

## BAB II

### LANDASAN TEORI

Bab ini membahas tentang teori-teori berdasarkan penelitian sebelumnya yang berkaitan dengan penelitian yang akan dilakukan terkait analisa perilaku pengguna internet.

#### 2.1. Perilaku Pengguna (*user behavior*) dan Analisis Perilaku

Perilaku penggunaan teknologi informasi sangat bergantung pada evaluasi pengguna dari sebuah sistem, dengan kata lain, penggunaan sistem adalah indikator dari kesuksesan dan penerimaan teknologi informasi (Ma & Zhang, 2015). Variabel perilaku penggunaan (*use behavior*) menggunakan data primer yang berasal dari tabel firewall dari sistem router OS Mikrotik. Sistem Operasi Mikrotik yang digunakan dalam penelitian ini akan mengukur dengan menggunakan instrumen data yang terdiri dari (1) Alamat IP Address user/workstation (2) Bandwidth yang digunakan (3) Intensitas Upload dan Download (4) Intensitas tiap jenis perilaku pengguna internet.

Analisis perilaku berfokus pada pemahaman bagaimana konsumen atau pengguna bertindak dan mengapa, memungkinkan prediksi yang akurat tentang bagaimana pengguna akan bertindak di masa depan. Hal ini memungkinkan administrator untuk membuat keputusan yang tepat untuk segmen koneksi data yang tepat pada waktu yang tepat.

Analisis perilaku memanfaatkan data mining pengguna yang ditangkap selama sesi di mana konsumen menggunakan aplikasi, permainan, atau situs web, termasuk lalu lintas data seperti jalur navigasi, even klik, interaksi media sosial, dan

jumlah waktu yang dihabiskan di situs web. Poin data ini kemudian dikumpulkandan dianalisis, sehingga dengan melihat perkembangan sesi dari ketika pengguna pertama kali memasuki platform sampai koneksi dibuat, analisis perilaku memungkinkan tindakan dan tren masa depan akan diprediksi berdasarkan pengumpulan data tersebut.

## **2.2. Trafik**

Trafik dapat diartikan sebagai perpindahan informasi dari satu tempat ke tempat lain melalui jaringan telekomunikasi. Besaran dari suatu trafik telekomunikasi diukur dengan satuan waktu, sedangkan nilai trafik dari suatu kanal adalah lamanya waktu kedudukan pada kanal tersebut. Salah satu tujuan perhitungan trafik adalah untuk mengetahui kerja suatu jaringan (*Network Perfomance*) dan mutu pelayanan jaringan (*Quality of Service*) (Riadi, 2011).

## **2.3. Data Mining**

### **2.3.1. Pengertian Data Mining**

Data Mining merupakan suatu knowledge yang menggunakan beberapa teknik, seperti teknik statistika, matematika, kecerdasan buatan, dan machine learning untuk menemukan suatu pola-pola yang dapat mengidentifikasi suatu informasi yang bermanfaat. Data Mining merupakan serangkaian teknik yang dapat memproses suatu data untuk menggali nilai tambah suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Mustafa et al., 2018).

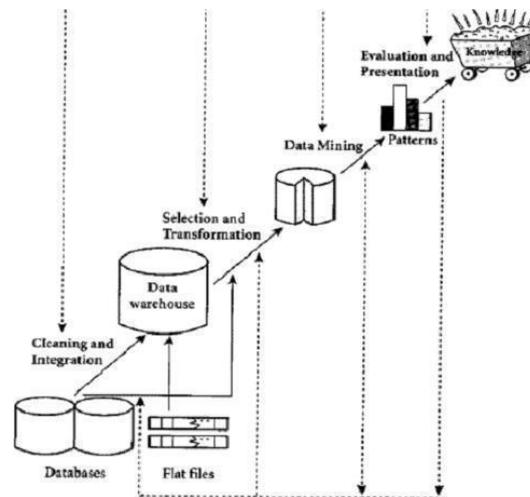
1) Beberapa teknik dan sifat data mining adalah sebagai berikut:

a. *Classification*

- b. *Clustering*
- c. *Association Rule*
- d. *Regression*
- e. *Deviation Detection*

## 2) Tahap-tahap Data Mining

Data mining adalah sebuah untaian proses, maka terbagi menjadi beberapa tahap. Tahapan tersebut akan bersifat interaktif, pengguna akan terlibat langsung atau dengan perantara KDD (Untari, 2010). Berikut tahapan data mining ditunjukkan pada gambar 2.1.



