PEMODELAN PREDIKSI POLA AKSES WEBSITE PEMERINTAH MENGGUNAKAN *CLASSIFICATION-VIA-REGRESSION*

MODELING AND TREND OF ACCESS PATTERN OF GOVERNMENT WEBSITE USING CLASSIFICATION-VIA-REGRESSION

Eni Yusriani[1] dan Yoyon K. Suprapto[2]
[1]Pemerintah Kota Madiun,
Jl. Pahlawan No.37, Madiun, Telp: 0351-46994
[2] Fakultas Teknik Elektro, Institut Teknologi Sepuluh November (ITS), Surabaya
Jl. Arif Rahman Hakim, Surabaya, Jawa Timur, Telp: 031-5994251
E-mail: enzhuhadi@gmail.com[1], yoyonsuprapto@ee.its.ac.id[2]

Naskah diterima tanggal 24 Juni 2016, direvisi tanggal 5 September 2016, disetujui pada tanggal 7 September 2016

Abstract

Government websites is one means of service to the public. The pattern of visitor access on government websites could be an indicator for the quality of its service. In addition, the pattern can be used to see the trends and predictions for the public service needs of the government website in the future. In this study, the method used is the classification by regression (ClassificationViaRegression), because it needs the multiclass classification with numeric attributes, so that it can transform the problem of classification into a prediction based on M5 decision tree and the greedy pruning so that it can provide test results with 80.42% prediction accuracy. The results showed that the level of public access to government websites are still lower than the other contents, especially social media, streaming, even far lower than pornographic content. So the government is expected to develop and improve the quality of its official website following this trend, for example by providing a channel streaming and open accounts in social media for more interactive and intensive communication, so it can accommodate more of the aspirations and participation of the public in order to improve quality of public services.

Keywords *:ClassificationviaRegression, Government Website, Trend, Prediction, Streaming.*

Abstrak

Website pemerintah merupakan salah satu sarana pelayanan kepada masyarakat. Pola akses pengunjung pada website pemerintah bisa menjadi indikator atas mutu pelayanannya. Selain itu, pola tersebut bisa dijadikan dasar untuk melihat tren dan prediksi kebutuhan pelayanan untuk masyarakat terhadap website pemerintah di waktu yang akan datang. Metode yang digunakan adalah klasifikasi melalui regresi (*ClassificationViaRegression*), karena klasifikasi yang dibutuhkan bersifat multiclass, dengan atribut bernilai numerik, sehingga bisa mentransformasikan masalah klasifikasi menjadi fungsi prediksi yang berbasis pohon keputusan M5, dengan proses *greedy pruning* sehingga bisa memberikan hasil pengujian dengan akurasi prediksi 80,42%. Hasil penelitian menunjukkan bahwa tingkat akses masyarakat pada website Pemerintah masih di bawah website dan konten yang lain terutama Media Sosial, streaming, bahkan jauh di bawah konten pornografi. Sehingga Pemerintah diharapkan bisa mengembangkan dan meningkatkan kualitas website resminya mengikuti tren yang terjadi di masyarakat, misalnya dengan menyediakan channel streaming dan membuka akun di banyak media sosial untuk komunikasi yang lebih interaktif dan intensif, sehingga bisa menyerap lebih banyak aspirasi dan partisipasi dari masyarakat dalam rangka meningkatkan kualitas pelayanan publik.

Kata Kunci :*ClassificationviaRegression*, Website Pemerintah, Tren, Prediksi, Streaming.

PENDAHULUAN

Jumlah pengguna internet dunia pada bulan November 2015 mencapai 3,249 milyar orang dari 7,382 milyar penduduk dunia, atau sekitar 44%. Dari jumlah tersebut, 88,1 juta orang di antaranya adalah pengguna internet dari Indonesia (Smart Bisnis, 2016). Data sebelumnya yaitu pada tahun 2013, pengguna internet di Indonesia sebesar 63 juta (Kementerian Komunikasi dan Informatika RI, 2013), dan pada tahun 2014 mencapai 82 juta (Kementerian Komunikasi dan Informatika RI,

2014). Perkembangan dan peningkatan jumlah pengguna internet ini terjadi sangat cepat dan bersifat global, hampir menyentuh semua lapisan masyarakat, dari berbagai golongan umur dan profesi. Hal ini tentu saja menciptakan kompetisi yang sangat tinggi antar penyedia layanan online.

Di tengah kompetisi tersebut, pemerintah harus mengambil peran penting, sebagai penyedia utama layanan masyarakat.Terkait dengan perkembangan dunia informasi yang demikian cepat, salah satu sarana layanan yang diandalkan adalah website, di mana semua K/L/D/I harus mempunyai alamat website resmi.

Selama ini, website pemerintah secara umum cenderung "sepi pengunjung", karena banyak faktor. Misalnya tentang tampilan yang kurang menarik (monoton dari waktu ke waktu), konten yang kurang *uptodate*, akses web yang lambat, maupun kurang terpenuhinya ekspektasi pengunjung terhadap informasi yang dibutuhkan. Tentu saja tidak semua website pemerintah mencerminkan hal-hal seperti yang disebutkan, dan penilaian tersebut hanya diambil secara acak dari komentar atau respon pengunjung secara umum. Sedangkan untuk mendapatkan informasi tentang bagaimana sebenarnya akses masyarakat terhadap website pemerintah, bisa menggunakan beberapa cara. Misalnya dengan metode survei, menggunakan *page rank* pada halaman web, maupun dengan melihat log aksesnya.

Pada penelitian ini, metode yang digunakan adalah dengan melihat dan membandingkan akses internet pengguna pada website pemerintah dan website yang lainnya. Sumber data yang digunakan adalah data log server proxy di instansi pemerintah, yang dimungkinkan seharusnya mempunyai tingkat akses yang lebih tinggi pada website pemerintah. Data log yang ada akan diklasifikasikan menjadi beberapa kelompok, salah satunya adalah kelompok pemerintahan.

