Algoritma K-Means untuk Diskretisasi Numerik Kontinyu Pada Klasifikasi Intrusion Detection System Menggunakan Naive Bayes

David Ahmad Effendy[1)], Kusrini[2)], Sudarmawan[3)]
Universitas AMIKOM Yogyakarta
Jl. Ringroad Utara Condong Catur Depok Sleman
Yogyakarta 55283 Indonesia
e-mail: bangdavid07@gmail.com[1)], kusrini@amikom.ac.id [2)], sudarmawan@amikom.ac.id[3)]

Abstrak

Intrusion Detection System (IDS) merupakan sebuah perangkat lunak atau perangkat keras yang dapat digunakan untuk mendeteksi adanya aktivitas yang tidak wajar dalam jaringan. Ada beberapa cara IDS bekerja. Cara umum adalah pendeteksian berbasis signature, mencocokannya dengan pola perilaku serangan yang telah didefinisikan dalam database. Kelemahan teknik ini tidak mampu mendeteksi jenis serangan yang dapat memodifikasi dirinya sendiri. Metode selanjutnya adalah Anomaly-based IDS, dilakukan dengan membandingkan lalu lintas yang sedang dipantau dengan lalu lintas normal. Kelemahannya adalah cara ini banyak memberikan pesan false positive. IDS membutuhkan performansi yang relatif cepat dengan tingkat false positif yang rendah. Penerapan metode machine learning sangat cocok untuk masalah ini, contohnya naive bayes. Naive bayes memerlukan atribut dengan nilai diskrit sehingga diperlukan proses diskritisasi untuk merubah atribut numerik kontinyu kedalam bentuk diskrit. Untuk menangani atribut numerik kontinyu digunakan algoritma K-Means Clustering. Hasil pengujian menunjukkan penerapan naive bayes dengan proses diskretisasi menghasilkan nilai akurasi 95,6%.

Kata kunci: *IDS, Diskretisasi, K-Means, Naive Bayes*

1. Pendahuluan

Perkembangan teknologi jaringan komputer menjadikan sistem keamanan menjadi sangat penting. Diperlukan sebuah sistem yang mampu mendeteksi, dan mengidentifikasi adanya aktifitas yang tidak wajar. Munculnya teknologi *Intrusion Detection System (*IDS*)* dapat memberikan solusi keamanan, sistem ini dapat diterapkan pada *host based IDS (*HIDS*)* dan *network based IDS (*NIDS*)* kedua teknik ini memiliki kemiripan namun hanya pada perangkat lunak atau perangkat keras tempat dari sistem IDS ini terpasang [1].

Secara umum cara kerja IDS dikembangkan dengan 2 cara yaitu menggunakan pendeteksian berbasis signature, yaitu mencocokkan dengan pola perilaku serangan yang telah didefinisikan dalam database. Teknik ini membutuhkan waktu eksekusi yang relatif singkat untuk proses pencocokan pola, namun memiliki kelemahan yaitu tidak mampu mendeteksi jenis serangan yang dapat memodifikasi dirinya sendiri. Cara selanjutnya adalah *Anomaly-based IDS,* teknik ini akan mendeteksi adanya aktifitas yang tidak wajar dari aktifitas jaringan pada kondisi biasanya. Teknik ini dapat mendeteksi adanya serangan dengan tipe baru namun definisi dari keadaan normal harus dapat dipastikan terlebih dahulu. Kelemahannya, adalah jenis ini banyak memberikan pesan *false positive* [2].

Intrusion Detection System (IDS) merupakan sebuah sistem perangkat lunak atau perangkat keras yang dapat digunakan untuk mendeteksi adanya aktivitas yang tidak wajar dalam sistem atau jaringan computer [1]. Permasalahan IDS ini telah didekati dengan beberapa algoritma kecerdasaran buatan seperti *decision tree*, *naive bayes*, *support vector machine*, jaringan syaraf tiruan dan algoritma lainnya. Penggunaan teknik kecerdasasn buatan yang dikenal sebagai data mining ataupun *machine learning* sebagai alternatif kemampuan manusia yang mahal dan berat. Teknik ini secara otomatis mempelajari data atau mengekstrak pola yang bermanfaat dari data sebagai referensi profil tingkah laku normal atau serangan dari data yang ada untuk klasifikasi trafik jaringan selanjutnya [3].

