e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 4 No. 2, Desember, 2020, **Hal. 40-49**
DOI: 10.29408/edumatic.v4i2.2433

Optimasi Algoritma *Naïve Bayes* Classifier untuk Mendeteksi *Anomaly* dengan Univariate Fitur Selection

Harianto[1], Andi Sunyoto[2], Sudarmawan[3].
[1,2,3] Program Studi Teknik Informatika, Universitas AMIKOM Yogyakarta
email: harianto.27@students.amikom.ac.id[1], email: andi@amikom.ac.id[2], email: sudarmawan@amikom.ac.id[3]

(Received: 29 Juli 2020/ Accepted: 19 Agustus 2020 / Published Online: 20 Desember 2020)

Abstrak

Keamanan sistem dan jaringan dari gangguan pihak yang tidak memilik hak akses terhadap sistem merupakan hal yang terpenting dalam sebuah sistem. Untuk mewujudkan sistem, data atau jaringan yang aman dari pengguna yang tidak sah atau gangguan lainnya, maka diperlukan sistem untuk mendeteksinya. *Intrusion-Detection System* (IDS) adalah sebuah cara yang dapat digunakan untuk mendeteksi aktivitas yang mencurigakan dalam sebuah sistem atau jaringan. Algoritma klasifikasi pada kecerdasan buatan bisaditerapkan pada permaalahan ini. Ada banyak algoritma klasifikasi yang bisa digunakan salah satunya adalah *Naïve Bayes*. Penelitian ini bertujuan untuk mengoptimasi *Naïve Bayes* menggunakan *Univariate Selection* pada *data set* UNSW-NB 15. Fitur yang digunakan hanya mengambil 40 fitur yang memiliki relevansi terbaik. Kemudian *data set* dibagi dua *data test* dan *data training* yaitu 10%:90%, 20%:70%, 30%:70%, 40%:60% dan 50%:50%. Dari percobaan yang dilakukan didapatkan seleksi fitur cukup berpengaruh terhadap nilai akurasi yang didapatkan. Nilai akurasi tertinggi didapatkan ketika *data set* dibagi menjadi 40%:60% baik pada seleksi fitur atau tanpa seleksi fitur. *Naïve Bayes* dengan fitur yang tidak diseleksi memperoleh nilai akurasi tertinggi 91.43% sedangkan dengan seleksi fitur 91.62%, dengan menggunakan seleksi fitur dapat meningkatkan nilai akurasi sebesar 0.19%.

Kata kunci: *Anomaly, Naïve Bayes, Univariate Selection.*

Abstract

System and network security from interference from parties who do not have access to the system is the most important in a system. To realize a system, data or network that is safe at unauthorized users or other interference, a system is needed to detect it. Intrusion-Detection System (IDS) is a method that can be used to detect suspicious activity in a system or network. The classification algorithm in artificial intelligence can be applied to this problem. There are many classification algorithms that can be used, one of which is Naïve Bayes. This study aims to optimize Naïve Bayes using Univariate Selection on the UNSW-NB 15 data set. The features used only take 40 features that have the best relevance. Then the data set is divided into two test data and training data, namely 10%: 90%, 20%: 70%, 30%: 70%, 40%: 60% and 50%: 50%. From the experiments carried out, it was found that feature selection had quite an effect on the accuracy value obtained. The highest accuracy value is obtained when the data set is divided into 40%: 60% for both feature selection and non-feature selection. Naïve Bayes with unselected features obtained the highest accuracy value of 91.43%, while with feature selection 91.62%, using feature selection could increase the accuracy value by 0.19%.

Keywords: *Anomaly, Naïve Bayes, Univariate Selection.*

PENDAHULUAN

Sistem keamanan dunia maya secara global digunakan untuk melindungi data, informasi dan komputer dari serangan, perusakan, dan akses yang tidak sah. Secara khusus, *Intrusion-Detection systems* (IDS) telah diusulkan sebagai alat yang efektif untuk memantau dan mengontrol aktivitas pada jaringan, untuk membantu dalam menentukan penggunaan yang sah dan tidak sah, untuk mengidentifikasi kerusakan sistem informasi, dan untuk melindungi sistem dari intrusi internal dan eksternal (intrusi dari dalam atau dari luar perusahaan) (Alhakami et al., 2019). *Anomaly* dapat didefinisikan dengan berbagai cara atau teknik misalnya penyimpangan dalam amplitudo, nilai yang dimasukkan secara acak, kurangnya data, data dari berbagai jenis dan asal data, dan tersirat(Karczmarek et al., 2020).

Serangan-serangan yang terjadi pada jaringan terdiri dari banyak jenis diantaranya seperti serangan pada protokol SSH (*Secure Shell*) (Arkaan & Sakti, 2019) dan *Maritime Anomaly Detection* (Zhen et al., 2017). Serangan atau *anomaly* sangat mempengaruhi dan berbahaya terhadap data dan sistem. Dengan besarnya pengaruh tersebut maka harus dilakukan deteksi dengan metode dan algoritma yang bermacam-macam *Anomaly detection with Density Estimation* (Nachman & Shih, 2020) dan masih banyak metode lainnya.