Dari masing-masing kelompok kemudian dibandingkan dan khusus untuk kelompok pemerintahan, dibuatkan suatu pemodelan untuk memprediksi pola akses dan faktor-faktor apa saja yang mempengaruhi akses pengunjung pada masa yang akan datang.

Rumusan permasalahan yang diselesaikan dalam penelitian ini adalah tentang bagaimana model akses pengunjung pada website pemerintah di waktu yang akan datang.

Adapun yang menjadi tujuan dari penelitian ini adalah mendapatkan prediksi pola akses masyarakat pada website pemerintah di masa yang akan datang sebagai sarana untuk meningkatkan mutu pelayanan pemerintah melalui website.

Kontribusi yang diharapkan bisa diberikan oleh penelitian ini adalah memberikan gambaran ke depan tentang prospek website pemerintah, yang harus mampu bersaing dengan website komersial dalam menarik minat masyarakat/pengunjung, sebagai salah satu indikator dari kepuasan masyarakat terhadap pelayanan pemerintah.

Penelitian Sebelumnya

Dalam penelitian-penelitian sebelumnya, untuk mengklasifikasikan akses website digunakan jenis aplikasi website sebagai dasar kategori, misalnya berdasarkan aplikasi HTTP, FTP, streaming, P2P, dan lain-lain. Sedangkan dalam penelitian ini dikhususkan pada aplikasi web berdasarkan alamat domain yang diakses atau alamat URL (pada log server). Dengan mengklasifikasi URL berdasarkan alamat domain diharapkan bisa mendapatkan alamat domain go.id dan konten-konten website yang mencerminkan bidang pemerintahan untuk dibandingkan dengan kelompok bidang yang lainnya.

Ada banyak penelitian yang menggunakan data log akses internet sebagai bahan/data sumber. Data yang ada bisa digali dari berbagai sisi, salah satunya adalah untuk klasifikasi dan analisa trafik. Analisa ini menjadi salah satu hal yang sangat penting untuk penyedia layanan internet, misalnya operator ISP. Dengan mendapatkan analisa trafik tersebut, akan memudahkan untuk melakukan monitoring dan evaluasi mutu pelayanan kepada konsumen. Selain itu, ada beberapa manfaat yang bisa didapatkan dari

hasil analisa trafik tersebut antara lain untuk mendeteksi ada tidaknya intrusi pada jaringan, menentukan alokasi *resource* jaringan, mendeteksi pola serangan dari luar, dan memberikan bahan referensi dari kebijakan jika terjadi kasus-kasus yang membutuhkan catatan trafik tertentu (Nguyen, 2008).

Pendekatan yang bisa dilakukan salah satunya adalah mengklasifikasikan trafik berdasarkan pengenalan pola statistik pada beberapa atribut yang diobservasi secara eksternal, dengan tujuan untuk mengklaster aliran trafik IP ke dalam kelompok-kelompok yang mempunyai kemiripan pola, atau mengklasifikasi satu atau lebih jenis aplikasinya. Jika klasifikasi dilakukan dengan menggunakan *Machine Learning* (ML) membutuhkan sejumlah langkah, antara lain menentukan fitur trafik, fitur maksimum dan minimum panjang paket atau fitur interval kedatangan paket. Setelah membuat kelas kemudian menggunakan salah satu metode dalam ML berdasarkan atribut-atribut yang sudah ditentukan.

Metode klasifikasi BLINC (*Blind Classification*) diperkenalkan oleh Thomas Karagiannis pada tahun 2005(Michalis Faloutsos, 2005). Metode ini dikatakan *blind* atau buta karena tidak ada informasi tentang paket *payload*, *port number* yang digunakan, dan informasi tambahan lain yang dibutuhkan, selain data yang disediakan oleh aplikasi kolektor trafik. Thomas mengatakan metodenya berbasis observasi dan identifikasi pola pada *transport layer* yang dibagi menjadi 3 level, yaitu sosial, fungsional dan level aplikasi. Sedangkan proses klasifikasinya dibagi dalam tiga kelompok web, P2P, *streaming*, *chat*, *game*, data ftp dan kelompok lain-lain.

Nguyen (Nguyen, 2008) menggunakan metode *Supervised Naive Bayes* untuk meneliti fitur panjang paket (min, max, *mean*, standar deviasi), statistik antar paket, statistik waktu kedatangan antar paket, dan kalkulasi melalui sejumlah kecil paket yang diklasifikasikan dan diambil dari bermacam-macam titik trafik yang signifikan, dimana ada tambahan trafik yang diteliti yaitu online game (*Enemy Territory*). Sedangkan algoritma *Expectation Maximization* digunakan oleh McGregor et.al (Lorier, McGregor, Hall, & Brunskill, 2004) dengan fokus pada fitur statistik paket data (min, max, kuartil), statistik interval kedatangan, *byte count*, durasi koneksi, jumlah transisi antar transaksi, dengan mengamati trafik campuran antara HTTP, SMTP, FTP, NTP, IMAP dan DNS.

Park et.al (CCJ, Tyan, & J, 2006) menggunakan algoritma *Naive Bayes* with *Kernel Estimation*, *Decision Tree* J48 dan *Reduce error Prunning Tree*, memfokuskan penelitian pada trafik WWW, Telnet, *Chat* (*Messenger*), FTP, P2P (Kazaa, Gnutella), Multimedia, SMTP, POP, IMAP, NDS, Oracle dan X11 untuk mendapatkan fitur durasi *flow*, jumlah aktual data paket, panjang paket, byte iklan, waktu interval antar paket dan total aliran paket.