Gambar 2.1 Tahapan Data Mining (Untari, 2010)

- 3) Tahapan data mining dibagi menjadi bagian-bagian yaitu :
  - a. Pembersihan Data (*Data Cleaning*)

Tahapan pertama yang dilakukan adalah *data cleaning* Han, Kamber & Pei menjelaskan *data cleaning* merupakan proses untuk dapat

mengatasi nilai yang hilang, *noise* dan data yang tidak konsisten (Han et al., 2011).

b. Integrasi Data (*Data Integration*)

*Data integration* merupakan proses menggabungkan data dari banyak sumber data. Proses ini dapat membantu mengurangi redundansi data dan data yang tidak konsisten yang disebabkan pengambilan data dari banyak sumber data. Hal ini tentu saja akan berpengaruh terhadap kecepatan dan akurasi saat melakukan data mining (Han et al., 2011).

c. Transformasi Data (*Data Transformation*)

*Data transformation* dilakukan untuk mengubah bentuk dan format data hal ini tentunya akan sangat membantu memudahkan pengguna dalam proses mining ataupun memahami hasil yang didapat (Han et al., 2011).

d. Proses *Mining*

Proses inti yang menggunakan sebuah metode tertentu untuk memperoleh sebuah pola dari data.

e. Evaluasi Pola (*Pattern Evaluation*)

Proses mengidentifikasi pola.

f. Presentasi Pengetahuan (*Knowledge Presentation*)

Proses yang dapat mempresentasikan informasi yang dibutuhkan dimana informasi yang telah didapatkan kemudian digunakan oleh pemilik data.

*Clustering* atau klasterisasi merupakan suatu metode atau teknik pengelompokan data. *Clustering* berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam *clustering*. *Clustering* tidak mencoba untuk melakukan klasifikasi ataupun memprediksi nilai dari variabel target. Akan tetapi, proses ini mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan. Teknik *clustering* banyak diterapkan dalam berbagai bidang (Tama, 2009).

#### A. Konsep Dasar Clustering

Hasil *clustering* yang baik akan menghasilkan tingkat kesamaan yang tinggi dalam satu kelas dan tingkat kesamaan yang rendah antar kelas. Kesamaan yang dimaksud merupakan pengukuran secara numerik terhadap dua buah objek. Nilai kesamaan antar kedua objek akan semakin tinggi jika kedua objek yang dibandingkan memiliki kemiripan yang tinggi. Begitu juga dengan sebaliknya. Kualitas hasil *clustering* sangat bergantung pada metode yang dipakai (Mustafa et al., 2018). Dalam *clustering* dikenal empat tipe data. Keempat tipe data pada tersebut ialah :

1. Variable berskala interval
2. Variable biner
3. Variabel nominal, ordinal dan rasio
4. Variable dengan tipe lainnya

Metode *clustering* juga harus dapat mengukur kemampuannya sendiri dalam usaha untuk menemukan suatu pola tersembunyi pada data yang sedang diteliti. Terdapat berbagai metode yang dapat digunakan untuk

mengukur nilai kesamaan antar objek-objek yang dibandingkan. Salah satunya ialah dengan *weighted Euclidean Distance*, *Euclidean distance* menghitung jarak dua buah point dengan mengetahui nilai dari masing-masing atribut pada kedua poin tersebut. Berikut formula yang digunakan untuk menghitung jarak dengan *Euclidean distance* (Omran et al., 2007):

$$Distance(p, q) = (\sum_k^N \mu_k |p_k - q_k|^r)^{1/r} \dots (1)$$

(Omran et al., 2007)

Keterangan :

N= Jumlah record data

K= Urutan field data

r= 2

$\mu_k$  = Bobot *field* yang diberikan *user*

Jarak adalah pendekatan yang umum dipakai untuk menentukan kesamaan atau ketidaksamaan dua vektor fitur yang dinyatakan dengan ranking. Apabila nilai ranking yang dihasilkan semakin kecil nilainya maka semakin dekat/tinggi kesamaan antara kedua vektor tersebut. Teknik pengukuran jarak dengan metode *Euclidean* menjadi salah satu metode yang paling umum digunakan (informatikalogi, 2018). Pengukuran jarak dengan metode *Euclidean* dapat dituliskan dengan persamaan berikut :

$$J(v_1, v_2) = \sqrt{\sum_{k=1}^N (v_1(k) - v_2(k))^2} \dots (2)$$

Dimana  $v_1$  dan  $v_2$  adalah dua vektor yang jaraknya akan dihitung dan N menyatakan panjang vector

