WEB USAGE MINING MENGGUNAKAN K-MEANS UNTUK MENGETAHUI KECENDERUNGAN AKSES PENGGUNA (Studi Kasus: ganto.co)

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana S-1
pada Program Studi Pendidikan Teknik Informatika

RIZKY MAULANA

2014/14076073

PROGRAM STUDI PENDIDIKAN TEKNIK INFORMATIKA
JURUSAN TEKNIK ELEKTRONIKA
FAKULTAS TEKNIK
UNIVERSITAS NEGERI PADANG

2018

PERSETUJUAN PEMBIMBING

SKRIPSI

Judul	: *Web Usage Mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)
Nama	: Rizky Maulana
NIM	: 14076073/2014
Program Studi	: Pendidikan Teknik Informatika
Jurusan	: Teknik Elektronika
Fakultas	: Teknik

Padang, 4 Oktober 2018

Disetujui oleh

Pembimbing

Nurindah Dwiyani, S.Pd., M.T.
NIP 197801182008122001

Ketua Jurusan

Drs. Hanesman, M.M.
NIP 196101111985031002

PENGESAHAN TIM PENGUJI

Nama : Rizky Maulana
NIM : 14076073/2014

Dinyatakan lulus setelah mempertahankan skripsi di depan Tim Penguji
Program Studi Pendidikan Teknik Informatika
Jurusan Teknik Elektronika
Fakultas Teknik
Universitas Negeri Padang
dengan judul

***Web Usage Mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)**

Padang, 4 Oktober 2018

Tim Penguji

		Tanda Tangan
1. Ketua	: Drs. Efrizon, M.T.	1.
2. Anggota	: Nurindah Dwiyani, S.Pd., M.T.	2.
3. Anggota	: Yeka Hendriyani, S.Kom., M.Kom.	3.

PERNYATAAN

Dengan ini saya menyatakan bahwa

1. karya tulis saya, tugas akhir berupa skripsi dengan judul "*Web Usage Mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)", adalah asli karya saya sendiri;

2. karya tulis ini murni gagasan, rumusan, dan penelitian saya sendiri, tanpa bantuan pihak lain, kecuali dari pembimbing;

3. di dalam karya tulis ini, tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan di dalam naskah dengan menyebutkan pengarang dan dicantumkan pada kepustakaan; dan

4. pernyataan ini saya buat dengan sesungguhnya dan apabila terdapat penyimpangan di dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai norma dan ketentuan hukum yang berlaku.

Padang, 4 Oktober 2018
Yang membuat pernyataan

Rizky Maulana
NIM 14076073/2014

ABSTRAK

Rizky Maulana. 2018. "*Web Usage Mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)".

Situs web Ganto, *ganto.co*, merupakan salah satu sumber informasi bagi mahasiswa, dosen, dan juga khalayak ramai. Setiap kunjungan ke situs web Ganto dicatat di dalam *server log*. Jika dibiarkan begitu saja, *server log* hanya akan memenuhi penyimpanan server. Oleh karena itu, perlu dilakukan tindakan seperti menerapkan *Web usage mining*.

Web usage mining, menggunakan *k-means clustering*, merupakan salah satu cara untuk menjamin kepuasan *user* yang mengakses situs web Ganto, yaitu dengan mengetahui kecenderungan akses *user* melalui analisis *server log*.

Cluster pertama, September 2017. *Cluster* kedua, Oktober 2017, November 2017, Desember 2017, Februari 2018, Maret 2018, dan April 2018. Sedangkan *cluster* ketiga, Juli 2017, Agustus 2017, Januari 2018, Mei 2018, dan Juni 2018. Hasil analisis, Berita merupakan rubrik yang paling sering di-*view*, kedua Artikel, dan ketiga E-Paper. Sedangkan rubrik lainnya (Info Kampus, Sastra Budaya, Ganto TV, dan Ganto Foto) tidak terlalu sering di-*view* oleh *user* atau pengunjung situs web Ganto.

Kata kunci: *Web usage mining*, *clustering*, *k-means*, Ganto

KATA PENGANTAR

Puji dan syukur ke hadirat Allah SWT yang telah melimpahkan rahmat-Nya, sehingga penyusunan skripsi, dengan judul "*Web usage mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)", ini dapat diselesaikan dengan baik.

Penyusunan skripsi ini bertujuan untuk memenuhi salah syarat untuk memperoleh gelar sarjana S-1 pada program studi Pendidikan Teknik Informatika Universitas Negeri Padang.

Penyusunan skripsi ini tidak terlepas dari kerja keras dan kontribusi berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada

1. Bapak Dr. Fahmi Rizal, M.Pd., M.T. selaku Dekan Fakultas Teknik Universitas Negeri Padang,
2. Bapak Drs. Hanesman, M.M. selaku Ketua Jurusan Teknik Elektronika Fakultas Teknik Universitas Negeri Padang,
3. Ahmaddul Hadi, S.Pd., M.Kom. selaku Ketua Program Studi Pendidikan Teknik Informatika Jurusan Teknik Elektronika Fakultas Teknik Universitas Negeri Padang,
4. Ibu Yeka Hendriyani, S.Kom., M.Kom. selaku penasihat akademik dan penguji skripsi,
5. Ibu Nurindah Dwiyani, S.Pd., M.T. selaku pembimbing skripsi,
6. Bapak Drs. Efrizon, M.T. selaku penguji skripsi,
7. Saudara Abdul Hamid selaku Pemimpin Umum Surat Kabar Kampus (SKK) Ganto tahun 2018,

8. Saudara/saudari pengurus Surat Kabar Kampus (SKK) Ganto,
9. Saudara Anton Hilman selaku administrator situs web Ganto (ganto.co),
10. Rekan-rekan seperjuangan pada program studi Pendidikan Teknik Informatika tahun masuk 2014, dan
11. Pihak-pihak lainnya yang tidak dapat penulis sebutkan.

Penulis mohon maaf apabila terdapat kekurangan ataupun kesalahan dalam penyusunan skripsi ini. Oleh sebab itu, diharapkan kritik maupun saran demi perbaikan dalam penyusunan karangan ilmiah serupa di masa yang akan datang.

Akhir kata, semoga Allah SWT membalas kebaikan seluruh pihak yang telah membantu dalam penyusunan skripsi ini. Semoga skripsi ini dapat memberikan manfaat bagi penulis maupun pembaca, *Aamiin*.