Penelitian ini bertujuan untuk menerapkan dan menganalisa *naive bayes* dengan dataset yang memiliki 4 kelompok jenis serangan yaitu DoS, Probe, R2L, dan U2R. Pengujian yang digunakan adalah mendeteksi besarnya akurasi klasifikasi serangan. Pemilihan fitur dari dataset yang akan digunakan adalah teknik *Information Gain* (IG) dan *Corellation Feature Selection* (CFS). Algoritma *naive bayes*

didasarkan pada tingkat probabilitas nilai dari suatu atribut terhadap kelasnya, teknik ini akan menghasilkan nilai probabilitas yang sangat kecil jika terdapat sangat banyak nilai yang berbeda dalam suatu atribut. Hal ini dapat disebabkan oleh penggunaan atribut dengan tipe kontinyu, sehingga nilai probabilitas menunjuk kemungkinan nilai yang sama keluar pada suatu kelas namun pada sisi lain rentang nilai pada atribut tersebut sangat besar sehingga nilai probabilitas dari nilai tersebut muncul kembali dalam suatu kelas akan sangat kecil. Hal inilah yang akan menyebabkan lemahnya performansi dari *naive bayes*. Untuk mengelompokkan atau mendeksritisasi data dengan tipe kontinyu dapat dilakukan dengan menggunakan K-Mean Clstering [4][5][6]. Pada penelitian ini K-Means Clustering akan digunakan untuk meningkatkan kinerja algoritma *naive bayes*.

2. Metode Penelitian

Penelitian ini adalah penelitian eksperimen. penelitian ini menyajikan penerapan *naïve bayes* dengan diskretisasi variabel bernilai numerik kontinyu menggunakan algoritma *K-Means* sekaligus dengan menggunakan seleksi atribut. Dalam mengklasifikasi intrusi jaringan komputer performa diukur berdasarkan tingkat *accuracy*. Dataset yang digunakan pada penelitian ini adalah NSL-KDD'99 yang diperoleh di *http://nsl.cs.unb.ca/NSL-KDD/* yang dikelompokkan menjadi 4 kategori serangan yaitu DoS, R2L, U2R, dan Probe.

2.1. NSL-KDD'99 *Dataset*

Data yang digunakan pada penelitian ini adalah dataset NSL-KDD Cup 1999. NSL-KDD adalah sebagai solusi dari permasalahan yang ada pada *dataset* KDD Cup 1999 (KDD-99). *Dataset* KDD-99 berusianya sudah lebih dari 15 tahun, namun masih umum digunakan dalam penelitian-penelitian sistem deteksi intrusi karena tidak banyak *dataset* alternatif yang tersedia dan dapat diakses publik [7].

Dataset ini terdiri dari kelas normal dan 39 jenis serangan. Pada penelitian ini jenis serangan yang terdapat pada dataset dikelompokkan menjadi 4 kategori yaitu DoS, R2L, U2R, dan Probe. Seperti yang terdapat pada Tabel 1. dan tipe atribut dataset NSL-KDD terdapat pada Tabel 2. dibawah ini.

Tabel 1. Kategori Serangan pada IDS

Normal (1825)	**Dos (3631)**	**Probe (3631)**	**R2L (2436)**	**U2R (56)**
Normal (1825)	Back (297) Land (4) Pod (35) Smurf (528) Teardrop (9) apache2 (615) Udpstorm (2) Processtable (572) Mailbomb (245) Neptune (1324)	Satan (614) Ipsweep (121) Nmap (58) Portsweep (133) Mscan (869) Saint (257)	guess_passwd (1031) ftp_write (2) imap (1) phf (2) multihop (13) warezmaster (506) xlock (8) xsnoop (3) snmpguess (282) snmpgetattack (152) httptunnel (110) sendmail (10) named (16)	buffer_overflow (16) loadmodule (2) rootkit (11) perl (2) sqlattack (2) xterm (10) ps (13)