Ada banyak metode klasifikasi yang populer dan banyak digunakan oleh para peneliti untuk mendeteksi *anomaly* dan kasus lainnya di antaranya adalah *K-Means Clastering* (Ridho & Kusuma, 2019), NBC dan SVM (Riadi et al., 2019), *Neural Network* (NN) (Ramdhani et al., 2018). *Naïve Bayes* digunakan untuk mengklasifikasikan *anomaly* IDS (*Intrusion-Detection System*) dan untuk pemilihan atribut dengan teknik korelasi (*correlation-based feature selection*) (Anwar et al., 2019). Penelitian yang telah dilakukan menggunakan koleksi data *Intrusion-Detection system* UNSW-NB15 yang terdiri dari 49 atribut dan 321.283 record data. Hasil evaluasi klasifikasi *anomaly* IDS menggunakan algoritma *Naïve Bayes* tanpa didahului atribut yang diseleksi dengan teknik korelasi diperoleh tingkat akurasi 71,2 %. Sedangkan hasil klasifikasi jika didahului dengan atribut yang diseleksi dengan teknik korelasi didapatkan akurasi 74,8% (Anwar et al., 2019).

NBC lebih banyak dan lebih tepat diterapkan pada data yang jumlahnya lebih besar dan dapat menangani data yang tidak lengkap (*missing value*) serta dapat menangani gangguan pada data dan kuat terhadap atribut yang tidak sesuai atau tidak relevan. Akan tetapi disamping NBC memiliki kelebihan, NBC juga memiliki kelemahan dimana sebuah probabilitas tidak bisa mengukur seberapa besar tingkat keakuratan sebuah prediksi. Selain itu, NBC juga memiliki kelemahan pada seleksi atribut sehingga dapat mempengaruhi nilai akurasi. *Naïve Bayes* masih tidak dapat memberikan kinerja yang memuaskan karena kurangnya jumlah sampel pelatihan yang cukup dengan *label* yang tepat dan fungsi distribusi probabilitas eksplisit dari lalu lintas dalam jaringan yang dikendalikan dan menghasilkan tingkat akurasi sebesar 76 % (Han et al., 2019). Oleh karena itu, NBC perlu dioptimasi dengan cara memberikan bobot pada atribut agar NBC dapat bekerja lebih efektif. Banyak metode optimasi yang digunakan untuk mengoptimasikan NBC menggonakan *Particle Swarm Optimization* (Koeswara et al., 2020), optimasi metode *Naïve Bayes* dengan algoritma genetika (Handayanna et al., 2017) dan metode *cross validation* (Samponu & Kusrini, 2018).

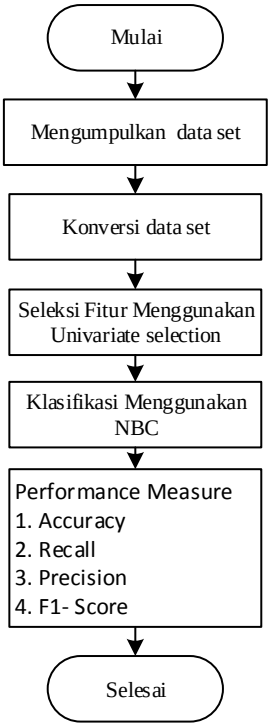
. Untuk mengatasi permasalahan tersebut, metode seleksi fitur dapat digunakan untuk melakukan pembobotan atribut yang akan digunakan pada proses klasifikasi untuk meningkatkan akurasi *Naïve Bayes.* Untuk melakukan seleksi fitur digunakan metode *Univariate features selection*. *Univariate features selection* secara umum bekerja dengan cara memilih fitur terbaik berdasarkan *test statistic univariate*. Hal ini dapat diketahui sebagai langkah *preprocess* sebuah *estimator*. *Select k-best* secara khusus bekerja dengan cara memilih sejumlah *k-features* terbaik berdasarkan pengujian statistik (Varoquaux et al., 2015). Algoritma NBC ketika menurunkan jumlah fitur terdapat kemungkinan terjadi kenaikan akurasi dengan menghilangkan fitur yang memiliki relevansi kecil (Rahmansyah et al., 2018).

Menghilangkan fitur yang memiliki relevansi kecil dihilangkan, karena berpengaruh terhadap nilai akurasi yang dihasilkan NBC. Sehingga pada penelitian ini dilakukan percobaan seleksi fitur dengan *data set* UNSW-NB15 untuk melihat pengaruh fitur yang memiliki relevansi kecil.