## B. Metode *Clustering*

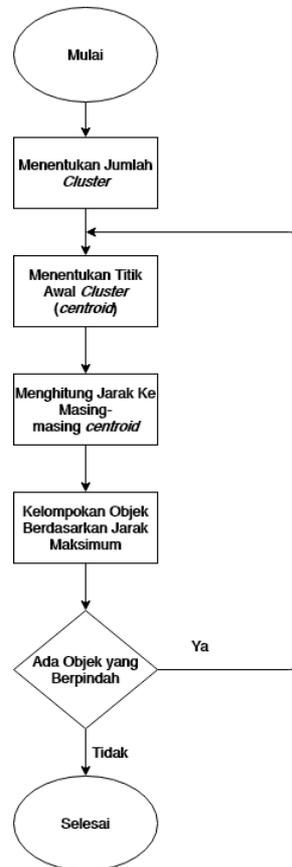
Metode atau teknik pengklasteran dapat dikelompokkan ke dalam empat kategori, yaitu (Bahauddin et al., 2021):

1. Metode Partitioning Melakukan pengklasteran melalui pencarian kluster yang terpisah secara langsung dan relokasi titik data secara iterative untuk memperbaiki kualitas kluster dari solusi awal
2. Metode Hirarkis: membangun kluster secara bertahap. yaitu melalui penggabungan kluster yang kecil menjadi kluster yang lebih besar, atau melalui pemecahan kluster yang lebih besar (menjadi kluster yang lebih kecil)
3. Metode density-based metode ini melakukan identifikasi kluster dengan menghubungkan daerah-daerah yang memiliki titik data yang cukup tinggi
4. Metode grid-based: untuk meningkatkan efisiensi pengklasterannya, metode ini membagi ruang data menjadi sejumlah sel yang membentuk struktur grid. kemudian menggunakan sel grid yang rapat untuk membentuk kluster.

### 2.3.2. Algoritma K-Means

Metode *K-Means* merupakan algoritma klasterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya. Ide dasarnya algoritma *K-*

*Means* sangatlah sederhana, yaitu meminimalkan *Sum of Squared Error* antara objek-objek data dengan sejumlah  $k$  titik pusat (Mustafa et al., 2018).



Gambar 2.2 Flowchart Algoritma K-Means (Rahayu et al., 2017)

1. Menentukan jumlah *cluster*. Pada pengelompokan bandwidth ini dibuat 3 *cluster*, menjadi tinggi, sedang dan rendah.
2. Inisialisasi titik pusat *cluster* awal biasa dilakukan dengan berbagai cara, namun yang paling sering dilakukan yaitu dengan cara random. Pusat-pusat *cluster* diberi nilai awal dengan angka-angka random pusat awal *cluster* didapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan random pusat awal dari data.

3. Alokasikan semua data ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek. Dan juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagai berikut :

$$D(i, f) = \sqrt{(X_{1j} - X_{1i})^2 + (X_{2j} - X_{2i})^2 + \dots + (X_{kj} - X_{ki})^2} \dots (3)$$

$D(i, f)$  = Jarak data i ke pusat *cluster* j

$X_{ki}$  = Data ke i pada atribut data ke k

$X_{kj}$  = Data ke j pada atribut data ke k

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat

Algoritma pengelompokan data:

- a. Ambil nilai jarak tiap pusat *cluster* dengan data
  - b. Cari nilai jarak terkecil
  - c. Kelompokkan data dengan pusat *cluster* yang memiliki jarak terkecil
4. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data dalam *cluster* tertentu.

5. Tugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai, atau kembali ke langkah 3 sampai pusat *cluster* tidak berubah lagi.

### 2.3.3. Klasifikasi

Klasifikasi dapat didefinisikan sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori (kelas) yang didefinisikan sebelumnya (Mustafa et al., 2018). Berbeda dengan *Clustering*, klasifikasi digunakan untuk memprediksikan data yang bersifat categorical. Jika pada metode *clustering* mengelompokkan objek ke dalam beberapa kelompok berdasarkan kemiripan antar objek, dimana dalam suatu *cluster* harus berisi objek yang saling mirip dan antar *Cluster* saling tidak mirip namun, klasifikasi ini memerlukan data pelatihan yang sudah diberi label kelompok atau kelas.

## **2.4.RapidMiner**

*RapidMiner* adalah software pengolah data. Menggunakan prinsip dan algoritma *data mining*, *RapidMiner* mengekstrak pola dari *big data*, menggabungkan metode statistik, kecerdasan buatan, dan *database*. *Rapid Miner* memudahkan penggunaannya untuk menghitung data dalam jumlah besar menggunakan operator. Operator ini digunakan untuk mengubah data. Data tersebut digabungkan dengan *node* di operator, dan kemudian Anda tinggal menghubungkannya ke node hasil untuk melihat hasilnya (Rachmat et al., 2017).

