KLASIFIKASI SELEKSI ATRIBUT PADA SERANGAN SPAM MENGGUNAKAN METODE ALGORITMA DECISION TREE

Aji Sudibyo [1]; Taufik Asra[2]; Bakhtiar Rifai [3]

[1]Program Studi Sistem Informasi
STMIK Nusa Mandiri Sukabumi
http://nusamandiri.ac.id/
aji.aby@nusamandiri.ac.id

[2]AMIK Bina Sarana Informatika
http://bsi.ac.id
taufik.tas@bsi.ac.id

[3]Program Studi Teknik Informatika
STMIK Nusa Mandiri Jakarta
http://nusamandiri.ac.id/
bakhtiar.bri@nusamandiri.ac.id

Abstract— *the Internet is very okay for right now, the Internet cannot be separated from its use of email, one of the threats that occur when using email is spam, spam is a message or email is unwanted by the recipient and sent in the masses.. Research on spam attacks are derived from the dataset as much spam 4601 records comprising 1813 records are considered spam and not spam data 278 with initial attributes as much as 57 attribute with attribute class 1, wants done on using select attribute with the decision tree becomes 15 attribute with attribute class 1 conducted 3 experiments testing with percentage attribute 30%, 50% and 70% select attribute obtained the results of the select attribute features of 70% better results were obtained from 30% or 50% with accuracy values of 92,469%.*

Keywords: Classification, Spam, Decision Tree.

Intisari—internet sangat biasa untuk sekarang ini, penggunaaan internetnya tak lepas dari penggunaan email, salah satu ancaman yang terjadi ketika menggunakan email adalah spam, spam merupakan pesan atau email yang tidak diinginkan oleh penerimanya dan dikirimkan secara massa.

Penelitian tentang serangan spam didapat dari dataset spam sebanyak 4601 record yang terdiri 1813 record dianggap spam dan 278 data bukan spam dengan atribut awal sebanyak 57 atribute dengan 1 atribute class, pada ekperimen yang dilakukan menggunakan select attribute dengan decision tree menjadi 15 atribute dengan 1 atribute class dilakukan 3 percobaan pengujian dengan persentase atribute 30%, 50% dan 70% select atribute didapat hasil fitur select atribute sebesar 70% didapat hasil lebih baik dari 30% ataupun 50% dengan nilai accuracy sebesar 92.469%.

Kata Kunci: *Classification, Spam, Decision Tree.*

PENDAHULUAN

Tidak diragukan lagi, pertumbuhan teknologi di masyarakat ini sangatlah pesat. Terlebih dalam penggunaan surat elektornik atau yang lebih sering disebut dengan nama email. Masyarakat dalam mengirim pesan lebih banyak memanfaatkan fasilitas email ini, karena lebih cepat sampai dan data pesan bisa diarsipkan dengan lengkap keterangan waktu, nama penerima nya dengan rapih. Namun dari keunggulan yang dimiliki oleh email ini dan juga sebagian orang yang memanfaatkannya untuk hal yang tidak baik, seperti mengirimkan pesan yang tidak kita inginkan.

Salah satu hal negatif dari email adalah spam, spam merupakan pesan atau email yang "tidak diinginkan" oleh penerimanya dan dikirimkan secara massal (Adisantoso & Rahman, 2009). Melakukan pengiriman email secara

massal atau berindikasi spam adalah sebuah pelanggaran terhadap *Acceptable Use Policy* (AUP) atau peraturan penggunaan yang bisa diterima pada hampir semua *Internet Service Provider* (ISP), dan dapat menghapus account pengirim tersebut (Hayuningtyas, 2017).

Dari kasus spam ini menimbulkan berbagai masalah di kalangan banyak pengguna email yang merasa terganggu dengan adanya spam (Hayuningtyas, 2017). Banyak waktu efektif yang terbuang untuk menghapus email spam dari inbox email, spam juga bisa menyebabkan pemborosan *bandwidth* karena Bandwidth merupakan daya tampung atau kapasitas dari sebuah Ethernet atau wireless agar dapat dilewati oleh paket data yang dilalui (Rifai, 2017) dan uang bagi user yang mengakses email dengan koneksi dial-up (Chandra, Indrawan, & Sukajaya, 2016)

Dalam mengatasi masalah yang terjadi maka diperlukan suatu filter untuk mengetahui email yang diterima itu terindikasi spam atau tidak. Banyak algoritma klasifikasi yang bisa digunakan untuk memfilter email.