Pada penelitian yang dilakukan ini, NBC digunakan untuk melakukan deteksi *anomaly* dengan atribut diseleksi terlebih dahulu menggunakan *Univariate selection*. Hasil seleksi atribut kemudian dilakukan beberapa kali percobaan kemudian dibandingkan hasil nilai yang didapatkan dengan percobaan tanpa seleksi fitur. *Univariate selection* akan diterapkan untuk pemilihan parameter atau pembobotan atribut pada NBC yang sesuai dan optimal, sehingga hasil klasifikasi *anomaly* lebih akurat. Hasil nilai percobaan seleksi fitur dan tanpa seleksi fitur di bandingkan sehingga mendapatkan hasil akhir seberapa besar pengaruh seleksi fitur terhadap NBC menggunakan *data set* UNSW-NB15 untuk mendeteksi *anomaly*.

METODE

Pada penelitian ini dilakukan percobaan untuk mendeteksi *anomaly* pada jaringan menggunakan *data set* UNSW-NB15. Algoritma NBC akan dilakukan optimasi dengan melakukan seleksi 40 fitur terbaik dan menghilangkan fitur yang memilikirelevansi kecil. Pada penelitian ini penulis menyusun desain penelitian berupa tahapan-tahapan penelitian agar penelitian dapat dilakukan secara sistematis. Penelitian ini dilakukan dengan mengumpulkan *data set* UNSW-NB15, melakukan konversi *data set*, seleksi fitur menggunakan *Univariate selection*, klasifikasi NBC dan terakhir melakukan perbandingan nilai akurasi. Alur penelitian bias dilihat pada gambar 1.



Gambar 1. Alur Penelitian

Pahapan pertaman pada penelitian ini mengumpulkan *data set*. *Data set* yang digunakan pada penelitian ini adalah *data set* UNSW-NB15 yang diambil pada https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-*Data sets*/. *Data set* yang diambil terdiri dari beberapa tipe data yaitu angka dan *string*. Jumlah data yang dibunakan adalah hanya 82.332. Tahap kedua yaitu konversi *data set***,** ini yang dilakukan adalah merubah data yang berjenis *string* atau *text* menjadi angka. Hal ini bertujuan supaya

program yang digunakan bias membaca datset tersebut ketika melakukan seleksi fitur. Konversi data dilakukan dengan cara merubah karakter menjadi kode ASCII kemudian menjumlahkannya. Hasil penjumlahan tersebut yang digunakan sebagai nilai dari atribut yang akan digunakan pada proses seleksi fitur. Tahapan ketiga seleksi fitur menggunakan teknik atau metode *Univariate Selection*. Teknik ini dilakukan dengan cara mencari nilai probabilitas dari sebuah *variable* dan melakukan perengkingan terhadap nilai yang didapatkan. Fitur yang diambil untuk melakukan klasifikasi sebanyak 40 fitur yang memiliki bobot atau skor yang terbaik. Tahapan keempat melakukan klasifikasi menggunakan NBC. Pada tahap ini dilakukan sebanyak 16 kali percobaan. Masing-masing hasil fitur yang sudah diseleksi dilakukan percobaan sebanyak 4 kali. Tahapan kelima yaitu membandingkan nilai akurasi, *recall*, *precision* dan *f1-score* terhadap masing-masing pengujian.

HASIL DAN PEMBAHASAN

Hasil

Penelitian yang dilakukan adalah melakukan klasifikasi *anomaly* pada jaringan apakah akses yang ada termasuk *anomaly* atau normal. Dari *data set* yang digunakan terdapat 45 fitur dan di seleksi 40 fitur terbaik yang memiliki relevansi tinggi. Seleksi fitur dilakukan menggunakan metode *Univariate selection.* Dari seluruh fitur dari *data set* pada tabel 1 dilakukan seleksi 40 fitur terbaik yang memiliki nilai bobot yang tinggi berdasarkan probabilitas dengan teknik *Univariate selection*. Berdasarkan hasil seleksi fitur yang dilakukan didapatkan 40 fitur terbaik ditunjukkan pada gambar 2.

	Specs	Score
22	dtcpb	1.172015e+13
21	stcpb	1.162998e+13
12	sload	6.400680e+11
13	dload	5.884417e+10
9	rate	2.383129e+09
0	id	1.696285e+08
8	dbytes	1.519980e+08
7	sbytes	1.292759e+08
16	sinpkt	6.076319e+07
18	sjit	3.124874e+07
30	response_body_len	2.014811e+07
28	dmean	1.895793e+06
20	swin	1.719167e+06
17	dinpkt	1.596148e+06
19	djit	1.496485e+06
23	dwin	1.422332e+06
10	sttl	1.191638e+06
3	service	1.063406e+06
43	attack_cat	6.158226e+05
6	dpkts	2.371761e+05

34	ct_src_dport_ltm	1.371216e+05
35	ct_dst_sport_ltm	1.218852e+05
11	dttl	1.138066e+05
36	ct_dst_src_ltm	1.127933e+05
27	smean	9.587950e+04
41	ct_srv_dst	9.534928e+04
31	ct_srv_src	8.932627e+04
15	dloss	7.984000e+04
40	ct_src_ltm	7.103211e+04
33	ct_dst_ltm	6.759774e+04
5	spkts	6.082648e+04
2	proto	3.712849e+04
44	label	3.700000e+04
32	ct_state_ttl	6.947356e+03
14	sloss	2.927817e+03
4	state	2.644695e+03
39	ct_flw_http_mthd	1.457141e+03
42	is_sm_ips_ports	1.122273e+03
24	tcprtt	4.387749e+02
25	synack	3.176365e+02