*RapidMiner* ditulis dalam bahasa pemrograman *Java* untuk menjalankan tugasnya. Pada dasarnya menggunakan proses *XML-File* yang dihasilkan oleh pengguna dan berisi urutan tugas yang diwakili oleh operator. Lebih dari 500 operator sudah termasuk dalam *RapidMiner*. Fungsi operator mencakup aspek utama dari analisis data seperti pemuatan dan transformasi data, *preprocessing* dan visualisasi data, pemodelan dan evaluasi model (Schlitter et al., 2013).

## **2.5.Davies-Bouldin Index (DBI)**

David L. Davies dan Donald W. Bouldin memperkenalkan sebuah metode yang diberi nama dengan nama mereka berdua, yaitu Davies-Bouldin Index (DBI) yang digunakan untuk mengevaluasi cluster (Nawrin et al., 2017). Evaluasi menggunakan DaviesBouldin Index ini memiliki skema evaluasi internal cluster, dimana baik atau tidaknya hasil cluster dilihat dari kuantitas dan kedekatan antar data hasil cluster (Bates & Kalita, 2016). Davies-Bouldin Index merupakan salah satu metode yang digunakan untuk mengukur validitas cluster pada suatu metode

pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar titik pusat cluster terhadap clusternya. Pengukuran dengan Davies-Bouldin Index ini memaksimalkan jarak inter-cluster antara cluster  $C_i$  dan  $C_j$  dan pada waktu yang sama mencoba untuk meminimalkan jarak antar titik dalam sebuah cluster. Jika jarak inter-cluster maksimal, berarti kesamaan karakteristik antar-masing-masing cluster sedikit sehingga perbedaan antar-cluster terlihat lebih jelas. Jika jarak intra-cluster minimal berarti masing-masing objek dalam cluster tersebut memiliki tingkat kesamaan karakteristik yang tinggi (Wani & Riyaz, 2017).

## **2.6.State of The Art**

### **2.6.1. Penelitian Terdahulu**

Penelitian terdahulu ini menjadi salah satu acuan dalam melakukan penelitian sehingga dapat memperkaya teori yang digunakan dalam mengkaji penelitian yang dilakukan. Ulasan penelitian terkait, dilakukan dengan maksud untuk menganalisis penelitian yang telah dilakukan sebelumnya dipaparkan juga pada tabel 2.1. Penelitian terdahulu dapat dilihat sebagai judul berikut:

#### **A. Implementasi Naive Bayes Pada Proxy Server Untuk Klasifikasi Pengguna Internet**

Nama penulis dari jurnal tersebut adalah Siwi Rahmat Januar, Rakhmadhany Primananda, Mochammad Hannats Hanafi Ichsan dengan tahun penerbitan jurnal yaitu Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 11, November 2018, hlm. 4462-4470, e-ISSN: 2548-964X.

Penelitian ini adalah untuk mengklasifikasikan pengguna internet menggunakan metode atau klasifikasi Naïve Bayes berdasarkan penggunaan internet dengan mengambil logging data dengan aplikasi squid analysis report generator pada proxy server. dan dapat melihat hasil akurasi dari penentuan pengelompokan data yang cukup signifikan yaitu 98,2%. Setelah ditemukan nilai dari kategori naïve bayes, maka pada router mikrotik akan dilakukan pemblokiran website yang mengandung konten negatif.

#### B. Analisis Perilaku Pengguna E-Learning BESMART Melalui Teknik Clustering dengan Algoritma K-Means

Nama penulis jurnal tersebut adalah Neni Miswaningsih dan Nur Insani dengan tahun penerbitan jurnal yaitu seminar nasional matematika dan pendidikan matematika uny 2015.

Jurnal tersebut berisi tentang analisis perilaku pengguna e-learning BESMART dengan metode klasifikasi algoritma K-means berdasarkan penggunaan BESMART di kampus Universitas Negeri Yogyakarta. Dengan menggali Riwayat pengguna BESMART dan hasil akademik mahasiswa, hasil yang diperoleh yaitu mata kuliah yang paling banyak diakses oleh pengguna BESMART adalah mata kuliah dari Fakultas Teknik, Jurusan Pendidikan Teknik Elektronika yaitu Media Digital, sedangkan aktivitas atau modul yang paling banyak dilakukan adalah course. Pada umumnya jumlah pengguna yang mengakses BESMART didalam jam perkuliahan lebih banyak daripada diluar perkuliahan, kecuali pada hari libur dimana jumlah waktu akses relatif sama. Untuk mempersiapkan Ujian

Tengah Semester, mahasiswa mempunyai kesadaran mengakses BESMART lebih rutin untuk mempersiapkan dan menghadapi ujian mereka.