Stefen Gebert pada tahun 2009(Schlosser, Gebert, & Heck, 2009) membuat pemodelan yang bisa digunakan untuk simulasi dan emulasi akses jaringan. Stefen mendapatkan tren akses aplikasi P2P dan file sharing yang sebelumnya mencapai 40% dari trafik, mulai mengalami penurunan, dan didominasi oleh aplikasi HTTP. Hal ini kemungkinan disebabkan oleh kebijakan tentang pemberian sanksi atas distribusi file-file video (biasanya film) tanpa ijin, sehingga pengguna beralih ke aplikasi web semacam Youtube atau RapidShare. Sehingga akses aplikasi HTTP menempati 60% dari trafik, sedangkan P2P hanya 14% saja.

Selain mengacu pada penelitian-penelitian di atas, penelitian ini juga merupakan pengembangan dari penelitian Penulis yang sudah pernah dilakukan sebelumnya (Yusriani, Suprapto, & Pratomo, 2014), di mana dalam penelitian tersebut, telah dilakukan klasifikasi trafik akses internet di instansi pemerintah, yang dikelompokkan ke dalam 8 kategori, berdasarkan *Top Level Domain* (TLD) dan kata kunci pada alamat URL, yaitu kelas Pemerintahan, Blog, Media Sosial, *Streaming*, Berita, Pendidikan, Email

dan kategori Lain-lain, dengan menggunakan metode Naïve Bayes. Hasil penelitian menunjukkan bahwa kelompok yang mempunyai peluang prediksi terbesar adalah Lain-lain dan *Streaming*. Kelompok Lain-lain mempunyai peluang terbesar karena kategori ini mempunyai peluang kata kunci yang paling besar di antara yang lain. Sedangkan *Streaming* lebih banyak didominasi oleh akses pada web Youtube.

Pengolaan Data Log Server

Data log server digunakan sebagai data sumber penelitian yang berisi data akses user pada halaman-halaman website. Untuk mengolah data log, menggunakan tahap-tahap web *data mining*, sebelum data tersebut bisa dimanfaatkan untuk membuat pemodelan. Salah satu yang menjelaskan pengolahan web *data mining* adalah Liu Bing, dengan tahapan sebagai berikut(Bing, 2006):

- *Data Cleaning*
  Pada tahap ini, item-item data log yang tidak dibutuhkan, bisa dihilangkan.
- *User Identification*
  Tahap untuk melihat dan menentukan IP user pengakses pada tiap-tiap halaman web yang dikunjungi.
- *Session Identification*
  Merupakan kumpulan *page* yang dikunjungi oleh *user* yang sama
- *Path Completion*
  Path page perlu disimpan untuk mengantisipasi jika ada *page* yang hilang setelah proses transaksi.

Jika semua proses penyiapan data sumber tersebut sudah dilakukan, maka dataset siap untuk digunakan sebagai bahan penelitian.

Analisis Regresi

Analisis regresi merupakan suatu proses statistik yang mengestimasi hubungan antara beberapa variabel, termasuk di dalamnya beberapa teknik pemodelan dan analisis beberapa variabel, antara satu atau beberapa variabel bebas dan suatu variabel terikat/independen. Biasanya digunakan untuk prediksi atau peramalan. Analisis regresi mengestimasi nilai variabel dependen yang diharapkan berdasarkan variabel independen yang diberikan, yaitu nilai rata-rata dari variabel dependen jika variabel independen ditetapkan.

Analisis regresi digunakan untuk prediksi dan peramalan, di mana penggunaannya secara substansi sering *overlap* dengan *machine learning*. Analisa regresi juga digunakan untuk mengetahui di antara variabel independen mana saja yang mempengaruhi variabel dependen, dan menemukan formulasi yang menghubungkan keduanya. Analisa regresi juga digunakan untuk menganalisa kausalitas antar variabel. Banyak teknik yang digunakan untuk menggambarkan analisis regresi. Salah satu metode yang cukup familiar adalah regresi linier yang merupakan regresi parametrik, di mana dalam fungsi regresi tersebut dipengaruhi sejumlah parameter tertentu.

Metode Klasifikasi Melalui Regresi

Metode klasifikasi melalui regresi (*classification via regression*) merupakan metode klasifikasi yang bisa mentransformasikan permasalahan menjadi fungsi regresi, menggabungkan prinsip algoritma pohon keputusan M5 dan regresi linier pada beberapa sub pohon (daun) yang dibangun. Metode ini mempunyai dua langkah utama (Wang & Witten, 1996) yaitu :

1. Membuat pohon keputusan biasa, dengan memaksimalkan pemisahan kriteria/parameter/ atribut dan variasinya sesuai dengan nilai target/output, serta kelas dibinerkan.
   Dalam membuat pohon keputusan, yang perlu diperhatikan adalah menghitung Standar Deviasi Reduksi (SDR) dengan persamaan:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \quad \text{(Pers. 1)}$$

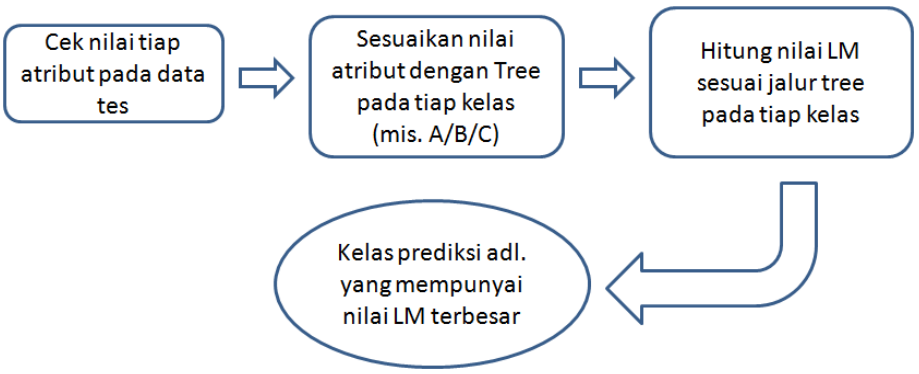
   dimana: sd(T) adalah standar deviasi dari T, T adalah dataset yang mencapai node pohon, dan T1, T2,... adalah himpunan

hasil dari pemisahan node berdasarkan atribut yang dipilih.

