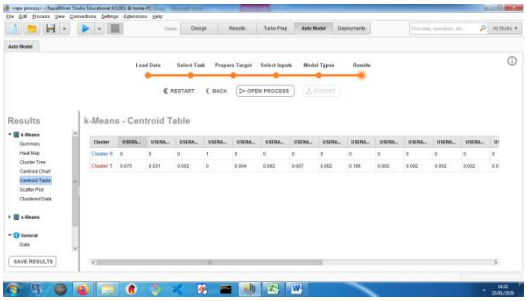
Gambar 10.*Heat Map*

Pola heat map digunakan untuk mengenali atribut yang paling penting dalam kelompok. Terlihat pada gambar 10 bahwa distribusi polanya merata artinya atributnya semua penting.



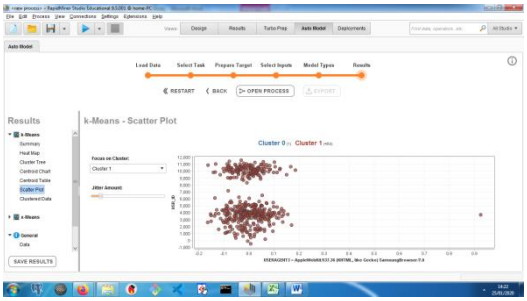
Gambar 11.*Cluster Tree*

Diagram cluster tree memperlihatkan perbedaan utama antara cluster-cluster yang telah dibentuk.



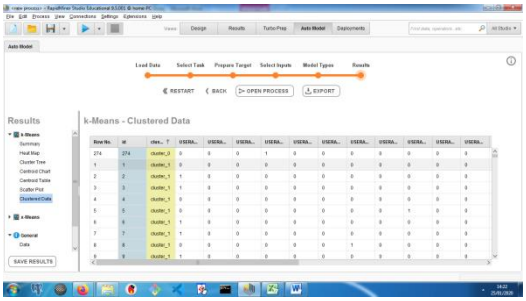
Gambar 12.*Centeroid Table*

Centeroid tabel, memperlihatkan nilai kluster-kluster centeroid pada tabel. Diperlihatkan pada gambar 9.



Gambar 13.*Scater Plot*

Gambar 10 memperlihatkan pola sctater-plot dimana titik-titiknya cenderung berkumpul pada satu sisi. Jadi polanya cenderung seragam.



Gambar 14.*Clustered Data*

Data kluster yang dihasilkan seperti pada gambar 11, bahwa hanya ada dua kelompok dimana kelompok pertama hanya beranggotakan satu data saja. Ini dapat diartikan bahwa dari 454 data yang dijadikan model, hanya satu data saja yang memiliki similaritas yang berbeda dari semua data yang dimodelkan. Jadi dapat diartikan bahwa hampir semua user yang memakai aplikasi di STMIK Jakarta STI&K menggunakan browser yang seragam dan tidak ada yang aneh dalam aktivitas pemakaiannya .

PENUTUP

Penggunaan pengelompokan K-Means dalam pendektesian aktivitas pengguna web, diharapkan bisa mendeteksi aktivitas yang tidak normal untuk mencegah masuknya malware atau virus. Atau bisa melihat kecendrungan pengguna web, browser apa yang sering dipakai, operating sistemnya, menurut informasi log User-Agent yang telah diketahui. Diharapkan untuk penelitian selanjutnya bisa ditambahkan analisa semantik untuk memisahkan karakterisitik dari log User-Agent sehingga variabel pembandingnya bisa lebih banyak. Dan untuk melihat kecendrungan pengamanan terhadap serangan malware dan virus, bisa menambahkan definisi malware yang bisa tedeteksi dari log file User_Agent, sehingga bisa dimanfaatkan untuk mentraining data log agar dapat otomatis mendeteksi serangan virus dan malware.

DAFTAR PUSTAKA

[1] Y. Wang, W. Cai, P. Lyu, and W. Shao, “A Combined Static and Dynamic Analysis Approach to Detect Malicious Browser Extensions,” *Secur. Commun. Networks*, 2018.

[2] Y. Zhang, H. Mekky, Z. L. Zhang, R. Torres, S. J. Lee, A. Tongaonkar, and M. Mellia, “Detecting malicious activities with user-agent-based profiles,” *Int. J. Netw. Manag.*, 2015.

[3] F. Arifien and M. Riastuti, “Model Interoperabilitas Web Service Feeder PDDIKTI Menggunakan Enterprise Javabeans (EJB) dan REST-API,” vol. 3, 2019.

[4] M. Santhanakumar and C. C. Columbus, “Web Usage Based Analysis of Web Pages Using RapidMiner Department of Computer Science and Engineering,” *WSEAS Trans. Comput.*, vol. 14, pp. 455–464, 2015.