#### C. Analisis Perilaku Pengguna Sistem E-Learning Universitas Gunadarma

Nama penulis jurnal ini adalah Dian Kusuma Ningtyas, Farah Virnawati, Prasetyo, Tirta Paramitta, dan I. Wayan Simri dengan tahun penerbitan yaitu Proceeding, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT 2008), Auditorium Universitas Gunadarma, Depok, 20-21 Agustus 2008, ISSN: 1411-6286.

Penelitian ini adalah analisis perilaku pengguna e-learning di universitas gundharma menggunakan data mining metode algoritma apriori. Universitas Gunadarma yang masih merintis penerapan E-Learning pada kegiatan pembelajarannya sehingga tahapan penyempurnaan merupakan hal yang mutlak. Dalam tulisan ini, kami menggunakan log dari sistem E-Learning sebagai sumber untuk mencari pola perilaku pengguna, kemudian meng-clusternya (metode dari Data Mining) berdasarkan kriteria yang diinginkan.

#### D. Manajemen Bandwidth Berdasarkan Prediksi Perilaku Pengguna Pada Jaringan TCP/IP Dengan Jaringan Syaraf Tiruan

Nama penulis jurnal ini adalah Rama Beta Herdian, Lie Jasa, Linawati dengan tahun penerbitan yaitu Majalah Ilmiah Teknologi Elektro, Vol. 19, No. 1, Januari - Juni 2020, DOI: <https://doi.org/10.24843/MITE.2020.v19i01.P11>.

Penelitian ini terkait tentang prediksi pengguna internet berdasarkan lalu lintas data atau bandwitch pada jaringan TCP/IP dengan menggunakan metode jaringan saraf tiruan (JST). Prediksi perilaku pengguna menggunakan jaringan

syaraf tiruan (JST) menunjukkan hasil yang akurat dalam memprediksi bandwidth yang digunakan pengguna, serta memberikan gambaran pola perilaku pengguna dalam beberapa hari kedepan. Administrator jaringan akan mudah melakukan kontrol jaringan serta membuat keputusan lebih cepat dalam menentukan koneksi prioritas maupun yang bukan prioritas. Penelitian yang dikembangkan menunjukkan hasil, dimana Sistem Aplikasi Pelayanan Kepegawaian(SAPK) adalah koneksi prioritas perilaku pengguna internet di Kantor Regional X BKN hanya memanfaatkan sumber daya internet sebesar 26,8% dari keseluruhan sumber daya jaringan internet yang tersedia. Hasil prediksi menunjukkan 40,3% sumber daya internet lebih banyak digunakan aktifitas streaming video, dan 37% digunakan aktifitas streaming audio. Berdasarkan hasil prediksi tersebut, administrator jaringan mengatur kembali sumber daya jaringan internet dengan alokasi bandwidth setiap jenis perilaku pengguna internet berdasarkan pola baru dari hasil prediksi, yakni alokasi sumber daya jaringan sebesar 50% untuk koneksi SAPK dan 50% berikutnya dialokasikan penggunaan selain SAPK. Kesimpulan penelitian ini menunjukkan bahwa perilaku pengguna dengan aktifitas streaming video merupakan pengguna bandwidth terbesar pada jaringan internet di Kantor Regional X BKN.

#### E. Hybrid Sistem Deteksi Perceraian dan Perilaku Pengguna Internet Menggunakan Algoritma Apriori dan K-Means

Nama penulis jurnal ini adalah Sofika Enggari, Halifia Hendri dengan tempat penerbitan Fakultas Ilmu Komputer, Universitas Putra Indonesia YPTK Padang.