Padang, Oktober 2018

Penulis

DAFTAR ISI

Halaman

ABSTRAK i
KATA PENGANTAR ii
DAFTAR ISI iv
DAFTAR GAMBAR v
DAFTAR TABEL vi
DAFTAR LAMPIRAN vii

BAB I PENDAHULUAN
A. Latar Belakang Masalah 1
B. Identifikasi Masalah 4
C. Batasan Masalah 4
D. Rumusan Masalah 5
E. Tujuan Penelitian 5
F. Manfaat Penelitian 5

BAB II KAJIAN PUSTAKA
A. Situs Web 7
B. Surat Kabar Kampus (SKK) Ganto 8
C. *Data Mining* 10
D. K-means 13
E. *Web Mining* 15
1. *Web Usage Mining* 15
F. RapidMiner Studio 18
G. Penelitian yang Relevan 22

BAB III METODE PENELITIAN
A. Desain Penelitian 24
B. Objek Penelitian 27
C. Jenis dan Sumber Data 27
D. Teknik Analisis Data 27

BAB IV HASIL DAN PEMBAHASAN
A. *Web Usage Mining* Menggunakan *K-means Clustering* 28
1. *Preprocessing* 29
2. *Pattern Discovery* 29
a. Perhitungan Manual 29
b. Perhitungan Menggunakan Aplikasi RapidMiner Studio 51
3. *Pattern Analysis* 56

BAB V PENUTUP
A. Simpulan 60
B. Saran 60

DAFTAR RUJUKAN 61

DAFTAR GAMBAR

Gambar **Halaman**

Gambar 1. *Flowchart* algoritme k-means 13
Gambar 2. Proses *Web usage mining* 16
Gambar 3. Logo retina RapidMiner 18
Gambar 4. Memilih *template* 20
Gambar 5. *Design tab* dengan *blank process template* 20
Gambar 6. *Result tab* 21
Gambar 7. *Auto Model tab* 21
Gambar 8. Menambahkan data ke dalam *Repository* 51
Gambar 9. Memilih tempat data tersimpan 52
Gambar 10. Mencari dan memilih data yang akan ditambahkan 52
Gambar 11. Menentukan *sheet*, *cell range*, dan *header row* 52
Gambar 12. Mengatur kolom *bulan* sebagai *label* 53
Gambar 13. Memilih *role* kolom yang dipilih sebagai *label* 53
Gambar 14. Hasil dari pengaturan kolom *bulan* sebagai *label* 53
Gambar 15. Memilih tempat dan memberi nama data yang akan disimpan 54
Gambar 16. Tinjauan data yang akan di *cluster* 54
Gambar 17. Desain proses *k-means clustering* 54
Gambar 18. *Parameter* dari *k-means clustering* 55
Gambar 19. *Cluster model* 55
Gambar 20. Perincian *cluster* dan anggotanya 56
Gambar 21. Sebaran anggota *cluster* 56
Gambar 22. *Cluster* pertama 57
Gambar 23. *Cluster* kedua 58
Gambar 24. *Cluster* ketiga 58

DAFTAR TABEL

Tabel **Halaman**

Tabel 1. Kepengurusan SKK Ganto 2018 9
Tabel 2. Sampel *server log file* 24
Tabel 3. Tabulasi sampel *server log file* 25
Tabel 4. Sampel *server log* setelah proses *data cleaning* 26
Tabel 5. *Server log* situs web Ganto Juli 2017–2018 28
Tabel 6. *Centroid* awal 29
Tabel 7. Hasil iterasi ke-1 35
Tabel 8. *Centroid* baru berdasarkan hasil iterasi ke-1 38
Tabel 9. Hasil iterasi ke-2 43
Tabel 10. *Centroid* baru berdasarkan hasil iterasi ke-2 44
Tabel 11. Hasil iterasi ke-3 50
Tabel 12. *Centroid* yang ditentukan di RapidMiner Studio 55
Tabel 13. Hasil *k-means clustering* 57

DAFTAR LAMPIRAN

Lampiran		**Halaman**
Lampiran 1.	Surat Izin Melakukan Penelitian (SKK Ganto)	62
Lampiran 2.	Balasan Surat Izin Melakukan Penelitian (SKK Ganto).....	63
Lampiran 3.	Surat Izin Melakukan Penelitian (BAK UNP/UPT PTIK UNP) ...	64
Lampiran 4.	Balasan Surat Izin Melakukan Penelitian (BAK UNP/UPT PTIK UNP)	65
Lampiran 5.	Kode-sumber WUM *k-means clustering* (Java)	66

BAB I

PENDAHULUAN

A. Latar Belakang Masalah

Perkembangan teknologi informasi saat ini sangat pesat dan berpengaruh signifikan terhadap kehidupan manusia. Seiring perkembangan tersebut, manusia juga semakin membutuhkan informasi. Informasi yang dibutuhkan dapat beragam, seperti informasi sosial, ekonomi, pendidikan, olahraga, politik hingga gaya hidup. Pada zaman sekarang, informasi tersebut dapat diperoleh dengan cukup mudah melalui surat kabar, majalah, televisi, radio, dan media informasi lainnya.

Di era digital seperti sekarang ini, orang lebih memilih mencari informasi di Internet melalui telepon genggam, komputer desktop, dan perangkat komputer yang sejenis. Informasi tersebut biasanya dimuat dalam situs-situs web sehingga mudah diakses kapan saja dan di mana saja–selama terhubung ke Internet. Aktivitas manusia sebagai pengguna (*user*) yang mengakses situs web semakin meningkat, hal tersebut dipengaruhi oleh kebutuhan informasi yang tinggi.

Situs web adalah keseluruhan halaman web dalam suatu domain yang mengandung informasi (Yuhefizar, dkk. dalam Syamnugroho dkk., 2014:29). Pada tahun 1991, Tim Berners-Lee (penemu WWW atau web) menyatakan bahwa sebuah situs web awalnya hanya untuk memudahkan tukar-menukar dan memperbarui informasi antar sesama peneliti di tempat mereka bekerja.

Situs web merupakan halaman web yang saling berhubungan dan pada umumnya berada di dalam server web. Sebuah situs web biasanya berisikan informasi yang disediakan perorangan, kelompok, atau organisasi. Ganto merupakan salah satu organisasi tersebut.