2. Memangkas pohon keputusan (*pruning*) tersebut pada beberapa sub pohon yang memungkinkan, dan mengisinya dengan fungsi regresi (Linear Model/LM) yang sesuai, biasanya pada daun/*leaf*.
   Dalam proses *pruning* hitung faktor pengali (pf) dengan formula sebagai berikut:

$$pf = \frac{n+v}{n-v} \quad \text{(Pers. 2)}$$

Adapun output dari metode ini adalah *tree* untuk tiap kelas, di mana di dalam *tree* tersebut terdapat model regresi linier (LM) yang disesuaikan dengan nilai dan variasi atribut. Jika terdapat banyak atribut, maka model linier yang digunakan adalah *multiple linear regression*. Alur prediksi kelas pada data tes ditunjukkan pada diagram Gambar 2.



Gambar 2. Blok diagram alur prediksi dengan regresi

Metode Penelitian

Penelitian ini dilakukan dengan beberapa tahapan, yaitu:

a. Klasifikasi Dataset

Untuk membuat klasifikasi dataset, dibuat aplikasi bantuan sebagai tahap pertama dalam bentuk visualisasi yang *user friendly*. Dengan bantuan aplikasi tersebut, diharapkan bisa memberikan sumber data yang siap pakai untuk proses analisa selanjutnya.

Tahap klasifikasi website adalah tahap memfilter data website yang diakses, untuk dimasukkan ke dalam beberapa kategori sebagai perbandingan, dimana untuk aplikasi website dikelompokkan dalam kategori Pemerintahan, Pendidikan, Email, Media Sosial, Blog/Belanja Online, *Streaming*, Berita, Pornografi dan Lain-lain. Alur kategorisasi URL ini bisa dilihat pada Gambar 3.

b. Penentuan Parameter

Tahap selanjutnya adalah penentuan parameter yang digunakan untuk mengklasifikasikan website. Parameter yang digunakan adalah:

- jumlah*user*; jumlah *user* yang mengakses suatu website dihitung totalnya dalam jangka waktu tertentu, misalnya untuk waktu seminggu, dua minggu, atau sebulan
- waktu akses; dalam hal ini yang dipergunakan adalah waktu/jam berapa saja website tersebut diakses oleh *user*, karena terkait jam kerja
- tanggal akses;
- intensitas akses; berapa kali website tersebut dikunjungi dalam jangka waktu tertentu
- *resource bandwidth*; jumlah *resource* data yang dibutuhkan untuk mengakses website tersebut.

c. Pemodelan Akses Website Menggunakan *ClassificationViaRegression*

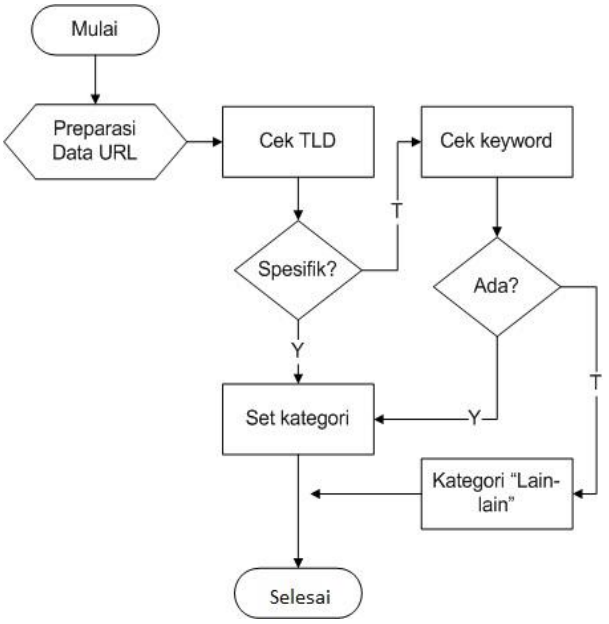
Untuk mengklasifikasikan URL dengan menggunakan regresi (*CvR*), ada empat tahap utama yang dilakukan, yaitu:

1. Pembangunan pohon keputusan biasa, dengan atribut tanggal, jam, intensitas, *user* dan *resource*.
   Dengan langkah-langkah:
   a. Hitung nilai deviasi masing-masing
   b. Hitung nilai SDR dengan menggunakan Pers. 1.
   c. Pilih SDR terbesar sebagai pemisah parameter/*node*

2. Pemangkasan pohon (*greedy pruning*)
   Langkah-langkah *pruning*:
   a. Hitung rata-rata selisih absolut antara nilai prediksi kelas/kategori dan nilai aktual semua data training yang mencapai *node* tertentu
   b. Hitung faktor pengali pf dengan Pers. 2.
   c. Buat model linier standar pada setiap *node* interior
   d. *Drop* salah satu parameter
   e. Hitung *error* dengan menggunakan pf
   f. Cek nilai *error* yang didapatkan dengan yang diharapkan, jika belum menurun ulangi lagi langkah (d).
   g. Tempatkan model linier pada *node* yang ditentukan
   h. Ulangi mulai langkah (a) lagi sampai tercapai *error* minimal dan semua *node* interior berisi model linier.

3. Penghalusan nilai prediksi (*smoothing*)
   a. Hitung nilai prediksi dari model linier standar pada suatu *node*
   b. Kirimkan nilai tersebut pada *node* di atasnya untuk dikombinasikan dengan nilai prediksi *node* tersebut
   c. Ulangi langkah (b) hingga mencapai root dari pohon.