Penelitian ini terkait perilaku pengguna internet dalam mencari informasi dengan kata kunci negatif merupakan salah satu faktor penyebab perceraian sebuah keluarga. Dalam penelitian ini kata kunci yang menjadi data sumber adalah kata Porno, Sex, Homo, dan Lesbian. Data di kumpulkan dari laporan google trend dan angka sidang perceraian dari pengadilan agama kota Padang. Tujuan dalam penelitian ini yaitu untuk mengungkapkan bagaimana algoritma apriori menghitung kemunculan datapencarian kata negatif Porno, Sex, Homo dan Lesbian secara individu maupun berpasangan. Kemudian penelitian ini juga bertujuan untuk mengungkapkan bagaimana algoritma K-Means dapat mengelompokkan pencarian kata negatif dan angka sidang perceraian dari sekelompok data yang ada. Selanjutnya penelitian ini juga bertujuan untuk mengungkapkan bagaimana mengkombinasikan algoritma apriori dengan metode K-Means secara bersama-sama (hybrid) sistem dalam deteksi fakta perceraian dan perilaku pengguna internet. Hasil dari penelitian ini adalah pertama dalam melakukan pengolahan data dengan algoritma apriori sebaiknya data diubah menjadi data angka 0 dan 1, angka diartikan tidak ada kemunculan data dan angka 1 diartikan ada kemunculan data, hal ini dilakukan untuk kemudahan pengolahan data. kedua dari 35 rule yang dihasilkan dengan metode algoritma apriori didapatkan lah 8 rule terbaik itu berarti 8 rule tersebut yang bernilai terbesar dan menggambarkan asosiasi sebenarnya. ketiga untuk menganalisa data hasil pengelompokkan (clustering) sebaiknya data diubah menjadi data bernilai kualitas agar lebih mudah.

#### F. Analisis Kinerja Pegawai Pusbindiklat Peneliti Lipi Berdasarkan Pola Pemanfaatan Internet Melalui Pendekatan Web Usage Mining

Nama penulis jurnal ini adalah Sutrisno Heru Sukoco, Imas Sukaesih Sitanggung, Heru Sukoco dengan tahun penerbitan yaitu Jurnal Penelitian Pos dan Informatika, Pusbindiklat Peneliti LIPI 771/AU1/P2MI-LIPI/08/2017 32a/E/KPT/2017, e-ISSN: 2476-9266 p-ISSN: 2088-9402 DOI:10.17933/jppi.2018.080205.

Penelitian ini terkait pengukuran kinerja pegawai dalam penggunaan layanan internet dapat dilakukan sebagai bagian dari penilaian kinerja. Pendekatan web usage mining melalui pengamatan rekam jejak akses internet yang tersimpan pada proxy server merupakan salah satu cara yang dapat diterapkan untuk memahami perilaku pengguna. Penelitian ini bertujuan untuk mendapatkan gambaran perilaku pegawai Pusbindiklat Peneliti LIPI dalam memanfaatkan layanan internet, mengukur level produktivitas pegawai berdasarkan lama waktu akses terhadap situs yang tidak mendukung pekerjaan dan memetakan kategori situs yang diakses apakah mendukung tugas fungsi jabatannya. Penerapan algoritme clustering K-Means digunakan untuk memudahkan memahami pola akses pengguna. Data yang digunakan adalah log proxy server dan nilai perilaku pegawai Pusbindiklat Peneliti LIPI periode Agustus-Desember 2016. Hasil penelitian menunjukkan pola pemanfaatan internet oleh pegawai Pusbindiklat Peneliti LIPI belum sepenuhnya mendukung tugas fungsi jabatannya. Sekitar 83% pegawai menggunakan internet untuk mengakses situs yang tidak mendukung pekerjaan berada pada level rendah (0-4 jam per minggu). Berdasarkan hasil tersebut dapat disimpulkan bahwa perilaku penggunaan internet yang dilakukan pegawai Pusbindiklat Peneliti LIPI tidak mempengaruhi produktivitas secara signifikan.

### G. Analisa Clustering Phising Dengan K-Means Dalam Meningkatkan Keamanan Komputer

Nama penulis jurnal ini adalah Suhardi Rustam, dengan tahun penerbitan ILKOM Jurnal Ilmiah Volume 10 Nomor 2 Agustus 2018.

Penelitian ini terkait aksi phishing ini semakin marak terjadi. Tercatat secara global, jumlah penipuan bermodus phishing 42% dari modus selain phishing yang dinyatakan dalam website Anti-Phishing Working Group (APWG) dalam laporan bulanannya, mencatat ada 12.845 e-mail baru dan unik serta 2.560 situs palsu yang digunakan sebagai sarana phishing. Selain terjadi peningkatan kuantitas, kualitas serangan pun juga mengalami kenaikan, perlu adanya usaha yang dilakukan oleh para administrator jaringan dalam meningkatkan pengawasan dalam memonitoring aktivitas di jaringan, dalam aksi pencurian data akan melakukan aksi memanipulasi seseorang dengan tampilan situs web tertentu. Dalam penelitian ini sekumpulan dataset akan diklasterisasi menggunakan k-means, algoritma K-Means akan mengelompokkan dataset, menghasilkan identifikasi phishing yang lebih akurat dan bermutu. dengan hasil penelitian ini iterasi=10, K-Fold=2 nilai index davis bouldin = 0.119.