Ganto berperan penting dalam menyediakan informasi yang edukatif, komunikatif, dan relevan bagi mahasiswa, dosen, maupun khalayak ramai. Setiap kunjungan ke situs web Ganto dicatat di dalam suatu *file* yang tersimpan–biasanya–di sisi server, yaitu *server log*. *File* tersebut berisi catatan kunjungan, seperti alamat IP, waktu akses, halaman yang diakses, dan-lain-lain.

Server log situs web Ganto hanya akan bertambah banyak dan memenuhi penyimpanan *server* jika dibiarkan begitu saja. Oleh karena itu, perlu dilakukan tindakan seperti menerapkan *Web usage mining. Web usage mining* berguna untuk mengetahui ketertarikan atau kecenderungan akses *user* yang mengakses situs web Ganto. Setelah dilakukan *Web usage mining*, maka *server log* tidak tersimpan begitu saja, tetapi menjadi lebih bermanfaat.

Web usage mining merupakan penerapan teknik *data mining* untuk menemukan pola penggunaan atau kecenderungan akses *user* terhadap sebuah situs web. Sumber data atau *data set* dalam *Web usage mining*, yaitu *server level collection*, *proxy level collection*, dan *client level collection*. Peneliti dapat menggunakan salah satu dari ketiga *data set* atau mengkombinasikan ketiganya.

Pada umumnya, sumber data yang digunakan adalah *server level collection* atau *server log* atau *web log file* karena *server log* banyak menyimpan riwayat akses atau *access log* dari *user* yang mengakses situs web. *Server log* atau *web log file* berisi data-data *user* yang mengunjungi situs web. Data-data tersebut seperti *ip address, log name, username, access time, request method, request status code, byte sent, user agent,* dan lain-lain. Situs web Ganto memiliki *server log* yang dapat dimanfaatkan untuk melakukan *Web usage mining*.

Dalam prosesnya, *Web usage mining* ini menggunakan metode *clustering. Clustering* merupakan suatu metode untuk mengelompokkan objek yang memiliki kemiripan karakter. Pengelompokan (*clustering*) merupakan bagian dari ilmu data mining yang bersifat tanpa arahan (*unsupervised*). Dalam *clustering*, data yang memiliki kesamaan dimasukkan ke dalam *cluster* yang sama, sedangkan data yang tidak memiliki kesamaan dimasukkan dalam *cluster* yang berbeda.

Clustering dapat dilakukan dengan menggunakan beberapa algoritma. Salah satu algoritma tersebut adalah *k-means*. *K-means* merupakan salah satu algoritma *clustering* yang paling mudah, cepat, tidak memerlukan banyak memori, serta bisa diaplikasikan pada *data set* berukuran kecil maupun besar.

Dari uraian latar belakang ini, maka penulis ingin melakukan penelitian terhadap *server log* situs web Ganto, *ganto.co,* dengan menerapkan *Web usage mining* menggunakan *k-means clustering*. Oleh karena itu, penelitian

ini diberi judul "*Web Usage Mining* Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co)".

B. Identifikasi Masalah

Berdasarkan latar belakang masalah, maka permasalahan diidentifikasi sebagai berikut:

1. Belum diketahui pola atau kecenderungan akses *user* yang mengakses situs web Ganto.
2. *Server log* situs web Ganto hanya akan bertambah banyak dan tidak berguna jika tidak dilakukan suatu tindakan untuk mencari dan mengetahui informasi apa yang dapat dimanfaatkan dari data-data tersebut.
3. Belum ada penelitian yang menerapkan *Web usage mining* untuk mengetahui pola atau kecenderungan akses *user* situs web Ganto.

C. Batasan Masalah

Berdasarkan identifikasi masalah, maka masalah dalam penelitian ini dibatasi sebagai berikut:

1. Situs web yang akan diteliti adalah situs web Surat Kabar Kampus (SKK) Ganto UNP.
2. *Cluster* yang akan dibentuk, yaitu berdasarkan rubrik atau menu utama yang ada di dalam situs web Ganto.
3. Penelitian ini menerapkan *Web usage mining* menggunakan *k-means clustering* untuk mengetahui kecenderungan akses *user* yang mengakses situs web Ganto.

D. Rumusan Masalah

Berdasarkan batasan masalah, maka dirumuskan permasalahan sebagai berikut:

1. Bagaimana kecenderungan akses *user* situs web Ganto?
2. Bagaimana *cluster* yang terbentuk berdasarkan rubrik atau menu utama dalam situs web Ganto?
3. Bagaimana hasil dari *Web usage mining* menggunakan *k-means clustering*?

E. Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mengetahui pola atau kecenderungan akses *user* situs web Ganto.
2. Mengetahui *cluster* yang terbentuk berdasarkan rubrik atau menu utama dalam situs web Ganto.
3. Mengetahui hasil dari *Web usage mining* menggunakan *k-means clustering*.

F. Manfaat Penelitian

Berdasarkan tujuan masalah, maka diharapkan penelitian ini dapat memberikan manfaat, diantaranya:

1. Memberikan sumbangan ilmu pengetahuan, memperkaya hasil penelitian yang telah ada, dan mengetahui informasi yang mungkin di dapat dari penerapan *Web usage mining* dengan *k-means clustering*.

2. Sebagai rujukan untuk penelitian serupa atau penelitian lanjutan tentang *Web usage mining*, *clustering*, ataupun *k-means*.
3. Sebagai saran ataupun masukkan demi peningkatan layanan dan kinerja dari situs web Ganto dalam memberikan informasi yang edukatif dan komunikatif.

BAB II

KAJIAN PUSTAKA

A. Situs Web

Situs web adalah keseluruhan halaman web yang terdapat dalam sebuah domain yang mengandung informasi (Yuhefizar, dkk. dalam Syamnugroho, 2014:29). Sedangkan menurut Sutarman dalam Hapsari (2010:49), situs web atau WWW (*World Wide Web*) adalah sebuah jaringan komputer *client-server* yang menggunakan *software* khusus.

Penemu situs web adalah Sir Timothy John "Tim" Berners-Lee, sedangkan situs web yang tersambung dengan jaringan pertama kali muncul pada tahun 1991. Maksud dari Tim ketika merancang situs web adalah untuk memudahkan tukar menukar dan memperbarui informasi pada sesama peneliti di tempat ia bekerja. Pada tanggal 30 April 1993, CERN (tempat di mana Tim bekerja) mengumumkan bahwa WWW dapat digunakan secara gratis oleh publik.