4. Menentukan prediksi kelas, dengan alur sebagaimana Gambar 3.



Gambar 3. Diagram Alir Kategorisasi Website/URL

HASIL DAN PEMBAHASAN

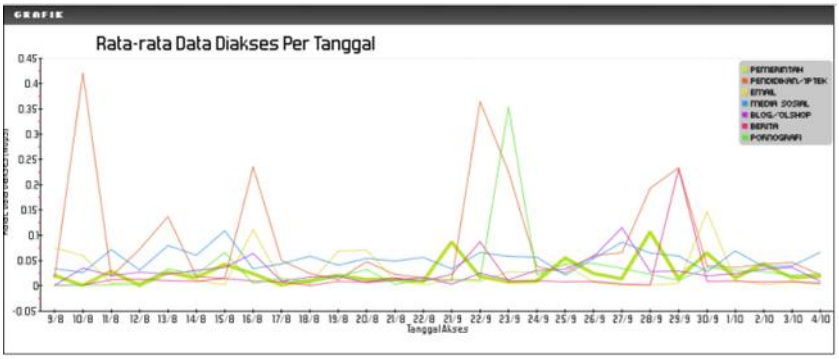
Sesuai dengan metodologi yang disampaikan pada bagian sebelumnya, berikut ini adalah pembahasan hasil berdasarkan tahap penelitian:

1. Hasil kategorisasi website berdasarkan URL
   Dalam proses kategorisasi, kategori website ditentukan berdasarkan *Top Level Domain* (TLD) terlebih dulu. Jika mempunyai TLD spesifik maka langsung dikategorikan seperti sesuai dengan isian pada tabel domain yang ditentukan sebelumnya. Sedangkan jika TLD website bersifat *General*, maka baru kemudian dicek pada *keyword* yang ada pada alamat URL, disesuaikan dengan *keyword* yang terekam dalam database.
2. Pola akses yang dihasilkan
   Pola trafik dihitung berdasarkan waktu per menit untuk mendapatkan tingkat akurasi yang mendekati aktual.
3. Analisis tren dengan regresi
   Analisis tren dicari dengan menghitung prediksi menggunakan regresi linier, regresi

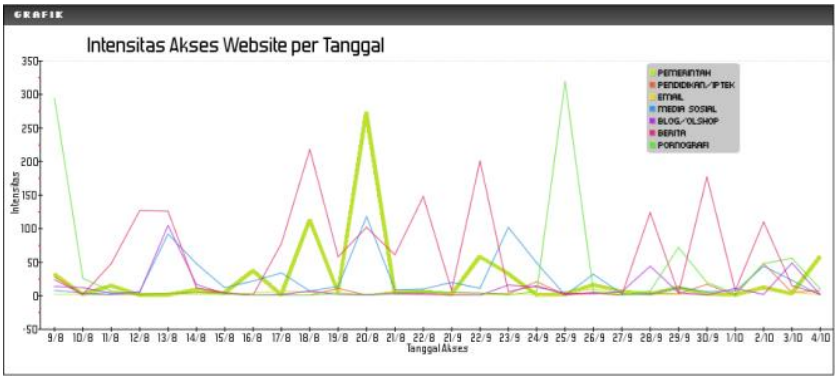
kuadratik dan eksponensial. Dari ketiga model regresi yang didapatkan, dihitung MSE masing-masing. Kemudian dibandingkan, formula prediksi yang mempunyai nilai MSE terkecil adalah yang mendekati kondisi aktual.

Pola Akses Internet Saat Ini

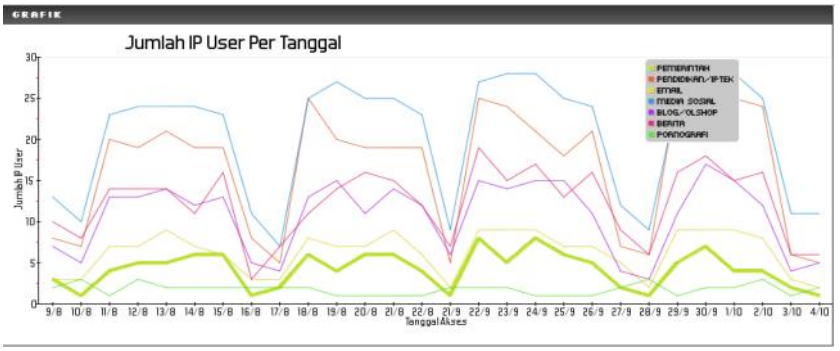
Pola pemanfaatan internet ditinjau dari sisi konsumsi data, jumlah *user* dan intensitas akses. Selama waktu pengamatan, diperoleh pola *resource* data dari masing-masing kategori seperti pada Gambar 4. Sedangkan pola pemanfaatan internet berdasarkan jumlah pengguna dan intensitas ditunjukkan pada Gambar 5 dan Gambar 6.



Gambar 4. Pola pemakaian *resource* data per kategori



Gambar 5. Pola intensitas akses per kategori



Gambar 6. Pola jumlah *user*/IP per kategori

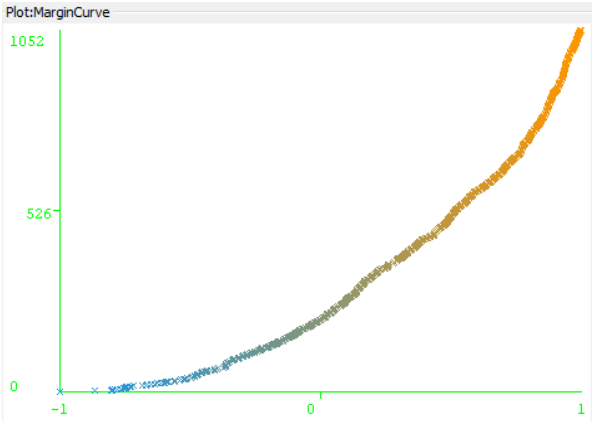
Berdasarkan pola yang didapatkan dan perhitungan rata-rata statistik, *user* lebih banyak mengakses website yang termasuk pada kategori Media Sosial, dengan rata-rata jumlah

pengguna sebesar 20,3929 orang/hari, rata-rata intensitas akses 747,143 kali/hari, dan rata-rata konsumsi *resource* data 48,92 Kbps/hari. Sedangkan untuk kategori Pemerintahan, rata-rata pengakses adalah 4,21 orang/hari, rata-rata intensitas akses 70,21 kali/hari dan rata-rata pamakaian *resource* data sebesar 24,19 Kbps/hari.