### H. Pengelompokan Pengguna Internet Dengan Metode K-Means Clustering

Nama penulis jurnal ini adalah Disi Amalia P., Bagus Sumargo dengan tahun penerbitan Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta Volume Issue 1, Juni 2019.

Penelitian ini terkait upaya untuk mengelompokkan provinsi berdasarkan persentase pengguna internet. Metode pengelompokan yang digunakan adalah

metode K-Means Clustering. Tujuan dari studi ini adalah mengelompokkan provinsi berdasarkan pengguna internet menjadi tiga cluster untuk mengetahui provinsi-provinsi mana yang memiliki infrastruktur yang baik, sehingga dapat mendukung penggunaan internet. Rata-rata persentase jumlah pengguna internet masing-masing cluster 1, cluster 2, dan cluster 3 adalah 40,91%; 72,27%; dan 57,60%. Perlu adanya pemerataan infrastruktur yang mendukung penggunaan internet pada masyarakat Indonesia.

#### I. Penggunaan Internet Dikalangan Siswa Sd Di Kota Ternate: Suatu Survey, Penerapan Algoritma Clustering Dan Validasi DBI

Nama penulis jurnal ini adalah Firman Tempola, Miftah Muhammad, Abdul Mubarak dengan tahun penerbitan Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK) Vol. 7, No. 6, Desember 2020, hlm. 1153-1160 Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019.

Penelitian ini terkait penggunaan internet dimasyarakat global terus tumbuh, tak hanya terjadi pada masyarakat dewasa melainkan juga pada anak-anak. Internet tidak hanya berdampak pada hal positif melainkan juga pada hal negatif. Di Ternate penggunaan internet terus tumbuh hal ini karena semakin mudah dalam mengakses internet. Namun laporan secara ilmiah mengenai penggunaan internet di Kota Ternate belum ada. Untuk itu, bagaimana mengetahui penggunaan internet dikalangan anak SD di kota Ternate. Penelitian itu bertujuan untuk mencari tahu penggunaan internet di Kota Ternate dengan cara survey secara langsung kepada kalangan anak SD di kota Ternate. Selain itu, data-data dari hasil survey kemudian di cluster dengan menggunakan algoritma k-means clustering. kemudian dilakukan

validasi clustering dengan davies bouldin index. Hasil dari penelitian ini dari 933 responden diperoleh 51,45 % siswa SD di kota Ternate aktif di jejaring sosial dengan 53,70% di whatsapp, 40,30% di instagram dan 27,80% di facebook. Untuk aktivitas ketika membuka youtube terdapat 61,60% sering menonton video di youtube dengan 61,60% video karton, komedi 49,80% dan konten edukasi 28,40%. Sedangkan untuk game online, yang aktif dalam bermain game online yaitu 49,41%. Untuk penerapan algoritma clustering k-means pada 32 sekolah SD di Kota Ternate diperoleh cluster terbaik saat pembagian 4 cluster, hal ini berdasarkan nilai davies bouldin index yang diperoleh sebesar 0,773 lebih kecil dibandingkan dengan pembagian cluster lainnya.

#### J. Analisis Perilaku Pengguna Digilib Unair Dengan Menggunakan Clustering Log Data Mining

Nama penulis jurnal ini adalah Yunus Abdul Halim, penelitian ini terkait Analisis terhadap sikap pengguna digilib UNAIR (ADLN) diketahui melalui akses *log web* ADLN yang berisi informasi tentang akses lokasi, periode waktu, isi informasi dan kapan pengguna mengaksesnya. Semua informasi itu telah disimpan di file *log-access.txt*. Penelitian ini menggunakan data nyata sebanyak 39.086 record pada log web ADLN bulan Mei

2008 hingga 2009. Clustering data mining digunakan untuk mengetahui sikap pengguna dalam analisis ini. Metode clustering ini menggunakan K-mean untuk clustering data. Berdasarkan hasil analisis clustering pada access log data mining Airlangga.Perpustakaan Digital menunjukkan 92,24% pengguna mengakses informasi melalui layanan ADLN di Perpustakaan Airlangga yang memakan waktu

lebih dari 10 menit. Selain itu, hasil analisis menunjukkan 75.64% akses pengguna lebih dari 10 menit dari 92.24%. Padahal, hampir semua pengguna membuka situs ADLN sekitar pukul 10.00 WIB hingga 15.00 WIB. Kondisi tersebut perlu mendapat perhatian jika perpustakaan ingin mengembangkan ADLN sebagai salah satu sumber rich file untuk penilaian di tingkat webometrics. Faktanya bahwa rich file telah memberikan kontribusi besar dalam ADLN, namun ketika kita melihat akses ADLN menunjukkan bahwa terdapat kurangnya akses ADLN melalui internet daripada intranet, sehingga berpengaruh pada penilaian webometrics dari titik akses.