BAHAN DAN METODE

Email adalah salah satu cara yang efektif dan juga efisien untuk melakukan komunikasi surat menyurat dengan Email mempunyai 3 jenis komponen (Hayuningtyas, 2017) antara lain Envelope Digunakan oleh *Mail Transport Agent* (MTA) untuk melihat rute atau jalur pesan. Header Sebagai informasi mengenai e-mail tersebut, mulai dari alamat pengirim, penerima, subjek dan lain-lain. Body Merupakan isi pesan dari pengirim ke penerima. Dalam mail body juga terdapat file attachment yang digunakan untuk mengirimkan e-mail berupa file (attachment). Spam Email Merupakan pesan email yang tidak kita inginkan ataupun yang kita minta (Hayuningtyas, 2017). Ada berbagai tipe email spam. Email mengandung unsur iklan promosi. email mengandung virus dan malware email terdekteksi phising email mengandung sifat penipuan atau scam email dengan pesan yang kurang berarti, seperti potongan pesan yang memenuhi inbox.

Data mining merupakan sebuah proses untuk mencapai hubungan, pola dan trend baru yang bertujuan untuk menyaring data yang sangat besar (Purnia & Warnilah, 2017).

Data mining merupakan teknik menganalisa data (untuk data berskala besar), sehingga menemukan hubungan yang jelas serta menyimpulkan yang belum diketahui sebelumnya dengan cara terkini sehingga mudah dipahami dan berguna untuk si pemilik data (Mulyadi, 2016).

Beberapa proses tahapan data mining agar tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantara. diantaranya sebagai Berikut (Hayuningtyas, 2017) Data Cleaning, membersihkan dari noise

Data Integration, gabungan dari berbagai database menjadi satu database. Data Selection, pengambilan data yang sesuai dengan kebutuhan

Data Transformation, pengubahan atau penggabungan format data yang sesuai untuk diproses dalam data mining.

Proses Mining, proses penerapan suatu metode Pattern Evaluation, untuk mengetahui pola-pola yang terdapat pada knowladge based yang dihasilkan. Knowledge Presentation, bentuk visualisasi dalam pengetahuan mengenai metode yang digunakan untuk memperoleh hasil ke pengguna.

A. Algoritma Decision Tree C4.5

Algoritma ini sudah sangat terkenal dan di sukai karena memiliki banyak kelebihan. Kelebihan ini misalnya dapat mengolah data numerik dan diskret, dapat menangani nilai atribut yang hilang satu yang tercepat dibandingkan dengan algoritma lain (Harryanto & Hansun, 2017). Ide dasar dari algoritma ini adalah pembuatan pohon keputusan berdasarkan pemilihan atribut yang memiliki prioritas tertinggi atau dapat memiliki gain tertinggi berdasarkan nilai entrophy atribut tersebut sebagai poros atribut (Harryanto & Hansun, 2017). Kemudian secara rekursif cabag-cabang ohon diperluas sehingga seluruh pohon terbentuk. entropy adalah jumlah data yang tidak relevan terhadap informasi dari suatu kumpulan data. (Harryanto & Hansun, 2017) Gain adalah informasi yang didapatkan dari perubahan entropy pada suatu kumpulan data, baik melalui observasi atau bisa juga disimpulkan dengan cara melakukan partisipasi terhadap suatu setdata. Menurut (Harryanto & Hansun, 2017) terdapat empat langkah dalam proses pembuatan pohon keputusan pada algoritma C4.5, yaitu:

1. Memilih atribut sebagai akar.
2. Membuat cabang untuk masing-masing nilai.
3. Membagi setiap kasus dalam cabang.
4. Mengulangi proses dalam setiap cabang sehingga semua kasus dalam cabang memiliki kelas yang sama.

Menurut (Harryanto & Hansun, 2017) data yang dimiliki harus disusun menjadi sebuah tabel berdasarkan kasus dan jumlah responden sebelum dilakukan perhitungan untuk mencari nilai entropy dan gain.

Rumus perhitungan entropy yang digunakan untuk menentukan seberapa informatif atribut tersebut.

$$Entropy\ (S) = \sum_{i=0}^{n} -pi * log^2 pi \quad \text{.....(1)}$$

Keterangan : S : Himpunan kasus
n : Jumlah partisi S
pi : Jumlah kasus pada partisi ke-i

Rumus yang digunakan dalam perhitungan gain setelah melakukan perhitungan entropy.

$$Gain\ (S, A) = Entropy\ (S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \qquad ....(2)$$

Keterangan : S : Himpunan kasus
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke – i
|S| : Jumlah kasus dalam S

Dengan mengetahui rumus-rumus diatas, data yang telah di peroleh dapat dimasukkan dan di proses dengan algoritma C4.5 untuk proses pembuatan decision tree.