Situs web biasanya ditempatkan pada server web. Sebuah server web umumnya telah dilengkapi dengan perangkat-perangkat lunak khusus untuk menangani pengaturan nama domain, serta menangani layanan atas protokol HTTP yang disebut sebagai HTTP *Server* seperti Apache HTTP *Server*, *Internet Information Services* (IIS), nginx, dan sebagainya.

Situs web dapat memiliki banyak fungsi dan dapat digunakan dalam berbagai bidang, seperti situs web pribadi, situs web korporat untuk sebuah perusahaan, situs web pemerintah, situs web organisasi, dan sebagainya. Situs

web biasanya didedikasikan untuk topik atau tujuan tertentu, mulai dari hiburan dan jejaring sosial hingga penyediaan berita dan edukasi.

B. Surat Kabar Kampus (SKK) Ganto

Surat Kabar Kampus (SKK) Ganto atau "Ganto" saja, merupakan sebuah koran kampus yang berbasis di Universitas Negeri Padang (UNP). Ganto yang awalnya bernama TRIDHARMA, merupakan sebuah media yang efektif untuk menyampaikan informasi, gagasan, dan juga aspirasi mahasiswa UNP. Menurut Rektor UNP, seperti dikutip dari situs web Ganto, "Ganto tidak hanya sebagai media penerbitan tetapi juga berfungsi sebagai sebuah wadah yang edukatif dan komunikatif".

Pada masa awalnya, Ganto menerbitkan koran sebulan sekali dengan 8 (delapan) halaman. Kemudian pada tahun 1990–1992, Ganto terbit sekali dalam sebulan dengan 12 (dua belas) halaman lalu menjadi 24 (dua puluh empat) halaman pada tahun 2002.

Tahun 1994 merupakan masa keemasan Ganto. Hal tersebut dibuktikan dengan penghargaan dari Dikti sebagai juara ke-2 Pers Mahasiswa Tingkat Nasional. Ketua penyunting saat itu adalah Abdul Maujud dengan anggota Ermanto, Yurnaldi, Syafwan, Nandik Suparyono, dan Zulkarnain.

Kepengurusan Ganto pada masa awalnya dijalankan oleh dosen dan juga mahasiswa. Dosen sebagai pengayom dan pendidik, sedangkan mahasiswa bertugas untuk mencari dan menulis berita. Keterlibatan mahasiswa diharapkan agar Ganto dapat menjadi tempat untuk menyuarakan aspirasi dan

juga menjadi media belajar dalam tulis-menulis. Adapun kepengurusan Ganto tahun 2018 adalah sebagai berikut.

Tabel 1. Kepengurusan SKK Ganto 2018

Bagian	Nama
Pemimpin Umum	Abdul Hamid
Sekretaris Umum	Putri Radilla
Pemimpin Redaksi	Laila Marni
Ka. Penelitian dan Pengembangan	Maida Yusri
Bendahara Umum	Oktri Diana Putri
Pemimpin Usaha	Debi Gunawan
Redaktur Pelaksana	Irza Suarni
Redaktur Berita	1. Dwi Agustini 2. Nadilla Aprisia
Redaktur Bahasa dan Sastra	Yanda Dwi Kurnia
Redaktur Tulisan	Deby Purnama Sari
Redaktur Artistik	Venny Sindya F
Layouter	M Agus
Fotografer	Akbar Wahyu
Riset	Monalisa
Perpustakaan dan Kearsipan	Ariska Novia Mariady
Iklan	Ariska Novia Mariady
Sirkulasi dan Percetakan	Hega Dwi Dian Dolla
Reporter	Niswasani

Sumber: *ganto.co*, diakses pada 7 Mei 2018.

Dalam perkembangannya, Ganto tidak hanya memuat berita atau informasi dalam bentuk media cetak (koran). Pada periode 2005–2006, Ganto merilis situs web dengan nama domain *ganto.unp.ac.id* di bawah pimpinan Adriyanto. Kemudian pada periode-periode selanjutnya, Ganto mengalami beberapa perubahan nama domain dari *ganto.unp.ac.id* menjadi *ganto.web.id*, *ganto.or.id*, dan sekarang menjadi *ganto.co*.

Sebagai koran kampus, Ganto tidak hanya memberikan informasi seputar kampus, tetapi juga informasi lainnya seperti sastra dan budaya, artikel, dan

sebagainya. Situs web Ganto menjadi alternatif bagi mahasiswa, dosen, warga kampus lainnya, dan juga masyarakat untuk mengakses berbagai informasi.

C. *Data Mining*

Larose (2014:2) menjelaskan bahwa ”*Data mining is the process of discovering useful patterns and trends in large data sets* (*Data mining* adalah proses untuk menemukan pola yang berguna dan kecenderungan dalam *data set* yang besar).” Sedangkan menurut Gartner Group dalam Larose (2014:xi)

> *Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.*

Data mining adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam *database*. Menurut Hand, *et al*., dalam Hamimi, 2014:5), *data mining* adalah analisis terhadap data (biasanya data yang berukuran besar) menggunakan metode tertentu untuk menemukan hubungan yang jelas serta menyimpulkan yang belum diketahui sebelumnya agar berguna bagi si pemilik data tersebut.

Penggalian data (*data mining*) adalah ekstraksi pola yang menarik dari data dalam jumlah besar. Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya, dan berguna. Pola yang disajikan haruslah mudah dipahami, berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, berguna, dan baru.

Penggalian data memiliki beberapa nama alternatif, meskipun definisi pastinya berbeda, seperti KDD (*Knowledge Discovery in Database*), analisis pola, dan sebagainya. Penggalian data diperlukan saat data yang tersedia

terlalu banyak (misalnya data yang diperoleh dari sistem basis data perusahaan, *e-commerce*, data saham, dan sebagainya), tetapi tidak diketahui pola yang bisa didapatkan.

Teknik *data mining* yang paling sering digunakan menurut Larose (2014:8), yaitu (1) *Description*, (2) *Estimation*, (3) *Prediction*, (4) *Classification*, (5) *Clustering*, dan (6) *Association*. Berikut adalah uraian singkat dari teknik-teknik *data mining* tersebut.

1. *Description*

   Adakalanya peneliti dan analis sekadar ingin mengetahui cara untuk *mendeskripsikan* pola dan kecenderungan yang ada di dalam suatu data. Deskripsi pola dan kecenderungan tersebut sering mempengaruhi penjelasan pola dan kecenderungan itu sendiri. Deskripsi berkualitas tinggi dapat diselesaikan menggunakan *explanatory data analysis*, yaitu eksplorasi data dengan metode grafis dalam pencarian pola dan kecenderungan.