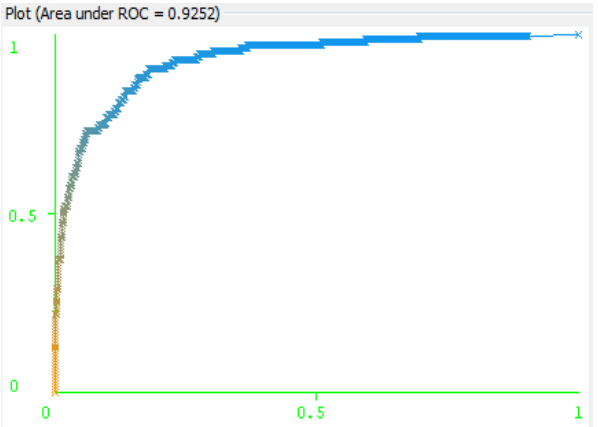
Prediksi Akses Internet Menggunakan Regresi

1. Analisis Regresi Kategori Pemerintahan Berdasarkan Waktu
   Pola yang didapatkan pada kondisi saat ini, bisa dijadikan dasar prediksi pola yang akan datang berdasarkan waktu menggunakan regresi. Setiap parameter dicari formula prediksinya menggunakan regresi linier, regresi kuadratik dan eksponensial.
   Parameter yang diprediksi dalam penelitian ini adalah konsumsi *resource* data, jumlah *user* dan intensitas akses. Setelah dihitung MSE pada tiap model regresi yang dihasilkan, didapatkan bahwa tren yang mendekati adalah tren kuadratik, dengan formula untuk parameter:
   - Kebutuhan *resource* data (*bytes*):
     y = 28, 2593 + 0, 2931x − 0, 0063x2
   - Jumlah *user* pengakses:
     y = 5, 15 + 0, 004x + 0, 004x2
   - Intensitas akses:
     y = 93, 199 + 0, 46x + −0, 086x2
2. Analisis Prediksi dengan *CvR*
   Data yang dianalisis adalah data trafik per menit, dengan atribut parameter yang digunakan adalah tanggal, jam akses, jumlah *user*, intensitas akses dan *resource* pemakaian data. Dari pengujian tanpa filter sebanyak 11 kali, diperoleh akurasi tertinggi sebesar 73% menggunakan data split dengan prosentase 80%. Kemudian pengujian dilakukan lagi dengan menambahkan filter *Resample*100% dan bisa mengikuti default = 0 pada praproses. Dari hasil pengujian ini, diperoleh peningkatan akurasi menjadi 78,13%. Sedangkan jika bias diset 1, dengan filter *Resample* yang sama (100%) , maka diperoleh akurasi sebesar 80,42%.

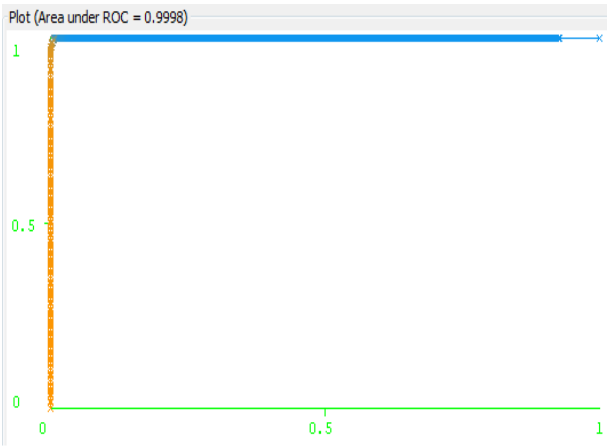
Kurva margin *error* yang dihasilkan adalah seperti pada Gambar 7. Untuk nilai ROC rata-rata diperoleh 96,9%. ROC terendah adalah kategori Pendidikan/Iptek sebesar 92,52% seperti diperlihatkan pada Gambar 8. Sedangkan ROC tertinggi adalah 100% untuk kategori Streaming seperti ditunjukkan pada Gambar 9. Sebagai fokus utama pengamatan penelitian ini, kategori Pemerintahan mempunyai nilai ROC yang cukup tinggi, yaitu 98,72% (Gambar 10).



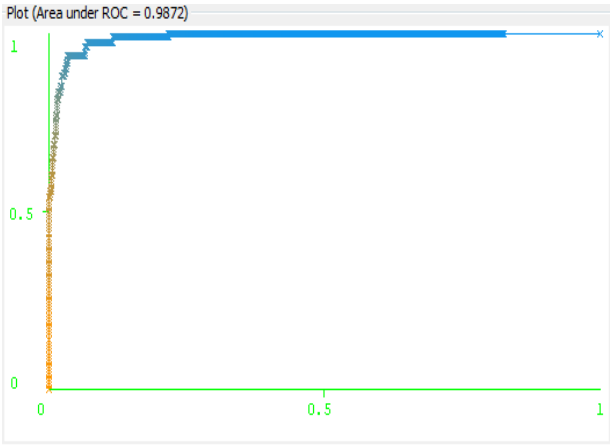
Gambar 7. Margin *Error* pada Pengujian dengan Akurasi Tertinggi



Gambar 8. ROC Kategori Pendidikan/Iptek



Gambar 9. ROC Kategori Streaming



Gambar 10. ROC Kategori Pemerintahan

Recall tertinggi kategori *Streaming* sebesar 99,4%, terendah pada kategori Berita sebesar 60,4%. Sedangkan *precision* tertinggi adalah kategori *Streaming* 96,9%. Sedangkan model hasil prediksi yang paling dominan digunakan dalam tahap pengujian data adalah seperti pada Tabel 1.