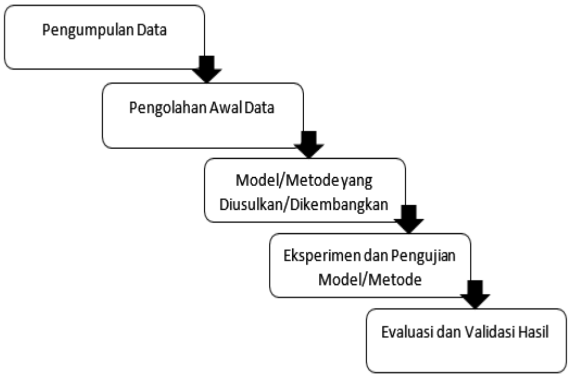
```
FormTree(T)
      (1) ComputeClassFrequency(T);
      (2) if OneClass or FewCases
               Return a leaf;
          Create a decision node N;
      (3) ForEach Attribute A
               ComputeGain(A);
      (4)  N.test = AttributeWithBestGain;
      (5) if N.test is Continuous
               Find Threshold;
      (6) ForEach T1 in the Splitting of T
      (7)      if T1 is Empty
                     Child of N is a leaf
               Else
                     Child of N = FormTree(T1);
      (8) ComputeErrors of N;
          Return N;
```

Sumber: Harryanto dan Hansun, (2017)
Gambar 1. Psedcode Algoritma C4.5.

B. Desain Penelitian

Penelitian ini dilakukan dengan data yang memiliki hasil spam dan bukan spam, data yang digunakan diambil dari UCI (Universitas of California, Irvine) Machine Learning Repository (George Forman,1999). Pada penelitian prediksi spam ini, akan diolah dengan algoritma decision tree. Sebelum dilakukan prediksi dengan algoritma decision tree (C4.5) data lebih dulu dilakukan penyeleksian atribut.

Dalam penelitian ini juga dilakukan beberapa langkah yang dilakukan dalam proses penelitian:



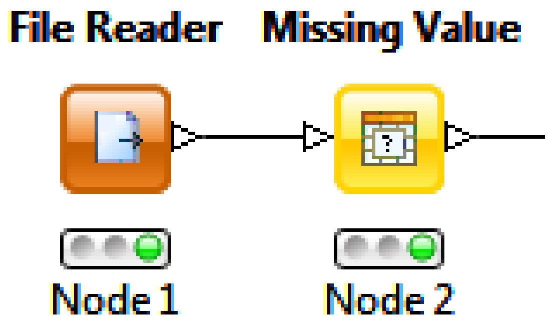
Sumber: (Sudibyo et al., 2018)
Gambar 2. Tahapan Penelitian

C. Pengumpulan Data

Dalam penelitian ini pengumpulan data adalah data masalah spam yang dianggap sebagai salah satu kejahatan yang terjadi pada dunia sistem informasi saat ini. Jumlah dataset ini memiliki total 4601 record berbeda, diantaranya 1813 data dianggap spam dan 2788 data bukan spam. Dan jumlah atribut data awal pada data ini memiliki 57 atribute di tambah 1 atribute class.

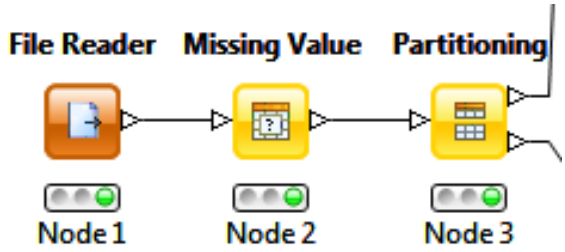
D. Pengolahan Awal Data

Data yang diperoleh untuk penelitian sebanyak 4601 record, data yang ada dengan format .csv. Proses pengolahan data awal ialah melakukan model missing value yang berfungsi untuk mengecek data, apakah masih mengandung duplikasi atau inkosisten data.



Sumber: (Sudibyo et al., 2018)
Gambar 3. Model Desain *Missing Value*

Setelah data selesai dilakukan proses missing value, tahapan selanjutnya yaitu data di proses dalam parttioning, pada parttioning ini tabel kelauran akan di bagi menjadi dua, untuk melatih dan menguji data.