2. *Estimation*

   Dalam estimasi, peneliti memperkirakan nilai dari variabel target numerik menggunakan sekumpulan variabel numerik dan/atau variabel prediksi kategori. Model dibangun menggunakan catatan *lengkap*, yang menyediakan nilai dari variabel target. Kemudian, untuk observasi baru, estimasi nilai variabel target yang terbentuk berdasarkan nilai dari variabel prediksi. Contoh, memperkirakan jumlah poin per gim yang dapat dicetak oleh LeBron James saat *double-teamed*.

3. *Prediction*

Prediksi serupa dengan klasifikasi dan estimasi, yang membedakan adalah hasilnya dapat dilihat di masa yang akan datang. Sejumlah metode dan teknik klasifikasi dan estimasi, dalam kondisi tertentu, juga digunakan dalam prediksi. Contoh, memprediksi pemenang Piala Dunia 2022 berdasarkan perbandingan statistik-tim.

4. *Classification*

Klasifikasi serupa dengan estimasi, yang membedakan adalah variabel target berupa kategori daripada numerik. Dalam klasifikasi, terdapat target variabel kategori, seperti *golongan pendapatan*, yang dapat dibagi ke dalam tiga kelas atau kategori: pendapatan tinggi, pendapatan menengah, pendapatan rendah.

5. *Clustering*

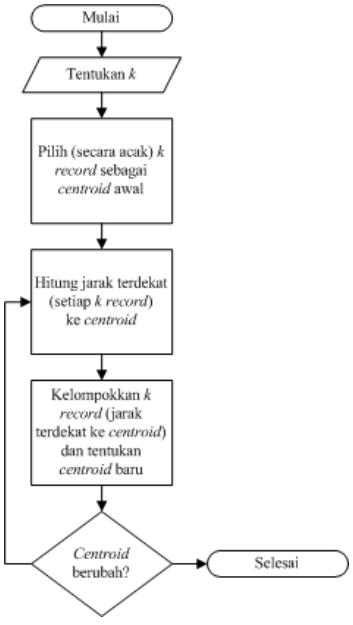
Clustering (penggugusan) berkenaan dengan pengelompokan *record*, observasi, atau kasus menjadi kelas dengan objek-objek yang mirip. Sebuah *cluster* (gugus) adalah kumpulan *record* yang serupa satu sama lain, tetapi takserupa dengan *record* dalam gugus lain. Penggugusan berbeda dengan klasifikasi karena tidak adanya variabel target. Proses penggugusan tidak mencoba untuk mengklasifikasikan, memperkirakan, atau memprediksi nilai dari variabel target. Namun, algoritme penggugusan mencari dan menyegmentasikan seluruh data ke dalam subgolongan atau gugus, memaksimalkan kemiripan *record* dalam satu gugus, dan meminimalkan kemiripan *record* antargugus.

6. *Association*

Teknik asosiasi dalam *data mining* adalah tugas untuk menemukan atribut mana yang *go together* (berkaitan). Tugas asosiasi dalam dunia bisnis umumnya, diketahui juga sebagai *affinity analysis* atau *market basket analysis*, digunakan untuk menemukan *rules* (kebiasaan), untuk mengukur hubungan antara dua atau lebih atribut. *Association rules* merupakan bentuk "*If* ***antecedent*** *then* ***consequent***", dan ukuran dari *support* dan *confidence* yang berasosiasi dengan *rule*.

D. K-means

Han, *et al.,* (2012:451) menjelaskan bahwa teknik partisi *centroid-based* (berbasis *centroid*) menggunakan *centroid* dari suatu *cluster*, $\boldsymbol{C_i}$, yang merepresentasikan *cluster* tersebut. Secara konseptual, *centroid* suatu *cluster* merupakan *center point* (titik pusat) dari *cluster* itu sendiri. Berikut ini *flowchart* algoritme k-means (Larose, *et al.,* 2014:215).



Gambar 1. *Flowchart* algoritme k-means

1. Tentukan banyak *cluster k* yang akan dibentuk dari *data set*.

   Larose menjelaskan bahwa banyaknya *cluster* yang akan dibentuk ditentukan oleh peneliti. Han, *et al.* (2012:452) menjelaskan bahwa "*'How does the k-means algorithm work?' ... First, it randomly selects k* (banyaknya *cluster* yang akan dibentuk) *of the objects in D* (*data set* yang akan diolah), ..."

2. Pilih *k record* untuk dijadikan sebagai lokasi pusat awal *cluster*.

   Penentuan *centroid* pada tahap awal, umumnya dilakukan secara acak (Santosa dalam Ong, 2013:14). Berdasarkan Larose, *et al*. (2014:215), pilih secara acak *k record* untuk dijadikan *cluster center* (m_i) awal. Sedangkan untuk iterasi selanjutnya menggunakan rumus berikut (Anonim dalam Hamimi, 2014:10).

   $$C(i) = \frac{X_1 + X_2 + \cdots + X_n}{\sum x} \quad \text{.....(1)}$$

   Keterangan:

   $C(i)$ = *centroid* dari *cluster* ke-i

   X_1 = nilai *record* ke-1

   X_2 = nilai *record* ke-2

   X_n = nilai *record* ke-n

   $\sum x$ = banyaknya anggota *cluster*

3. Cari jarak terdekat ke pusat *cluster* pada setiap *record*.

   Jarak terdekat dengan *centroid* yang digunakan adalah *Euclidean Distance*, dicari dengan menggunakan persamaan berikut (Santosa dalam Ong, 2013:14).

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

Keterangan:

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

X_{ki} = data ke *i* pada atribut data ke *k*

X_{kj} = titik pusat ke *j* pada atribut ke *k*

4. Hitung rasio antara *Between-Cluster Variation* (BCV) dan *Within-Cluster Variation* (WCV).

   $BCV = d(m_1 - m_2) + d(m_1 - m_3) + d(m_2 - m_3)$ .....(3)

   $\frac{BCV}{WCV}$ .....(4)

5. Untuk setiap *k cluster*, temukan *cluster centroid*-nya, dan perbarui lokasi *cluster* berdasarkan nilai *centroid* baru.