Tabel 2. Tipe Atribut

Nominal	Biner	Numerik
protocol_type(2) service(3) flag(4)	land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21), is_guest_login(22)	duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23), srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29), diff_srv_rate(30), srv_diff_host_rate(31),

2.2. Naive Bayes Classifier

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas [8]. *Naive Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu [9]. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya

membutuhkan jumlah data yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan. Persamaan metode naive bayes di definisikan dengan persamaan (1)[10].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

dimana :

X : data dengan *class* yang belum diketahui
H : hipotesis data merupakan suatu *class* spesifik
P(H|X) : Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probabilitas)
P(H) : probabilitas hipotesis *H* (prior probabilitas)
P(X|H) : probabilitas *X* berdasarkan kondisi pada hipotesis *H*
P(X) : probabilitas X

2.3. Normalisasi

Normalisasi disini merupakan proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tertentu. Normalisasi dilakukan bertujuan untuk mengurangi adanya kesalahan pada proses data mining. Metode Siqmoid ini sangat berguna pada saat data-data yang ada melibatkan data-data outlier. Data outlier data yang keluar jauh dari jangkauan data lainnya seperti atribut *duration, src-byte, dst-byte, hot, num_root,* dan *num_compromised.* Secara umum fungsi normalisasi siqmoid dituliskan dengan persamaan (2 dan 3) [11] sebagai berikut :

$$y = \frac{x_{ik} - \bar{x}_k}{r\,\sigma_k} \quad (2)$$

$$\bar{x}_{ik} = \frac{1}{1 + e^{-y}} \quad (3)$$

dimana :

x_{ik} : dataset tupel
$\bar{x}_k$: nilai mean dari seluruh record dataset
σ : standart deviasi
r : faktor value, r nilai didefinisikan oleh user, di set dengan nilai 1
$\hat{x}_{ik}$: hasil normalisasi
e : konstanta, di set dengan nilai 2.71

2.4. Diskretisasi

Proses diskretisasi dilakukan dengan algoritma K-Means. K-means adalah salah satu metode clustering non-hirarkis yang berusaha memecah data yang ada menjadi satu atau lebih cluster. Metode ini memposisikan data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam cluster dan data yang sama yang memiliki karakteristik berbeda dalam pengelompokan ke dalam kelompok lain. Secara umum algoritma dasar K-Means Clustering adalah sebagai berikut:

- Tentukan jumlah cluster
- Alokasikan data ke dalam cluster secara random
- Hitung centroid/rata-rata dari data yang ada di masing-masing cluster
- Alokasikan masing-masing data ke centroid/rata-rata terdekat
- Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan.

Penentuan centroid secara acak ditentukan dengan rumus persamaan yang dituliskan persamaan (4) [12] sebagai berikut:

$$c_i = \min + \frac{(i-1) * (\max - \min)}{n} + \frac{(\max - \min)}{2 * n} \quad (4)$$

dimana :

c_i : centroid dari kelas ke-i
min : nilai terkecil dari data kelas kontinyu
max : nilai terbesar dari data kelas kontinyu
n : jumlah kelas diskret

Distance space digunakan untuk menghitung jarak antara data dan centroid. Adapun persamaan yang dapat digunakan salah satunya yaitu *Euclidean distance space*. *Euclidean distance space* sering

digunakan dalam perhitungan jarak, hal ini dikarenakan hasil yang diperoleh merupakan jarak terpendek antara dua titik yang diperhitungkan, persamaan dituliskan (5) [13] sebagai berikut:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \{x_{ik} - x_{jk}\}^2} \quad (5)$$

dimana

d_{ij} : jarak objek antara obyek i dan j

P : dimensi data

2.5. Seleksi Fitur

Seleksi fitur adalah salah satu teknik penting dan sering digunakan dalam *pre-processing*. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi [14]. Pada penelitian ini proses *feature selection* akan dilakukan dengan *tools* WEKA. Tujuan utama dari seleksi fitur ialah memilih fitur terbaik dari suatu kumpulan fitur data [15].