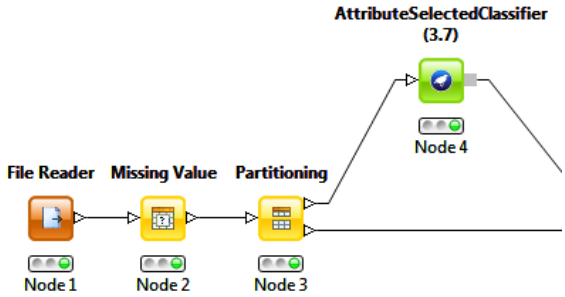


Sumber: (Sudibyo et al., 2018)

Gambar 4. Model Desain *Parttioning*

E. Metode Yang Diusulkan

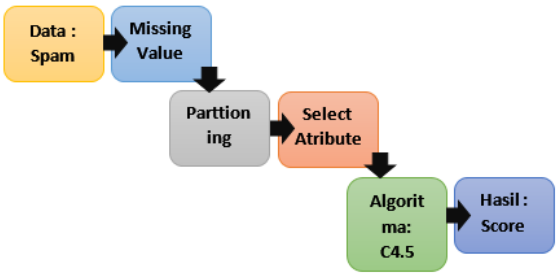
Pada tahap modeling ini dilakukan pemrosesan data training sehingga akan membahas metode algoritma yang diujikan dengan memasukkan data spam kemudian dilakukan analisa, tapi sebelum melakukan prediksi data dilakukan dulu proses penyeleksian *atribute.*



Sumber: (Sudibyo et al., 2018)

Gambar 5. Model desain Penyeleksian *Atibute*

Berikut ini bentuk gambaran metode algoritma yang akan diuji:



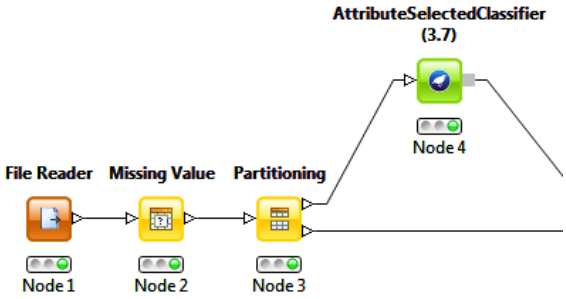
Sumber: (Sudibyo et al., 2018)

Gambar 6. Model Yang Diusulkan

HASIL DAN PEMBAHASAN

A. Eksperimen dan Pengujian Model

Pada tahapan awal penelitian ini melakukan penyeleksian atribut dengan menggunakan *attribute selected classifier* dengan *algorithm decision tree*. Data awal yang ada memiliki sebanyak 57 *attribute* di tambah 1 atribute class dan setelah dilakukan *attribute selected classifier* dengan *algorithm decision tree* menjadi 15 *attribute* di tambah 1 *attribute class*.



Sumber: (Sudibyo et al., 2018)

Gambar 7. Model *attribute selected classifier* dengan *algorithm decision tree*

Untuk table yang telah didapat dari 15 atribute yang dipilih setelah melakukan *attribute selected classifier* dengan *algorithm decision tree* adalah sebagai Berikut:

Tabel 1. Hasil *attribute selected classifier*

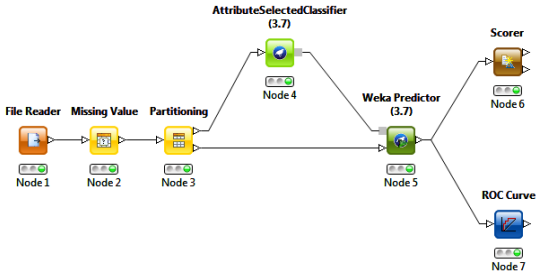
No	Nama Atribute
1	Word freq our numeric
2	Word freq remove numeric
3	Word freq will numeric
4	Word freq free numeric
5	Word freq you numeric
6	Word freq credit numeric
7	Word freq your numeric
8	Word freq font numeric
9	Word freq hp numeric
10	Word freq 1999 numeric
11	Word freq edu numeric
12	Char freq !
13	Char freq $
14	Capital run length average numeric
15	Apital run length longes numeric
16	Class {0,1}

Sumber: (Sudibyo et al., 2018)

B. Evaluasi Dan Validasi Hasil

Setelah melakukan *attribute selected classifier* dengan *algorithm decision tree* hasil dari *attribute* itu akan di lakukan evaluasi berdasarkan tingkat akurasinya untuk melihat dari kinerja metode yang kita gunakan.