6. Ulangi langkah ke-3–ke-5 hingga *convergence* (*centroid* tidak berubah).

E. *Web Mining*

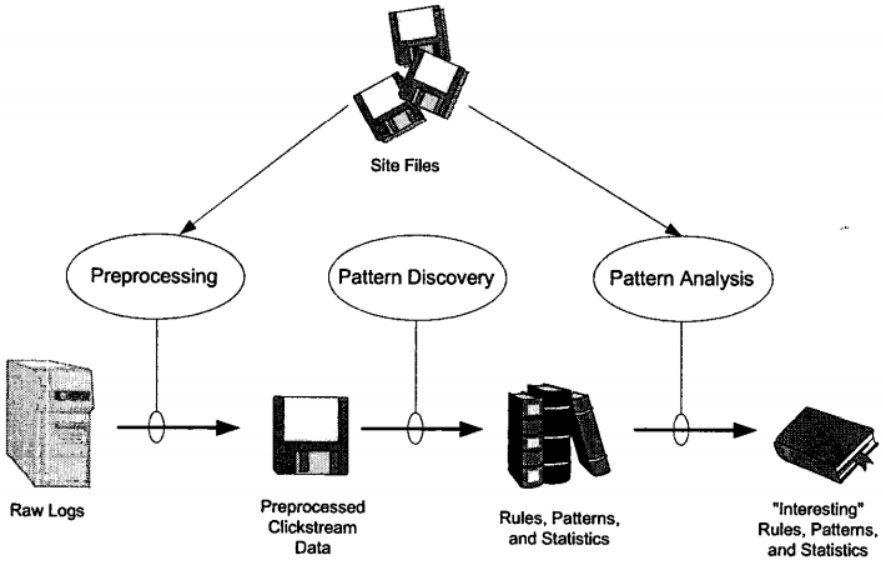
Web mining merupakan ekstraksi pola-pola penting dan bermanfaat namun tersimpan secara implisit pada kumpulan data yang relatif besar pada layanan *World Wide Web*. *Web mining* dibagi menjadi tiga, yaitu *Web content mining, Web structure mining,* dan *Web usage mining*.

1. *Web Usage Mining*

Web usage mining merupakan penerapan teknik *data mining* untuk menemukan pola penggunaan atau pun kecenderungan akses *user* terhadap sebuah situs web. Sumber data yang akan di-*mining*, yaitu *server level collection, proxy level collection,* dan *client level collection*.

Pada umumnya, sumber data yang digunakan adalah *server level collection* atau *server log* atau *web log file* karena *server log* banyak menyimpan riwayat *access log* dari *user* yang mengakses situs web. *Server log* atau *web log file* berisi data-data *user* yang mengunjungi situs web. Data-data tersebut seperti *ip address*, *log name*, *username*, *access time*, *request method*, *request status code*, *byte sent*, *user agent*, dan lain-lain.

Web usage mining menjadi hal yang sangat penting untuk pengelolaan situs web, menciptakan situs web yang adaptif, layanan dukungan dan bisnis, personalisasi, serta analisis alur trafik pada jaringan. Secara garis besar, proses *Web usage mining* terbagi menjadi tiga, yaitu (1) *preprocessing*, (2) *pattern discovery*, (3) *pattern analysis* (Srivastava, *et al*., 2000:12).



Gambar 2. Proses *Web usage mining* (Srivastava, *et al*., 2000:15)

a. *Preprocessing*

Tahap ini merupakan langkah awal untuk memulai proses keseluruhan dari *Web usage mining*. Tahap ini sangat penting dikerjakan untuk melakukan standardisasi data dan menghilangkan bagian–bagian data yang tidak diperlukan dalam proses *mining*. Berikut adalah hal-hal yang dilakukan dalam *preprocessing*.

1) *Data cleaning*

Tahap ini bertujuan untuk membersihkan *server log* dari data yang tidak relevan dengan proses mining, seperti data multimedia (misalnya *.png, *.gif, *.pdf, dan lain-lain), *http request method* selain GET, *http status code* atau *request status* selain 200, kolom *rfc1413*, *authuser*, dan juga *bytes*.

2) *User identification*

Proses untuk mengidentifikasi *user* yang mengakses suatu situs web. Hal tersebut ditentukan dengan menyeleksi IP dan UA (*User Agent*) yang sama. Jika IP dan UA-nya sama, maka dihitung sebagai satu pengguna.

3) *Session identification*

Proses untuk mengidentifikasi *session* dari tiap-tiap *user* yang mengakses situs web. Jika seorang *user* mengakses suatu halaman (termasuk *page loading* dan membaca konten) web dalam waktu tidak lebih dari 30 menit, maka itu dihitung sebagai satu *session*.

b. *Pattern discovery*

Tahap kedua dari *Web usage mining*, yaitu pencarian pola akses yang dilakukan oleh *user* (pengguna) situs web. Tahap ini sangat penting dan menentukan keluaran dari proses *Web usage mining*. Beberapa teknik yang sering digunakan dalam proses *pattern discovery* ini adalah *statistical analysis*, *association rules*, *clustering*, *classification*, *sequential pattern*, dan *dependency modeling*.

c. *Pattern analysis*

Proses ini merupakan akhir dari *Web usage mining*. Pada tahap ini dilakukan proses visualisasi dari *pattern discovery*. Hasilnya disajikan sesuai dengan kebutuhan dari situs web yang dilakukan *Web usage mining*.

Penyajian hasil tersebut dapat mempermudah untuk mengambil tindakan, misalnya mengubah tampilan situs web, meningkatkan dan menambah layanan situs web, meningkatkan kinerja web server, atau bahkan mengurangi dan meminimalisasi menu-menu pada situs web agar terlihat lebih rapi.