Tabel 1. Model linier yang paling dominan dalam pengujian data

Index Kelas	Jml Model	Model yang sering dipakai
0 (Berita)	164	LM160 (647 kali)
kategori = 0*x1 + 0,0002*x2 + 0*x3 - 0,0001*x4 - 0,0001*x5 +0,0014		
1 (Blog/Olshop)	147	LM147 (633 kali)
kategori = 0*x1 + 0*x3 - 0,0001*x4 - 0*x5 + 0,0045		
2 (Email)	116	LM115 (468 kali)
		LM116 (1249 kali)
kategori = -0,0002*x1 - 0,0005*x2 + 0,0001*x3 - 0,0004*x4 - 0,0003*x5 +0,0165		
kategori = -0,0001*x1 - 0*x2 - 0*x4 - 0*x5 +0,0033		
3 (Medsos)	88	LM20 (524 kali)
kategori = 0,0004*x2 + 0,0001*x3 - 0,0003*x4 - 0*x5 +0,002		
4 (Pemerintahan)	131	LM77 (113 kali)
		LM131 (1064 kali)
kategori = 0,001*x1 - 0,0004*x2 - 0,0245*x3 + 0*x4 + 0,0065*x5 + 0,1137		
kategori = 0*x1 - 0,0002*x2 + 0*x3 + 0*x4 - 0*x5 +0,0043		
5 (Pendidikan)	154	LM144 (251 kali)
kategori = 0,0005*x1 - 0,0004*x2 - 0,0004*x3 + 0,0022*x4 - 0,0002*x5 + 0,003		
6 (Pornografi)	123	LM1 (159 kali)
kategori = 0,0025*x1 + 0,0041*x2 - 0,0079*x3 - 0,0004*x5 - 0,0001*x5 + 0,022		
7 (Streaming)	16	LM1 (3192 kali)
		LM14 (464 kali)
kategori = 0*x1 + 0*x2 - 0*x3 + 0*x4 + 0*x5 +0		
kategori = 0,0001*x1 + 0,0003*x2 - 0,0001*x3 - 0,0004*x4 + 0,0001*x5 + 0,977		

dimana: *x1* adalah tanggal akses (ctanggal), *x2* adalah jam akses (jam), *x3* adalah intensitas akses, *x4* adalah jumlah pengakses (*user*) dan *x5* adalah jumlah konsumsi *resource* data (*resource/bytes*).

Pada model yang dominan di kategori Berita, parameter tanggal, jam, dan intensitas mempunyai koefisien positif yang menandakan bahwa parameter tersebut memberikan pengaruh positif pada prediksi kategori. Parameter jumlah pengakses dan jumlah data akses (*resource*) kurang memberikan pengaruh yang signifikan, dan bisa jadi berbanding terbalik dengan hasil prediksi karena adanya faktor *intercept* bernilai positif. Hal ini juga ditunjukkan oleh *rule model* pada kategori Blog/Olshop. Sedangkan pada kategori Email, semua atribut menunjukkan koefisien negatif, (bahkan atribut intensitas tidak ada dalam model yang dominan) kecuali nilai *intercept*-nya. Sehingga prediksi untuk kategori ini tidak dipengaruhi oleh atribut yang digunakan.

Untuk kategori Media Sosial, atribut tanggal tidak ada pada model dominan yang dipilih, koefisien jam dan intensitas menunjukkan nilai positif, sedangkan jumlah *user* dan jumlah data akses bernilai negatif. Sehingga masing-masing memberikan pengaruh yang seimbang dalam penentuan prediksi ditambah dengan nilai positif dari *intercept*. Kategori Pemerintahan menunjukkan representasi model yang sebaliknya, di mana jumlah pengakses, tanggal akses dan jumlah data akses mempunyai nilai koefisien positif, sehingga memberikan peluang berbanding lurus dengan nilai prediksi. Atribut jam akses dan intensitas mempunyai kontribusi negatif.

Perbandingan Hasil Prediksi dengan Metode Naïve Bayes

Pada kajian pustaka yang dilakukan pada paper (Nguyen, 2008), proses klasifikasi yang sudah dilakukan lebih banyak menggunakan metode *Naive Bayes*. Selain akurasi, yang dijadikan acuan evaluasi adalah *Trust* atau *Recall*, khususnya pada klasifikasi *multiclass*. Pada proses klasifikasi *multiclass*, biasanya tingkat akurasi semakin menurun sesuai dengan jumlah klas yang dicari. Sehingga dengan menggunakan acuan *Recall/Trust*, maka yang dilihat adalah *recall* pada masing-masing kelasnya.

Dalam penelitian ini, pengujian dengan *Naive Bayes* menggunakan filter *KernelEstimator* dimaksudkan untuk meningkatkan akurasinya. Berdasarkan hasil *Recall* (R), *Precision* (P) dan ROC tiap kelas yang dihasilkan oleh kedua metode, didapatkan perbandingan sebagaimana ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan Per Kelas Antara Metode CvR dan NB

Kategori	Metode CvR			Metode NB		
	Akurasi : 80,42%			Akurasi : 43,58%		
	%R	%P	%ROC	%R	%P	%ROC
Berita	60,5	76,4	93,6	20,9	39,3	76,1
Blog/Olshop	70,7	69	95,9	17,9	31,6	73,2
Email	76	74,8	95,7	30,6	22,7	75,5
Media Sosial	88,7	88,1	99,2	55,9	52,1	87,1
Pemerintahan	81,5	83,8	98,7	32,1	47,1	79,7
Pendidikan/Iptek	68,8	68,8	92,5	7,4	34,3	67,9
Pornografi	92,5	80,4	99,1	92,8	36,5	90,4
Streaming	99,4	96,9	100	84,4	89	97,2

Kategori yang mempunyai ROC tertinggi pada kedua metode adalah sama, yaitu kategori *Streaming*. Demikian juga untuk ROC terendah, adalah pada kategori Pendidikan/Iptek. Kategori Pemerintahan pada metode regresi mempunyai akurasi lebih tinggi yaitu 98,7% sedangkan pada metode *Naive Bayes* adalah 79,7%.