Penelitian ini selain untuk melakukan seleksi atribute atau menyederhanakan *attribute* juga melihat akurasi dari prediksi spam, nilai apakah dari kriteria file atau pesan itu spam atau bukan spam.



Sumber: (Sudibyo et al., 2018)

Gambar 8. Model Prediksi Spam atau Bukan Spam pada attribute selected classifier dengan algorithm decision tree

Setelah melakukan pemodelan seperti gambar 8 dimana pada node 5 menggunakan weka predictor dimana hasil keluaran score berupa akurasi dari data spam dan ROC Curve berupa hasil dari ROC class 0 dan ROC class 1. Dan untuk hasil dari pengukuran yang dilakukan pada data data spam adalah seperti berikut ini.

Pengujian dengan melakukan 30% attribute selected classifier dengan algorithm decision tree pada dataset Spam didapat correct classified 2.937 dengan wrong classified 284 dengan nilai accuracy 91,183%

Class \ Prediction (Class)	1	0
1	1097	167
0	117	1840

Correct classified: 2.937 Wrong classified: 284

Accuracy: 91,183 % Error: 8,817 %

Cohen's kappa (κ) 0,814

Sumber: (Sudibyo et al., 2018)

Gambar 9. Nilai Akurasi 30% attribute selected classifier

Sedangkan untuk pengujian dengan melakukan membagi data testing sebesar 50% didapat hasil correct classified 2.119 dengan wrong classified 182 serta untuk nilai accuracy sebesar 92.09%

Class \ Prediction (Class)	1	0
1	801	100
0	82	1318

Correct classified: 2.119 Wrong classified: 182

Accuracy: 92,09 % Error: 7,91 %

Cohen's kappa (κ) 0,833

Sumber: (Sudibyo et al., 2018)

Gambar 10. Nilai Akurasi 50% attribute selected classifier

Untuk pengujian dengan nilai data testing 70% didapat hasil sebagai berikut:

Class \ Prediction (Class)	1	0
1	474	51
0	53	803

Correct classified: 1.277 Wrong classified: 104

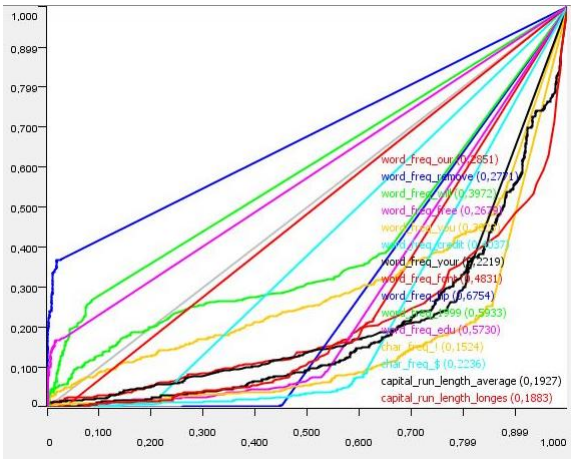
Accuracy: 92,469 % Error: 7,531 %

Cohen's kappa (κ) 0,84

Sumber: (Sudibyo et al., 2018)

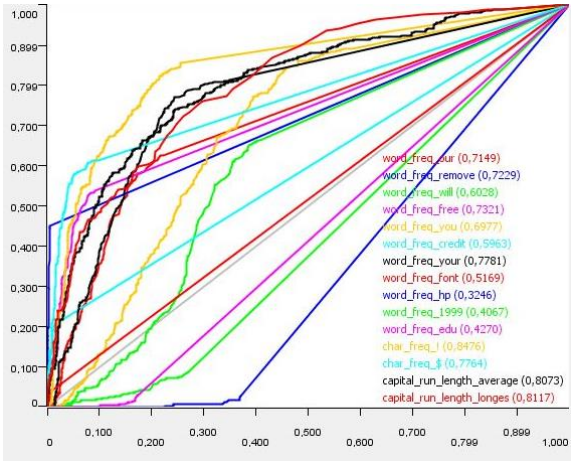
Gambar 11. Nilai Akurasi 70% attribute selected classifier