F. RapidMiner Studio

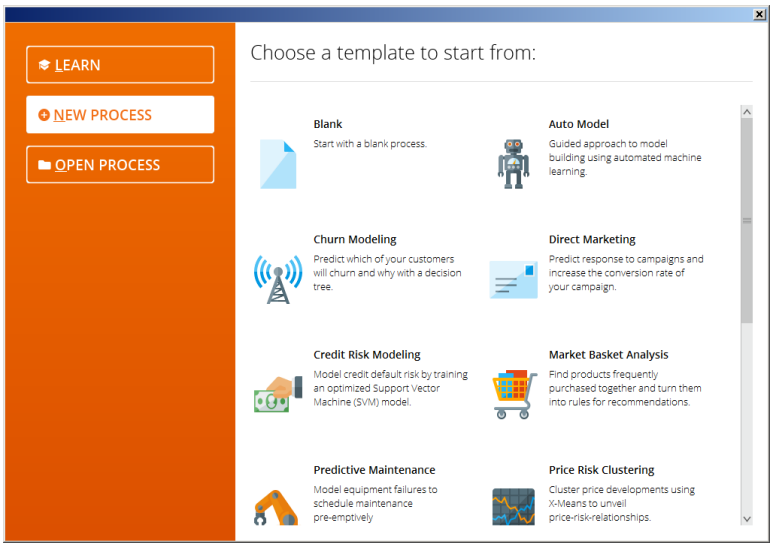
Gambar 3. Logo retina RapidMiner

RapidMiner Studio adalah *visual workflow designer* yang membantu peneliti data agar lebih produktif, yaitu dari membuat prototipe gagasan dengan cepat hingga perancangan model prediksi tugas kritis (dikutip dari *rapidminer.com*).

RapidMiner merupakan platform perangkat lunak ilmu data yang dikembangkan oleh perusahaan bernama sama (RapidMiner), dengan menyediakan lingkungan terintegrasi untuk persiapan data, pembelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis yang bersifat prediksi.

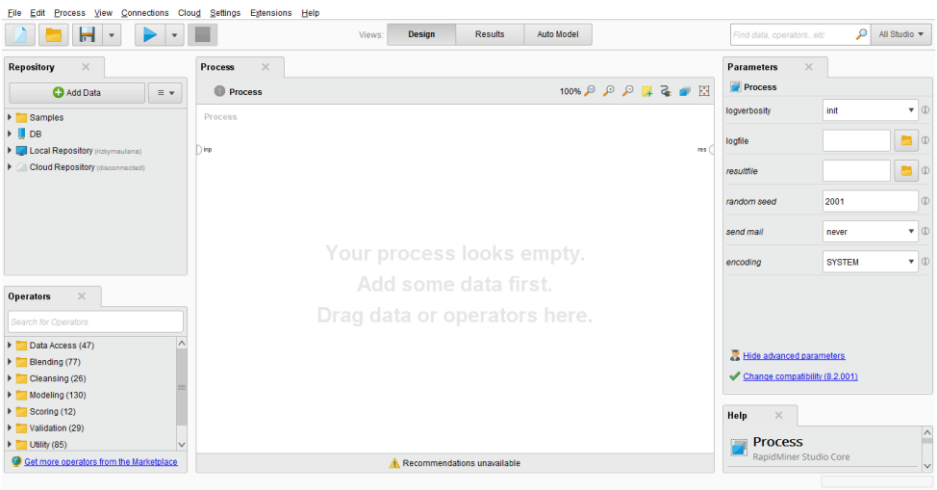
RapidMiner sebelumnya dikenal sebagai YALE (*Yet Another Learning Environment*), mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer dari Unit Kecerdasan Buatan Universitas Teknik Dortmund. RapidMiner Studio yang digunakan adalah versi gratis, yang terbatas untuk 1 *logic processor* dan 10.000 baris data, tersedia di bawah lisensi AGPL.

Berikut adalah beberapa *screenshot* fitur-fitur yang ada di dalam RapidMiner Studio.



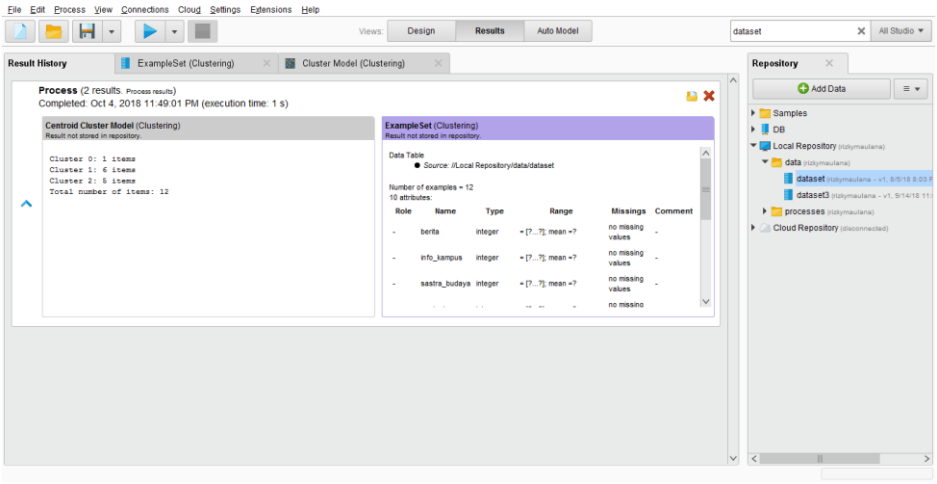
Gambar 4. Memilih *template*

Pengguna dapat memilih *template* yang hendak digunakan, *blank process*, *auto model*, *market basket analysis*, dan lain-lain.



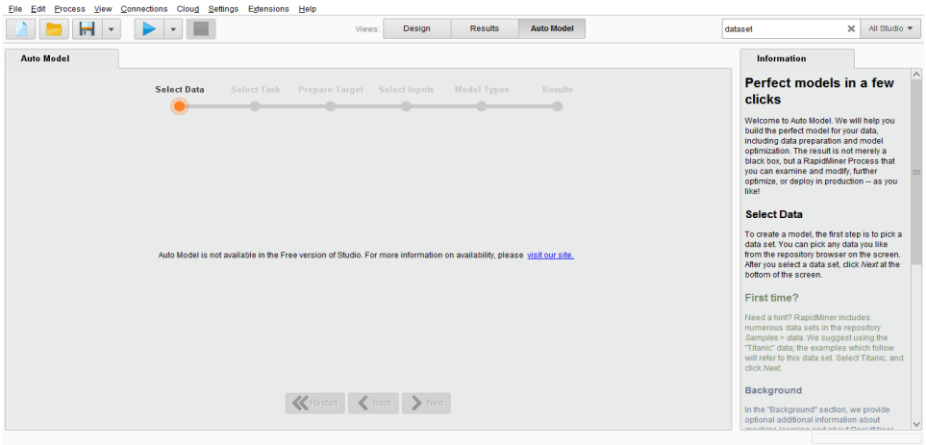
Gambar 5. *Design tab* dengan *blank process template*

Anatarmuka awal RapidMiner Studio–*design tab*–terdapat panel *repository*, panel *operators*, panel *parameters*, dan panel *help*.