Dari Tabel 2 terlihat kedua metode mempunyai persamaan pada tiga kategori dengan tingkat ROC tertinggi, yaitu Streaming, Pornografi, dan Media Sosial. Hal ini menunjukkan bahwa ketiga kategori tersebut mempunyai jumlah user, intensitas akses user dan tingkat konsumsi data yang paling tinggi,dan hal ini sesuai dengan hasil survey (Smart Bisnis, 2016) bahwa dari 88 juta pengguna internet di Indonesia, 79 juta orang di antaranya adalah pengguna aktif media sosial. Pemerintah harus bisa memanfaatkan kondisi itu untuk menunjang peningkatan pelayanan masyarakat.

PENUTUP

Kesimpulan yang diperoleh dari penelitian ini adalah bahwa untuk kondisi sekarang, akses terhadap website pemerintah masih lebih rendah jika dibandingkan dengan kategori website yang lain, bahkan di instansi pemerintah itu sendiri, koneksi internet lebih banyak digunakan untuk mengakses media sosial daripada pemerintahan. Yang paling banyak diakses adalah kategori *Streaming, Pornografi dan Media Sosial*. Akan tetapi dalam hal ini perlu penelitian lebih lanjut tentang konten yang lebih spesifik dari streaming dan media sosial yang diakses. Karena di dalamnya dimungkinkan masih ada akses konten negatif/pornografi. Sehingga bisa didapat hasil yang lebih akurat pada masing – masing kategori tersebut.

Sedangkan model prediksi kategori Pemerintahan yang dihasilkan memperlihatkan bahwa parameter waktu akses (tanggal dan jam) dan *resource* data lebih banyak berkorelasi positif terhadap hasil prediksi, dan berkebalikan dengan parameter jumlah pengguna dan intensitas akses, berarti akses website pemerintah lebih banyak dipengaruhi oleh waktu, misalnya di saat jam kerja, sedangkan jumlah pengunjungnya tidak terlalu tinggi.

Sebagai salah satu bentuk tindak lanjut dari penelitian ini adalah menjadikan output penelitian sebagai salah satu input dalam manajemen website Pemerintah. Pada kategori Streaming dan Sosial Media, menunjukkan bahwa saat ini masyarakat lebih tertarik untuk mengakses informasi melalui bentuk visual, interaktif dan *mobile* yang lebih mudah dan cepat. Sehingga sangat disarankan untuk mengembangkan website Pemerintahan yang dinamis, didukung oleh ketersediaan *chanel streaming* dan media sosial, mengikuti tuntutan dan kebutuhan masyarakat khususnya dalam hal diseminasi informasi. Misalnya dengan membuat *channel streaming* di Youtube, akun Facebook, Twitter, Instagram, dan media sosial yang lain yang lebih cepat dan mudah diakses masyarakat. Sehingga diharapkan bisa lebih menyerap dan menyampaikan misi pelayanan yang diemban.

DAFTAR PUSTAKA

Bing, L. (2006). *Web Data Mining*. New York: Springer.

CCJ, K., Tyan, H., & J, P. (2006). Internet Traffic Classification for Scalable QoS Provision. *IEEE International.*

Kementerian Komunikasi dan Informatika RI. (2014, May 8). *Kemkominfo: Pengguna Internet di Indonesia Capai 82 Juta*. Dipetik Agustus 30, 2016, dari Website Kemkominfo: https://www.kominfo.go.id/content/detail/3980/kemkominfo-pengguna-internet-di-indonesia-capai-82-juta/0/berita_satker

Kementerian Komunikasi dan Informatika RI. (2013, November 7). *Kominfo : Pengguna Internet di Indonesia 63 Juta Orang*. Dipetik Agustus 30, 2016, dari Website Kemkominfo: https://www.kominfo.go.id/content/detail/3415/kominfo-pengguna-internet-di-indonesia-63-juta-orang/0/berita_satker

Lorier, P., McGregor, A., Hall, M., & Brunskill, J. (2004). Flow clustering using machine learning techniques. *Development and Society* .

Michalis Faloutsos, T. K. (2005). BLINC: Multilevel Traffic Classification in the Dark. *SIGCOMM'05.*

Nguyen, G. A. (2008). A survey of techniques for Internet Traffic Classification Using Machine Learning. *IEEE Communication and Survey and Tutorial, 10(4).*

Schlosser, D., Gebert, S., & Heck, R. P. (2009). Internet Access Traffic Measurement and Analysis. *University of Wurzburg.*

Smart Bisnis. (2016, February 19). *Statistik Pengguna Internet dan Media Sosial di Indonesia*. Dipetik Agustus 30, 2016, dari Website Smart Bisnis: http://www.smartbisnis.co.id/content/read/belajar-bisnis/statistik-pengguna-internet-dan-media-sosial-di-indonesia

Wang, Y., & Witten, I. H. (1996). Induction of Model Tree for Predicting Continous Classes. *Department of Computer Science The University of Waikato* .

Yusriani, E., Suprapto, Y. K., & Pratomo, I. (2014). Analisis Trafik Pemanfaatan Internet di Instansi Pemerintah. *Prosiding Temu Alumni Penerima Beasiswa Kementerian Komunikasi dan Informatika ke-5* , 456.