Gambar 2. Fitur yang diseleksi

Dari 40 fitur yang diseleksi, berarti terdapat 5 fitur yang memiliki bobot yang rendah. Adapun fitur-fitur tersebut adalah, is_ftp_login, dur, ackdat, ct_ftp_cmd dan trans_depth. Percobaan yang dilakukan dengan menggunakan hasil fitur yang diseleksi menggunakan teknik *Univariate selection* dan non seleksi dilakukan percobaan dengan membagi *data set* menjadi *data test* dan *data training* sebanyai 5 kali. Sedangkan untuk hasil percobaan untuk *data set* seleksi fitur yaitu 40 fitur terbaik bisa dilihat pada tabel 4.

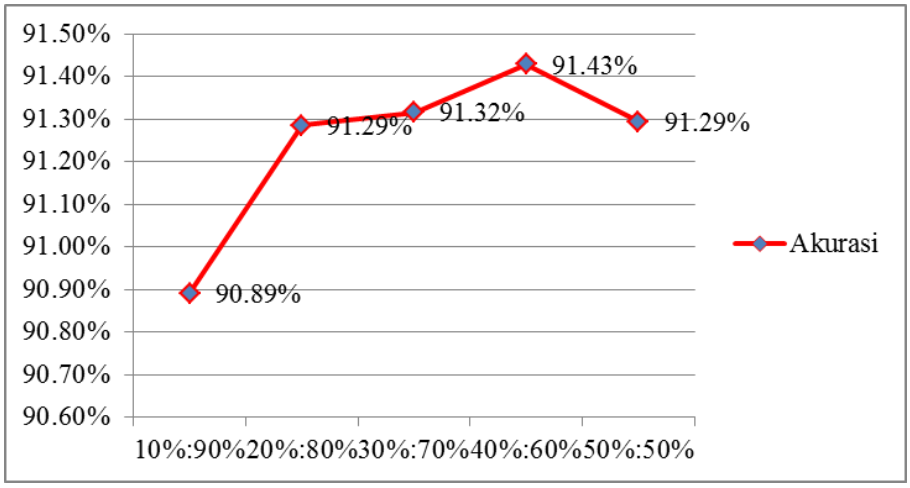
Tabel 1. *Confusion Matrix, Accuracy, Recall Precision* dan *F1-Score* fitur tanpa seleksi

No.	*Data set*	*Confusion Matrix*		*Accuracy*	*Recall*	*Precision*	*F1-Score*
1	10%:90%	3,324	352	90.89%	89.31%	90.42%	89.86%
		398	4,160				
2	20%:80%	6,627	711	91.29%	90.15%	90.31%	90.23%
		724	8,405				
3	30%:70%	9,897	1,096	91.32%	90.42%	90.03%	90.22%
		1,049	12,658				
4	40%:60%	13,206	1,478	91.43%	90.76%	89.93%	90.34 %
		1,345	16,904				
5	50%:50%	16,578	1,852	91.29%	90.54%	89.95%	90.24%
		1,732	21,004				

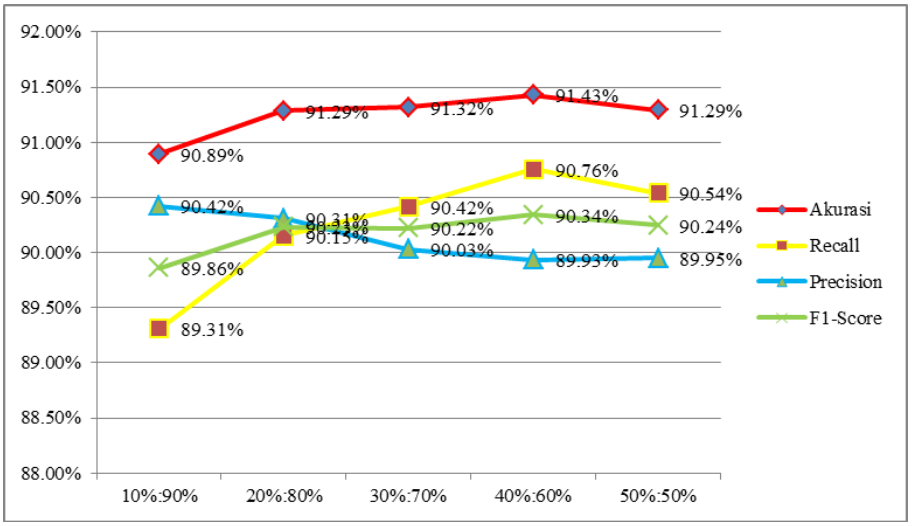
Tabel 2. *Confusion Matrix, Accuracy, Recall Precision dan F1-Score seleksi 40 fitur*