Metode seleksi fitur yang umum digunakan adalah *Corellation Feature Selection* (CFS) dan *Information Gain*. *Information Gain* merupakan teknik seleksi fitur yang memakai metode *scoring* untuk nominal ataupun pembobotan atribut kontinyu yang didiskretkan menggunakan maksimal entropy. Suatu entropy digunakan untuk mendefinisikan nilai *Information Gain*. Secara matematis dituliskan dengan (6 dan 7) [12].

$$Entropy\ (S) = -\sum_{i=1}^{n} \frac{|s_i|}{|s|} \log \frac{|s_i|}{|s|} \quad (6)$$

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \quad (7)$$

dimana :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

$|s_i|$: jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Fitur yang dipilih adalah fitur dengan nilai *Information Gain* yang tidak sama dengan nol dan lebih besar dari suatu nilai *threshold* tertentu. Atribut terpilih dari proses ini adalah {3, 29, 4, 28, 22, 39, 2, 26, 34, 40, 27, 23, 32, 33, 30, 24, 38, 37, 31, 25, 1, 21, 35, 10, 11, 12, 36, 6, 5, 14, 13, 16, 8, 9, 20, 18, 19, 7, 15, 17} dengan 42 atribut. Ide dibalik *Information Gain* untuk memilih fitur adalah menyatakan fitur dengan informasi yang paling signifikan terhadap kategori.

Metode *Corellation Feature Selection* (CFS) merupakan bagian evaluasi heuristik yang memperhatikan manfaat fitur individu untuk prediksi kelas bersama-sama dengan level antar-korelasi di antara mereka. CFS menempatkan skor tinggi sebagai subset data yang mengandung fitur dengan korelasi tinggi dengan kelas tetapi memiliki antar-korelasi rendah satu dengan yang lain [16]. Berdasarkan pada sifat tersebut CFS dapat dedifinisikan pada persamaan 8 [17].

$$CFS = \max \left(\frac{(r_{cf1} + r_{cf1} + \cdots + r_{cfn})}{\sqrt{k + 2(r_{f1f2} + r_{fifj} + \cdots + r_{fkf1}}} \right) \quad (8)$$

CFS akan memilih sejumlah atribut yang memiliki nilai terbesar yang menyatakan kelayakan dari atribut tersebut untuk dipilih, r_{cf1} menyatakan korelasi atribut $f1$ dengan kelas c, dan r_{f1f2} adalah nilai korelasi antara fitur 1 dan fitur 2 dalam dataset tersebut. Atribut yang terpilih dari proses *feature selection* ini dengan menggunakan CFS adalah atribut { 2,3,4,11,14,17,29,30} dengan atribut 42.

3. Hasil dan Pembahasan

Dataset pada penelitian ini adalah dataset NSL-KDD'99, dataset ini telah melalui beberapa proses sebelumnya. Data yang digunakan adalah sebanyak 10.000 data yang terdiri dari 39 jenis serangan. Dari 39 jenis serangan tersebut dikelompokkan menjadi 4 kategori yaitu DOS, PROBE, R2L dan U2R. Metode pengujiannya dilakukan dengan cross-validation, pembagian kedalam subset 10 k-fold. Hasil ekseperimen akan menggambarkan kemampuan dari setiap pengujian dalam mengklasifikasikan suatu data kedalam 5 kelas tersebut beserta dengan hasil analisisnya. Berikut merupakan hasil percobaan pada setiap pengujian yang dilakukan.

a. Pengujian pertama

Pengujian pertama dilakukan tanpa proses normalisasi dan diskretisari variabel. Dalam percobaan ini telah melewati pembersihan data dan data bernilai kontiyu dengan fungsi densitas gaus. Berdasarkan hasil ujicoba yang dilakukan diperoleh hasil tingkat akurasi klasifikasi serangan sebesar 80% dengan confusion matrik disajikan pada Tabel 3.