Tabel 2.2 Matriks Penelitian (Lanjutan 1)

No	Peneliti / Tahun	Judul	Ruang Lingkup Penelitian													
			Metode				Algoritma				Sumber Data					
			Clustering	Klasifikasi	Regresi	Jaringan Syaraf Tiruan	K-Means	Apriori	Backpropagation	Naive Bayes	C.45	Log Website	Log Proxy Server	Log Router	Kuisisioner	Google Trends
3	Disi Amalia P dan Bagus Sumargo [2019]	<i>Pengelompokan Pengguna Internet Dengan Metode K-Means Clustering</i>	v				v									v
4	Suhardi Rustam [2018]	<i>Analisa Clustering Phising Dengan K-Means Dalam Meningkatkan Keamanan Komputer</i>	v				v									v
5	Sutrisno Heru Sukoco, Imas Sukaesih Sitanggung dan Heru Sukoco [2017]	<i>Analisis Kinerja Pegawai Pusbindiklat Peneliti Lipi Berdasarkan Pola Pemanfaatan Internet Melalui Pendekatan Web Usage Mining</i>	v				v					v				

Tabel 2.3 Matriks Penelitian (Lanjutan 2)

No	Peneliti / Tahun	Judul	Ruang Lingkup Penelitian													
			Metode				Algoritma				Sumber Data					
			Clustering	Klasifikasi	Regresi	Jaringan Syaraf Tiruan	K-Means	Apriori	Backpropagation	Naive Bayes	C.45	Log Website	Log Proxy Server	Log Router	Kuisisioner	Google Trends
6	Sofika Enggari dan Halifia Hendri [2018]	<i>Hybrid Sistem Deteksi Perceraian dan Perilaku Pengguna Internet Menggunakan Algoritma Apriori dan K-Means</i>	v	v	v		v	v							v	
7	Rama Beta Herdian, Lie Jasa, Linawati [2020]	<i>Manajemen Bandwidth Berdasarkan Prediksi Perilaku Pengguna Pada Jaringan TCP/IP Dengan Jaringan Syaraf Tiruan</i>				v		v						v		

Tabel 2.4 Matriks Penelitian (Lanjutan 3)

No	Peneliti / Tahun	Judul	Ruang Lingkup Penelitian													
			Metode				Algoritma				Sumber Data					
			Clustering	Klasifikasi	Regresi	Jaringan Syaraf Tiruan	K-Means	Apriori	Backpropagation	Naive Bayes	C.45	Log Website	Log Proxy Server	Log Router	Kuisisioner	Google Trends
8	Siwi Rahmat Januar, Rakhmadhany Primananda, Mochammad Hannats [2018]	<i>Implementasi Naive Bayes Pada Proxy Server Untuk Klasifikasi Pengguna Internet</i>		v					v				v			
9	Neni Miswaningsi dan Nur Insani [2015]	<i>Analisis Perilaku Pengguna E-Learning BESMART Melalui Teknik Clustering dengan Algoritma K-Means</i>	v										v			



Pada tabel 2.1, penelitian serupa yang menggunakan metode clustering dilakukan sebanyak 8 penelitian. Sedangkan, metode yang paling terbanyak dari beberapa literatur ditemukan sejumlah 8 penelitian yang menggunakan metode K-Means. Berbeda halnya dengan metode klasifikasi, regresi, dan jaringan syaraf tiruan yang minoritas digunakan oleh peneliti sebelumnya. Begitupun dengan algoritmanya yang minoritas digunakan, yaitu naïve bayes dan apriori. Seperti yang dijelaskan sebelumnya bahwa hal tersebut menjadi salah satu alasan penulis mengangkat penelitian ini.

Dan untuk sumber data yang serupa dengan penelitian yang penulis lakukan sebanyak 1 penelitian, yaitu menggunakan sumber data log router. Hal tersebut menunjukkan bahwa penggunaan sumber data dengan log router jarang dilibatkan. Adapun sumber data mayoritas yang digunakan oleh penelitian sebelumnya adalah sumber data log website, yaitu sebanyak 3 penelitian. Sumber data log proxy server, dataset, dan google trends jarang dilibatkan oleh peneliti sebelumnya.