No.	*Data set*	*Confusion Matrix*		*Accuracy*	*Recall*	*Precision*	*F1-Score*
1	10%:90%	3,329	347	91.09 %	89.59%	90.56%	90.07%
		387	4,171				
2	20%:80%	6,633	705	91.45 %	90.42%	90.39%	90.40%
		703	8,426				
3	30%:70%	9,916	1,077	91.54%	90.74%	90.20%	90.47%
		1,012	12,695				
4	40%:60%	13,232	1,452	91.62 %	91.00%	90.11%	90.56%
		1,308	16,941				
5	50%:50%	16,600	1,830	91.40 %	90.66%	90.07%	90.36%
		1,710	21,026				

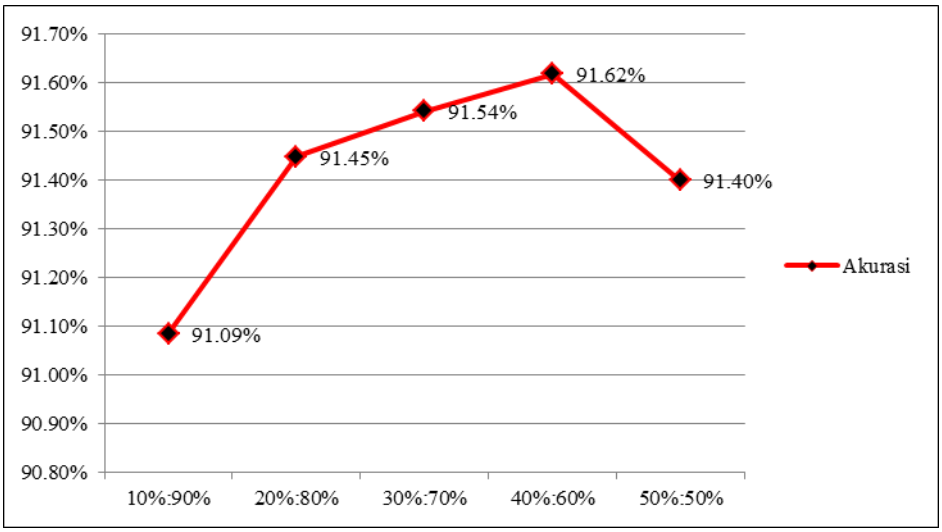
Pada percobaan yang dilakukan sebanyak 5 kali pada *data set* tanpa seleksi fitur, nilai akurasi tertinggi didapatkan pada saat *data set* dabagi menjadi 40% *data test* dan 60% *data training*. Adapun nilai akurasinya yaitu 91.43%. Hasil perbandingan nilai akurasi percobaan yang dilakukan pada *data set* tanpa seleksi fitur dapat dilihat pada gambar 3. Nilai *recall* tertinggi 90.76% disaat *data set* dibagi menjadi 40%:60%. Nilai *precision* tertinggi disaat *data set* dibagi menjadi 10%:90% yaitu 90.42%. sedangkan untuk nilai *F1-Score* tertinggi pada saat *data set* dibagi menjadi 40%:60% dengan nilai 90.34%. Untuk perbandingan nilai akurasi tanpa seleksi fitur *recall, precision dan F1-Score* bisa dilihat pada gambar 4. Sedangkan untuk hasil perbandingan nilai akurasi percobaan yang dilakukan pada *data set* seleksi fitur dapat dilihat pada gambar 3 dan perbandingan nilai akurasi *recall, precision dan F1-Score* seleksi fitur bisa dilihat pada gambar 6.



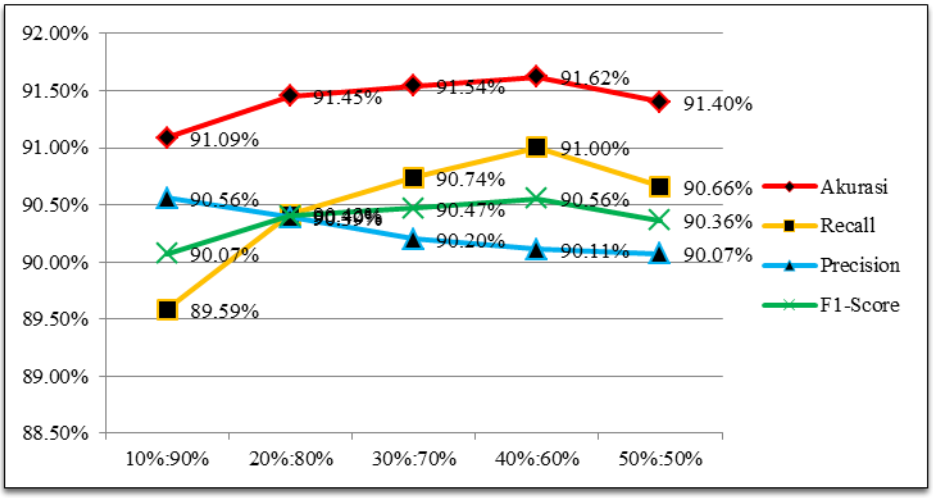
Gambar 3. Perbandingan nilai akurasi berdasarkan *data set* untuk tanpa seleksi fitur