Berdasarkan pengukuran yang dilakukan didapat hasil untuk nilai akurasi yang didapat sebesar 92,469 %, dengan *correct classified* 1277 dengan 104 *wrong classified* dengan nilai error 7,531 %, *cohen's kappa* 0,84



Sumber: (Sudibyo et al., 2018)

Gambar 12. ROC Class Value 0



Sumber: (Sudibyo et al., 2018)

Gambar 13. ROC Class Value 1

C. Analisis Hasil Penelitian

Dari penelitian yang telah dilakukan dengan beberapa kali pengujian dengan jumlah

persentase data yang berbeda-beda dapat dirangkum pada tabel 2 berikut ini:

Tabel 2. Hasil Percobaan

Relative	Akurasi	Cohen's kappa (K)	Error
30 %	91,183 %	0,814	8,817 %
50 %	92,09 %	0,833	7,91%
70 %	92,469 %	0,84	7,531 %
100 %	83,67 %	0,652	16,33 %

Sumber: (Sudibyo et al., 2018)

Hasil dari perhitungan pada tabel 2. dengan penerapan klasifikasi performance keakurasian AUC maka dapat diklasifikasikan menjadi lima kelompok, antara lain:

1. 0.50 – 0.60 = klasifikasi salah
2. 0.60 – 0.70 = klasifikasi buruk
3. 0.70 – 0.80 = klasifikasi cukup
4. 0.80 – 0.90 = klasifikasi baik
5. 0.90 – 1.00 = klasifikasi sangat baik

Berdasarkan pengelompokan pada tabel 2. Dengan membandingkan akurasi terlihat bahwa penelitian terbaik berada pada saat data relative 70% dengan nilai akurasi sebesar 92,469 %. Dan melihat dari keakurasian AUC/Kappa mendapatkan nilai 0.80-0.90 yang menandakan memiliki klasifikasi baik.

KESIMPULAN

Pada penelitian ini dimana dataset spam yang diperoleh sebanyak 4601 record yang terdiri dari 1813 record dianggap spam dan 278 data bukan spam dengan atribut awal sebanyak 57 atribute dengan 1 atribute class, pada ekperimen yang dilakukan menggunakan select atribute menjadi 15 atribute dengan 1 atribute class dilakukan 3 percobaan pengujian dengan persentase 30%, 50% dan 70% attribute selected classifier didapat hasil untk percobaan dengan Select atribute 30% didapat accuracy sebesar 91.183%, sedangkan untuk percobaan 50% atribute didapat 92.09% dan untuk 70% didapat accuracy sebanyak 92.469% maka dapat disimpulkan pengujian dengan fitur select atribute sebesar 70% didapat hasil lebih baik dari 30% ataupun 50% dengan nilai accuracy sebesar 92.469.

REFERENSI

Adisantoso, J., & Rahman, W. (2009). Pengukuran Kinerja Spam Filter Menggunakan Graham's Naive Bayes Classifier. *Jurnal Ilmu Komputer Agri-Informatika*, *2*(Spamhaus), 1–8. Retrieved from http://journal.ipb.ac.id/index.php.jika

Chandra, W. N., Indrawan, G., & Sukajaya, I. N. (2016). Spam Filtering Dengan Metode Pos Tagger Dan Klasifikasi Naïve Bayes. *Jurnal Ilmiah Teknologi Informasi Asia (JITIKA)*, *10*(1), 47–55.

Harryanto, F. F., & Hansun, S. (2017). Penerapan Algoritma C4 . 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE. *Jatisi*, *3*(2), 95–103.

Hayuningtyas, R. Y. (2017). Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes. *IJCIT (Indonesia Journal on Computer and Information Technology*, *2*(1), 53–60.

Mulyadi. (2016). Penerapan Algoritma Naive Bayes Untuk Klasifikasi Penerima Beasiswa Prestasi. *JURNAL SISTEM INFORMASI STMIK ANTAR BANGSA*, *V*(2), 139–145.

Purnia, D. S., & Warnilah, A. I. (2017). Implementasi Data Mining Menggunakan Algoritma Apriori. *Prosiding SINTAK 2017*, *2*(2), 31–39.

Rifai, B. (2017). Management Bandwidth Pada Dynamic Queue Menggunakan Metode Per Connection Queuing. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *2*(2), 73–79. Retrieved from http://ejournal.nusamandiri.ac.id/ejurnal/index.php/jitk/article/view/246

Sudibyo, A., Asra, T., & Rifai, B. (2018). *Laporan Hasil Penelitian*.