Gambar 6. *Result tab*

Menampilkan hasil dari proses atau tugas yang dilakukan (pada gambar di atas: hasil *k-means clustering*).



Gambar 7. *Auto Model tab*

Jika hendak menggunakan *auto model*, yaitu *user* tidak harus merancang proses, melainkan dilakukan secara otomatis. Namun, fitur ini tidak tersedia di RapidMiner Studio versi gratis.

G. Penelitian yang Relevan

Berikut adalah beberapa penelitian yang terkait dengan *k-means clustering* dan *Web usage mining*.

1. Fauzanu, dkk. (2017), berjudul "Analisis *Web Usage Mining* Menggunakan Teknik *K-means Clustering* dan *Association Rule* (Studi Kasus: www.olexa.com)". Hasil penelitian tersebut adalah (a) nilai *k* (banyaknya *cluster* yang akan dibentuk) optimum sama dengan delapan, (b) pola kunjungan *user* dengan frekuensi paling tinggi, yaitu dari menu Layanan menuju menu Tentang Kami (*cluster* 1 dengan *confidence* 60%, *cluster* 4 dengan *confidence* 82%, dan *cluster* 7 dengan *confidence* 60%), (c) pengunjung cenderung membutuhkan informasi mengenai Owlexa beserta layanannya, dan (d) kategori halaman yang tidak pernah diakses adalah Provider Kesehatan.
2. Bianda, dkk. (2016), dengan judul "*Web Usage Mining* untuk Mengetahui Pola Akses Pengguna Menggunakan Algoritma *Improved K-means* (Studi Kasus: kitabisa.com)". Adapun simpulan dari penelitian tersebut, yaitu
    a. nilai *k* sama dengan enam,
    b. *cluster* 1 dan *cluster* 2 mempunyai anggota paling banyak, yaitu 41% dan 29% dari seluruh matriks transisi,
    c. halaman yang paling sering dikunjungi adalah Home dan Halaman Proyek, dan
    d. halaman yang paling jarang dikunjungi adalah Liputan Media.

3. Halim (2011), berjudul "Analisis Perilaku Pengguna Digilib UNAIR dengan Menggunakan *Clustering Log Data Mining*". Simpulan dari penelitian tersebut adalah
    a. *log file* yang digunakan, yaitu satu tahun (Mei 2008–Mei 2019);
    b. *cluster* yang akan dibentuk berdasarkan alamat pengakses (k = 3), waktu pengakses (k = 3), dan lama akses (k = 3);
    c. sebanyak 92% pengguna Digilib UNAIR (adln.lib.unair.ac.id) melakukan akses melalui layanan ADLN di perpustakaan dengan lama akses lebih dari 60 menit; dan
    d. waktu kunjungan paling padat, yaitu dari pukul 10.00 WIB–15.00 WIB.

BAB V

PENUTUP

A. Simpulan

Berikut adalah simpulan dari penelitian yang telah dilakukan.

1. Berita merupakan rubrik yang paling sering dikunjungi.
2. Banyaknya *cluster* yang dibentuk berjumlah tiga (k = 3) dan iterasi dilakukan sebanyak tiga kali.
3. Perincian *cluster* yang terbentuk, yaitu
   a. *cluster* pertama, September 2017;
   b. *cluster* kedua, Oktober 2017, November 2017, Desember 2017, Februari 2018, Maret 2018, dan April 2018; dan
   c. *cluster* ketiga, Juli 2017, Agustus 2017, Januari 2018, Mei 2018, dan Juni 2018.

B. Saran

1. Ganto dapat menjadikan hasil penelitian ini sebagai bahan evaluasi untuk memperbaiki atau meningkatkan layanan situs web mereka agar penyediaan berita atau informasi lebih informatif dan edukatif serta menarik.
2. *Web usage mining* dapat diterapkan pada situs web selain Ganto, asalkan pihak atau pemilik situs web terkait berkenan untuk dilakukan penelitian.
3. Gunakan catatan *server log* dengan rentang waktu yang lebih lama, misalnya dua tahun atau lebih dari lima tahun–jika tersedia–agar hasil lebih akurat.

DAFTAR RUJUKAN

Hamimi, Hafilah. 2014. "Analisis Data Anggaran Pendapatan Belanja Daerah Menggunakan *Clustering* K-means dan *Forecasting* (Studi Kasus pada DPKA Kota Padang)". *Jurnal Vokasional Teknik Elektronika dan Informatika (VOTEKNIKA),* II(1).

Han, Jiawei., M. Kamber, & J. Pei. 2012. *Data Mining: Concepts and Techniques*. 3rd *ed*. Massachusetts: Morgan Kaufmann.

Hapsari, Septika. 2010. "Pembuatan *Website* pada *Google Original Movie Rental* Pacitan". *Journal Speed: Sentra Penelitian Engineering dan Edukasi,* II(2), halaman 48–54.

Larose, D. T. & C. D. Larose. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. 2nd *ed*. New Jersey: John Wiley & Sons, Inc.

Ong, J. O. 2013. "Implementasi Algoritma *K-means Clustering* untuk Menentukan Strategi *Marketing President University*". *Jurnal Ilmiah Teknik Industri*, Juni, XII(1), halaman 10–20.

Srivastava, *et al*. 2000. "*Web Usage Mining*: *Discovery and Applications of Usage Patterns from Web Data*". *SIGKDD Explorations,* Januari, I(2), halaman 12–23.

Syamnugroho, I. T. & Rissal Efendi. 2014. "Pengembangan *Website* SMK Negeri 4 Semarang". *Jurnal Teknologi Informasi dan Komunikasi,* Agustus, V(2).

_____. data-mining-blog.com/cloud-mining/rapidminer-cebit-2010, diakses 12 Agustus 2018.

_____. ganto.co/profil/1/tentang-kami.html, diakses 7 Mei 2018.

_____. ganto.co/profil/4/redaksi.html, diakses 7 Mei 2018.

_____. kdnuggets.com/2010/02/f-interview-rapid-i-founders.html, diakses 12 Agustus 2018.

_____. rapidminer.com/products/studio, diakses 9 Agustus 2018.

_____. techcrunch.com/2013/11/04/german-predictive-analytics-startup-rapid-i-rebrands-as-rapidminer-takes-5m-from-open-ocean-earlybird-to-tackle-the-u-s-market, diakses 12 Agustus 2018.