Gambar 4. Perbandingan nilai *accuracy*, *recall*, *precision* dan *F1-Score* berdasarkan *data set* untuk seleksi fitur

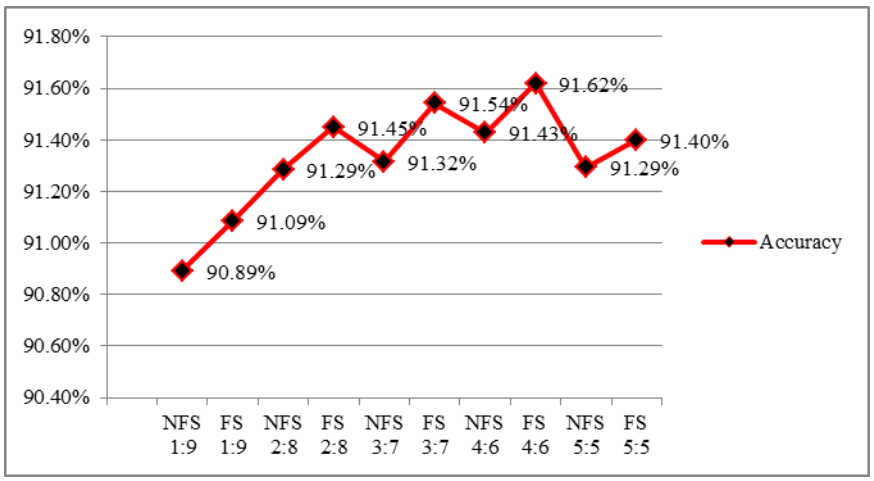


Gambar 5. Perbandingan nilai akurasi berdasarkan *data set* untuk seleksi fitur



Gambar 6. Perbandingan nilai akurasi, *recall*, *precision* dan F1-Score seleksi fitur

Setelah melakukan perbandingan nilai yang didapakan antara percobaan dengan *data set* yang tanpa seleksi fitur dan percobaan dengan *data set* seleksi fitur, selanjutnya dilakukan perbandingan. Perbandingan nilai akurasi yang dilakukan adalah perbandingan nilai akurasi dasi semua percobaan. Dari seluruh percobaan nilai akurasi tertinggi 91.62% didapatkan pada NBC dengan seleksi fitur dengan *data set* 40% *data test* dan 60% *data training*.Perbandingan nilai akurasi bisa dilihat pada gambar 7. Untuk perbandingan selisih dan atau peningkatan nilai akurasi yang didapatkan menggunakan NBC dengan seleksi fitur dan tanpa seleksi fitur bisa dilihat pada tabel 3.



Gambar 7. Perbandingan nilai akurasi seluruh percobaan

Tabel 3. Selisih Peningkatan nilai akurasi NBC dengan NBC + seleksi fitur

Data set	NFS	FS	Selisih/Peningkatan
10%:90%	90.89%	91.09%	0.19%
20%:80%	91.29%	91.45%	0.16%
30%:70%	91.32%	91.54%	0.23%
40%:60%	91.43%	91.62%	0.19%
50%:50%	91.29%	91.40%	0.11%

Pembahasan

Koleksi data yang dipakai pada penelitian ini ialah UNSW-NB15 tahun 2015. UNSW-NB15 merepresentasikan sembilan besar mayoritas serangan dengan menggunakan IXIA *Perfect Storm Tool* dari simulasi yang dilakukan dengan periode waktu 16 jam pada 22 Januari 2015 dan 15 jam pada 17 Februari 2015 untuk merekam 100 GBs data. Ada 49 atribut yang telah dihasilkan dengan menggunakan *Argus*, *Bro-IDS tool* dan dua belas algoritma yang dibangun dengan bahasa C# yang mencakup karakteristik paket jaringan (Moustafa & Slay, 2015). Dari *data set* awal sebanyak 2.540.044 *record* diambil sampling sebanyak 82.332 *record* data. Dari data tersebut terdapat 45 atribut. Koleksi data yang diapakai adalah data secara keseluruhan tipe dengan jumlah 82.332. distribusi data yang digunakan dapat dilihat pada tabel 5.

Tabel 4. Atribut *Data set* UNSW-NB15

No.	Nama	Tipe	15	dloss	*Integer*	30	sinpkt	*Float*
1	dur	*Float*	16	Sjit	*Float*	31	dinpkt	*Float*
2	proto	*Nominal*	17	djit	*Float*	32	ct_state_ttl	*Integer*
3	service	*Nominal*	18	swin	*Integer*	33	ct_dst_ltm	*Integer*
4	state	*Nominal*	19	dwin	*Integer*	34	ct_src_dport_ltm	*Integer*
5	spkts	*Integer*	20	stcpb	*Integer*	35	ct_dst_sport_ltm	*Integer*
6	dpkts	*Integer*	21	dtcpb	*Integer*	36	ct_dst_src_ltm	*Integer*
7	sbytes	*Integer*	22	tcprtt	*Float*	37	is_ftp_login	*Binary*
8	dbytes	*Integer*	23	synack	*Float*	38	ct_ftp_cmd	*Integer*
9	rate	*Float*	24	ackdat	*Float*	39	ct_flw_http_mthd	*Integer*
10	sttl	*Integer*	25	smean	*Integer*	40	ct_src_ltm	*Integer*
11	dttl	*Integer*	26	dmean	*Integer*	41	ct_srv_dst	*Integer*
12	sload	*Float*	27	trans_depth	*Integer*	42	is_sm_ips_ports	*Binary*
13	dload	*Float*	28	Response _body_len	*Integer*	43	attack_cat	*Nominal*
14	sloss	*Integer*	29	ct_srv_src	*Integer*	44	label	*Binary*