Tabel 3. Hasil pengujian pertama

Klasifikasi	NORMAL	PROBE	DOS	R2L		U2R
NORMAL	1019	234	54	201		317
PROBE	29	1890	38	0		95
DOS	672	895	1645	277		142
R2L	44	135	8	638		1611
U2R	0	0	0	1		55

Terlihat hasil pada Tabel 3. hasil pengujian tanpa melalui normalisasi dan diskretisasi variabel kemampuan klasifikasi serangan dengan *naive bayes* tingkat akurasinya hanya berkisar pada 80%. Hal ini terjadi karena atribut dengan nilai kontinyu akan memiliki probabilitas kemunculan yang sangat kecil dalam dataset sehingga data tersebut tidak dapat diklasifikasikan dengan baik.

b. Pengujian kedua

Pengujian kedua dilakukan dengan proses normalisasi dan diskretisasi variabel tanpa seleksi fitur. Dalam percobaan ini telah melewati pembersihan data dan diskretisasi variabel bernilai kontinyu. Berdasarkan hasil ujicoba yang dilakukan diperoleh tingkat akurasi klasifikasi serangan sebesar 94% dengan confusion matrik disajikan pada Tabel 4.

Tabel 4. Hasil pengujian kedua

Klasifikasi	NORMAL	PROBE	DOS	R2L	U2R
NORMAL	1344	86	82	288	25
PROBE	234	1784	27	7	0
DOS	541	24	3025	39	2
R2L	422	112	11	1876	15
U2R	0	0	1	27	28

Bersadarkan confusion matrik terlihat hasil pada pengujian kedua dengan nilai akurasi 94,4% meningkat dari hasil pengujian sebelumnya.

c. Pengujian ketiga

Pengujian ketiga dilakukan dengan proses normalisasi dan diskretisari variabel dengan seleksi fitur. Dalam percobaan ini telah melewati proses normalisasi, diskretisasi variabel dan seleksi fitur menggunakan *Information Gain* (IG) dan *Corellation based Feature Selection* (CFS). Berdasarkan hasil ujicoba menggunakan seleksi fitur *Information Gain* diperoleh tingkat akurasi klasifikasi serangan sebesar 94,6%, meningkat 0.2% dari pengujian kedua tanpa seleksi fitur dengan confusion matrik yang disajikan pada Tabel 5. dan Tabel 6.

Tabel 5. Hasil ujicoba dengan seleksi fitur IG

Klasifikasi	NORMAL	PROBE	DOS	R2L	U2R
NORMAL	1343	86	82	286	28
PROBE	235	1782	27	7	1
DOS	452	24	3024	39	2
R2L	422	112	11	1879	12
U2R	0	0	1	28	27

Tabel 6. Hasil ujicoba dengan seleksi fitur CFS

Klasifikasi	NORMAL	PROBE	DOS	R2L	U2R
NORMAL	1173	99	315	231	7
PROBE	118	1824	92	18	0
DOS	51	15	3565	0	0
R2L	433	67	21	1909	6
U2R	0	0	9	24	23

Hasil ujicoba dengan seleksi fitur *Corellation based Feature Selection* diperoleh tingkat akurasi klasifikasi sebesar 95,6% dan meningkat 1.2% lebih tinggi dari seleksi fitur menggunakan *Information Gain.*

d. Perbandingan nilai klasifikasi serangan pada setiap pengujian

Bagian ini dibahas mengenai hasil perbandingan terhadap kemampuan klasifikasi serangan dari setiap pengujian yang telah dilakukan. Klasifikasi serangan yang dimaksudkan adalah sistem yang disusun mampu menentukan apakah pola tersebut masuk kedalam pola normal atau serangan tanpa meperhitungan jenis serangan yang dideteksi. Hasil perbandingan disajikan pada Tabel 7.