Tabel 5. Distribusi *Data set* UNSW-NB15

Type	*Records*
Normal	37.000
Fuzzers	6.062
Analysis	677
Backdoor	583
DoS	4.089
Exploits	11.132
Reconnaissance	3.496
Shellcode	378
Worms	44
Generic	18.871

Percobaan yang dilakukan dengan *data set* UNSW-NB15 yang dilakukan sebanyak 5 kali percobaan pada masing-masing *data set*. Lima kali percobaan pada *data set* dengan tanpa seleksi fitur dan 5 kali dengan *data set* seleksi fitur. Percobaan dengan *data set* tanpa seleksi fitur dilakukan sebanyak 5 kali dan dengan *data set* seleksi fitur 5 kali, sehingga total

percobaan sebanyak 10 kali. Masing-masing di pecah menjadi 5 kali percobaan dengan membagi *data set* menjadi *data test* dan *data training* 10%:90%, 20%:80%, 30%:70%, 40%:60% dan 50%:50%.

Percobaan pertama *data set* tanpa seleksi fitur dibagi menjadi *data test* 10% dan *data training* 90% menghasilkan nilai akurasi 90.89%. Sedangkan pada *data set* dengan seleksi fitur mendapatkan nilai akurasi 91.09%. Seleksi fitur lebih unggul dibandingkan dengan tanpa seleksi fitur sebesar 0.19%. Percobaan kedua *data set* tanpa seleksi fitur dibagi menjadi 20% *data test* dan 80% *data training* menghasilkan nilai akurasi 91.29%. Sedangkan pada *data set* dengan seleksi fitur mendapatkan nilai akurasi 91.45%. Seleksi fitur lebih unggul 0.16% dibandingkan dengan tanpa seleksi fitur. Percobaan ketiga dilakukan dengan membagi *data set* menjadi 30% *data test* dan 70% *data training*. Tanpa seleksi fitur mendapatkan 91.32% dan seleksi fitur mendapatkan 91.54%. Jika ibandingkan maka NBC dengan seleksi fitur 0.23% lebih unggul dibandingkan dengan tanpa seleksi fitur. Percobaan keempat *data set* 40% dan *data training* 60%. NBC dengan tanpa seleksi fitur mendapatkan nilai akurasi 91.43% sedangkan seleksi fitur mendapatkan 91.62%. bila dibandingkan maka NBC dengan seleksi fitur lebih unggul 0.19% dibandingkan NBC dengan tanpa seleksi fitur. Selanjutnya percobaan yang terakhir dengan *data set* 50% *data test* dan 50% *data training*. NBC dengan tanpa seleksi fitur mendapatkan 91.29% dan NBC dengan tanpa seleksi fitur mendapatkan 91.40%. pada percobaan yang terakhir ini NBC dengan seleksi fitur lebih baik 0.11% dibandingkan NBC tanpa seleksi fitur. Perbandingan antar NBC tanpa seleksi fitur dan NBC seleksi fitur bisa dilihat pada tabel 3 dan gambar 7.

Perbandingan nilai akurasi bisa dilihat pada tabel 3 dan pada gambar 7, NBC dengan seleksi fitur mendapatkan peningkatan sebesar 0.23% yaitu pada saat *data set* dibagi menjadi 30% *data test* dan 70% *data training*. Kemudian dari seluruh percobaan yang dilakukan baik pada NBC dengan tanpa seleksi fitur nilai akurasi tertinggi didapatkan saat *data set* dibagi menjadi 40%:60% dengan nilai akurasi 91.62%. Sedangkan pada percobaan yang dilakukan tanpa fitur yang diseleksi nilai akurasi tertinggi didapatkan pada saat *data set* dibagi menjadi 40%:60% dengan nilai akurasi 91.43%. Penelitian yang dilakukan (Han et al., 2019) menggunakan NBC tanpa fitur seleksi dari lalu lintas dalam jaringan yang dikendalikan dan menghasilkan tingkat akurasi sebesar 76%. Sedangkan pada penelitian ini dengan menggunakan seleksi fitur dengan teknik *Univariate* maka akurasi bisa ditingkatkan sebesar 0.19% dapat dilihat pada tabel 3. Pada penelitian yang dilakukan (Anwar et al., 2019) hasil evaluasi klasifikasi *anomaly* IDS menggunakan algoritma NBC tanpa didahului atribut yang diseleksi dengan teknik korelasi diperoleh tingkat akurasi 71.2%. Sedangkan hasil klasifikasi jika didahului dengan atribut yang diseleksi dengan teknik korelasi didapatkan akurasi 74.8%.