Tabel 7. Hasil perbandingan dengan beberapa ujicoba

Ujicoba	Akurasi
Pengujian pertama	80%
Pengujian kedua	94,4%
Pengujian ketiga dengan IG	94,6%
Pengujian ketiga dengan CFS	95,6%

Pada Tabel 7. terlihat bahwa tingkat akurasi terbaik adalah pada pengujian ketiga yang telah melalui proses normalisasi, diskretisasi dengan K-Means dan seleksi fitur menggunakan *Corellation based Feature Selection* (CFS) dengan nilai akurasi klasifikasi serangan sebesar 95,6%. Maka dari hasil pembahasan penelitian ini dapat memberikan manfaat bahwa dengan teknik/metode yang digunakan dalam penelitian ini memiliki kemampuan yang baik untuk sistem pendeteksi pada Intrusion Detection System.

4. Simpulan

Metode yang diusulkan pada penelitian ini memberikan hasil sangat baik. Klasifikasinya meningkat jika dibandingkan dengan proses klasifikasi tanpa melakukan diskretisasi variabel. Jadi dapat disimpulkan bahwa proses diskritisasi dengan K-Means Clustering menjadikan algoritma naive bayes dapat meningkatkan kinerjanya pada klasifikasi serangan.

Daftar Pustaka

[1] K. Scarfone, P. Mell. *Guide to Intrusion Detection and Prevention Systems (IDPS)*, Computer Security Resource Center (National Institute of Standards and Technology). 2007; 800–94.

[2] Neethu, B. Classification of Intrusion Detection Dataset Using Machine Learning Approaches. *International Journal of Electronics and Computer Science Engineering*. 2012; pp.1044-51.

[3] Olusola, A.A., Oladele, A.S. & Abosede, D.O. *Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Feature*. In World Congress on Engineering and Computer Science. San Fransisco. 2010.

[4] Kusrini, K. *Grouping of Retail Items by Using K-Means Clustering*. The Third Information Systems International Conference. Surabaya. 2015; Vol. 72, Pages 495–502.

[5] Kusrini, K, Iskandar, M.D., Wibowo F. W. *Multi Features Content-Based Image Retrieval Using Clustering And Decision Tree Algorithm*. TELKOMNIKA Telecommunication, Computing, Electronics and Control. Yogyakarta. 2016; Vol 14 No 4; Halaman : 1480-1492.

[6] Kusrini. *Pendiskritan Kelas Kontinyu dengan Algoritma K-Mean Cluster*. Jurnal Dasi. 2010; Vol 11 No 4.

[7] Hettich. *The UCI KDD Archive.* California: Department of Information and Computer Science. 1999

[8] Patil, T. R., Sherekar, M. S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal of Computer Scienceand Applications*. 2013; Vol. 6, No. 2, Hal 256-261.

[9] Ridwan, M., Suyono, H., Sarosa, M. *Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier*, *Jurnal EECCIS.* 2013; Vol 1, No. 7, Hal. 59-6.

[10] Bustami. Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, *TECHSI : Jurnal Penelitian Teknik Informatika*. 2013; Vol. 3, No.2, Hal. 127-146.

[11] Prasetyo, E. *Data Mining : Mengolah data menjadi informasi menggunakan matlab*. Yogyakarta. 2014.

[12] Kusrini, E. T. *Algoritma Data Mining*. Yogyakarta: Penerbit Andi. 2009.

[13] Putra, Darma. *Pengolahan Citra Digital*, Yogyakarta : Penerbit Andi. 2010.

[14] Djatna, Taufik & Morimoto, Yasuhiko. *Pembandingan Stabilitas Algoritma Seleksi Fitur Menggunakan Transformasi Ranking Normal*. Jurnal Ilmiah Ilmu Komputer, Institut Pertanian Bogor. 2008; Vol. 6 No. 2, ISSN : 1693 -1629.

[15] Abadi, Delki. *Perbandingan Algoritme Feature Selection Information Gain Dan Symmetrical Uncertainty Pada Data Ketahanan Pangan*. Skripsi, Institut Pertanian Bogor, Bogor. 2013.

[16] ZHAO Jing-Jing, HUANG Xiao-Hong, SUN Qion, MA Yan, Real time feature selection in traffic classification, *The Journal of China Universities of Posts and Telecommunication*, 2008.

[17] Bolón-Canedo, "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," *Expert Systems with Applications.* 2011; No. 38, p. 5947–5957.