SIMPULAN

Berdasarkan hasil pengujian algoritma *Naïve Bayes* untuk klasifikasi *anomaly* IDS yang diawali pemilihan atribut diperoleh kesimpulan bahwa seleksi fitur cukup berpengaruh terhadap nilai akurasi yang didapatkan. Dengan menambahkan fitur seleksi pada algoritma NBC dapat meningkatkan kinerja algoritma NBC. Dari hasi yang telah dilakukan, bahwa NBC dengan seleksi fitur menggunakan *Univariate selection* lebih unggul dibandingkan dengan NBC tanpa fitur seleksi.

REFERENSI

Alhakami, W., Alharbi, A., Bourouis, S., Alroobaea, R., & Bouguila, N. (2019). Network *Anomaly* Intrusion Detection Using a Nonparametric Bayesian Approach and Feature Selection. *IEEE Access*, *7*, 52181–52190. https://doi.org/10.1109/ACCESS.2019.2912115

Anwar, S., Septian, F., & Septiana, R. D. (2019). Klasifikasi Anomali Intrusion Detection System (IDS) Menggunakan Algoritma Naïve Bayes Classifier dan Correlation-Based Feature Selection. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *2*(4), 135–140. https://doi.org/10.32493/jtsi.v2i4.3453

Arkaan, N., & Sakti, D. V. S. Y. (2019). Implementasi Low Interaction Honeypot Untuk Analisa Serangan Pada Protokol SSH. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *5*(2), 112–120. https://doi.org/10.25077/teknosi.v5i2.2019.112-120

Han, W., Xue, J., & Yan, H. (2019). Detecting anomalous traffic in the controlled network based on cross entropy and support vector machine. *IET Information Security*, *13*(2), 109–116. https://doi.org/10.1049/iet-ifs.2018.5186

Handayanna, F., Rinawati, Arisawati, E., & Dewi, L. S. (2017). Prediksi Penyakit Diabetes Menggunakan Naive Bayes Dengan Optimasi Parameter Menggunakan Algoritma Genetika. *KNiST (Konferensi Nasional Ilmu Sosial & Teknologi)*, 71–76.

Karczmarek, P., Kiersztyn, A., Pedrycz, W., & Al, E. (2020). K-Means-based isolation forest. *Knowledge-Based Systems*, *195*, 105659–105673. https://doi.org/10.1016/j.knosys.2020.105659

Koeswara, T. S. N., Mardiyanto, M. S., & Ghani, M. A. (2020). Penerapan Particle Swarm Optimization (Pso) Dalam Pemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosispenyakit Hepatitis Dengan Metode Naive Bayes. *Journal Speed – Sentra Penelitian Engineering Dan Edukasi*, *12*(1), 1–10.

Nachman, B., & Shih, D. (2020). *Anomaly* detection with density estimation. *Physical Review D*, *101*(7), 075042–075057. https://doi.org/10.1103/PhysRevD.101.075042

Rahmansyah, A., Dewi, O., Andini, P., Hastuti, T., Ningrum, P., & Suryana, M. E. (2018). Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, 1–7. Yokyakarta: Universitas Islam Indonesia

Ramdhani, Y., Susanti, S., Adiwisastra, M. F., & Topiq, S. (2018). *Penerapan Algoritma Neural Network Untuk Klasifikasi Kardiotokografi*. *5*(1), 43–49.

Riadi, I., Umar, R., & Aini, F. D. (2019). Analisis Perbandingan Detection Traffic *Anomaly* Dengan Metode Naive Bayes Dan Support Vector Machine (Svm). *ILKOM Jurnal Ilmiah*, *11*(1), 17–24. https://doi.org/10.33096/ilkom.v11i1.361.17-24

Ridho, F., & Kusuma, A. A. (2019). Deteksi Intrusi Jaringan dengan K-Means Clustering pada Akses Log dengan Teknik Pengolahan Big Data. *Jurnal Aplikasi Statistika & Komputasi Statistik*, *10*(1), 53-66. https://doi.org/10.34123/jurnalasks.v10i1.202

Samponu, Y. B., & Kusrini, K. (2018). Optimasi Algoritma Naive Bayes Menggunakan Metode Cross Validation Untuk Meningkatkan Akurasi Prediksi Tingkat Kelulusan Tepat Waktu. *Jurnal ELTIKOM*, *1*(2), 56–63. https://doi.org/10.31961/eltikom.v1i2.29

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33.

Zhen, R., Jin, Y., Hu, Q., Shao, Z., & Nikitakos, N. (2017). Maritime *Anomaly* Detection within Coastal Waters Based on Vessel Trajectory Clustering and Naïve Bayes Classifier. *Journal of Navigation*, *70*(3), 648–670.