Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 3 No. 2 (2019) 227 - 232 ISSN Media Elektronik: 2580-0760

Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting

Bobby Suryo Prakoso[1], Didi Rosiyadi [2] ,Heru Sukma Utama[3],Dedi Aridarma[4]
[1]Magister Ilmu Komputer, Fakultas Ilmu Komputer, STMIK Nusa Mandiri Kramat
[2]Fakultas Ilmu Komputer, STMIK Nusa Mandiri Kramat
[2] Fakultas Teknik Informasi, Universitas Bina Sarana Informatika
[2] Pusat Penelitian Informatika LIPI
[3,4]Magister Ilmu Komputer, Fakultas Ilmu Komputer, STMIK Nusa Mandiri Kramat
[1]14002107@nusamandiri.ac.id

Abstract

This research is part of text mining for the classification of news content that already has labels based on the category of news on the site detik.com. The process carried out is to do data modeling and processing, start the pre-processing process, information gain feature selection process, and the application of the Naive Bayes Classifier algorithm model with Bayesian Boosting. The results obtained from the model get an evaluation value of accuracy, recall, and precision of 73.2%. Whereas the more concise model is the Naive Bayes Classifier algorithm model, with Bayesian Boosting getting the same evaluation value of 73.2%. The evaluation of the results of the evaluation model that has been implemented concludes that the application of the Information Gain feature does not have a large effect on the increase in performance results on the condition of the Polynomial label..

Keywords: *Information Gain, Naive Bayes Classifier, Boosting, Bayesian Boosting*

Abstrak

Penelitian yang dilakukan ini merupakan bagian dari *text mining* untuk klasifikasi konten berita yang telah memiliki label berdasarkan katagori berita pada situs detik.com . Proses yang dilakukan adalah melakukan permodelan dan pengolahan data, mulai proses *pre-processing*, proses seleksi fitur *information gain*, dan penerapan model algoritma *Naive Bayes Classifier* dengan *Bayesian Boosting*. Hasil yang diperoleh atas model tersebut mendapatkan nilai evaluasi terhadap akurasi, *recall*, dan presisi sebesar 73.2%. Sedangkan dengan model yang lebih ringkas yaitu model algoritma *Naive Bayes Classifier*, dengan *Bayesian Boosting* mendapatkan nilai evaluasi yang sama besar yaitu 73.2%. Penilaian atas hasil evaluasi model yang telah terlaksankan berkesimpulan bahwa penerapan seleksi fitur *Information Gain* tidak berpengaruh besar atas kenaikan hasil performa terhadap kondisi label *Polynomial*.

Kata kunci: *Information Gain, Naive Bayes Classifier, Boosting, Bayesian Boosting*

1. Pendahuluan

Klasifikasi teks banyak mengenalnya sebagai bagian mendefinisikan satu atau lebih katagori untuk dokumen berbahasa natural. Jika pada dasarnya klasifikasi dokumen secara manual atau dengan aturan klasifikasi otomatis yang terumuskan oleh manusia, banyak algoritma machine learning digunakan untuk mengklasifikasikan teks ataupun berita secara otomatis yang tidak terlihat berdasarkan data training atau data testing yang telah diberi label oleh manusia.

Mengingat semakin banyaknya file dokumen online yang tersedia melalui internet diantaranya berita online, email dan perpustakaan digital, tugas ini sangat penting secara praktis dapat mengklasifikasikan lebih terstruktur.

Pemanfaatan internet saat ini yang menarik adalah untuk bidang sosial – politik dengan presentase penggunaan sebagai berikut berita sosial lingkungan sebesar 50,26 %, berita informasi agama 41,55%, berita politik 39,94% dan kegiatan amal sebesar 16,31%.

Diterima Redaksi : 22-07-2019 | Selesai Revisi : 29-07-2019 | Diterbitkan Online : 03-08-2019

Berdasarkan data teserbut bahwa pemberitaan online pun menjadi media yang digunakan untuk memperoleh informasi tersebut[1].

Penelitian yang dilakukan oleh Shuo Xu, yang mengklasifikasikan berita menggunakan algoritma *Naive Bayes Classifier* dengan menggabungkan metode *Gausian Event Model* dan *Multinomial Event Model*. Terhadap 20 katagori berita yang digunakan, dengan masul untuk gausian *Gausian Event Model* memiliki hasil yang lebih baik setelah diperbandingkan, dengan hasil akurasi 88%[2].

Selanjutnya tentang analisis sentimen dengan menggunakan algoritma *Naive Bayes Classifier* dan ditambahkan fitur seleksi *Chi Square* untuk menentukan rekomendasi atas lokasi restoran dengan tema tradisional. Berdasarkan perhitungan yang dilakukan untuk penelitian tersebut terdapat beberapa skenario pengujuan yang diantaranya untuk nilai *Chi Square* 25% mendapatkan akurasi 81%, *Chi Square* 50% mendapatkan 80%, *Chi Square* 75% mendapatkan akurasi 77%, dan yang terkahir *Chi Square* 100% mendapatkan akurasi 80%. Dengan kesimpulan untuk penggunaan *Chi Square* tidak terlalu berpengaruh[3].

Kemudian tentang penelitian yang dilakukan oleh Dio Ariadi dan Kartika Fithriasari, yang melakukan klasifikasi berita dengan menggunakan metode *Naive Bayes Classifier* dengan menambahkan *confix stripping stemmer*. Detail tentang penelitian ialah mengunakan 12 katagori berita dengan menggunakan 100 berita setiap masing masing katagori. Hasil akhir atas penelitian tersebut untuk algoritma *Naive Bayes Classifier* performa akurasi, presisi, dan *recall* sebesar 82,2% 83,9% dan 82,2%. Sedangkan untuk algoritma *Support Vector Machine* akurasi, presisi, dan *recall* adalah 88,1%, 89,1%, dan 88,1%[4].

Berikutnya masih membahas untuk klasifikasi teks yang berbasis analisa sentimen terhadap pariwisata pada kota malang, dengan data yang diambil dari situs tripadvisor.com pada bagian komentar yang diberikan pada tempat wisata. Metode yang digunakan adalah algoritma *Naive Bayes Classifier* dengan seleksi fitur *Query Expansion Ranking* untuk mengurangi jumlah fitur pada proses klasifikasi, yang mendapatkan hasil akurasi sebesar 86.6%.[5]

Berdasarkan rujukan penelitian yang ada, tentang kombinasi penggunaan seleksi pembobotan seleksi fitur dan boosting untuk model perhitungan dalam klasifikasi pengolahan teks. Baik dari penggunaan

Atas dasar landasan yang diberikan maka mucul beberapa pertanyaan, diantaranya adalah bagaimana mendapatkan data, proses *pre-processing* data, seleksi fitur yang akan digunakan, dan terakhir metode algoritma yang digunakan. Hal inilah yang menjadi tujuan dari penelitian ini.

2. Metode Penelitian

Berdasarkan penjelasan yang telah diutarakan pada bagian pendahuluan, berikut dijelaskan alur prosesnya.

2.1. Pengambilan Dataset

Dalam hal pengambilan data, pada penelitian ini menggunakan *tool* telah disediakan oleh *google*, dengan menggunakan *google spreadsheet*[6]. Dalam hal ini pengambilan menggunakan suatu formula yang dapat mengambil data baik JSON, nilai dalam XPATH, nilai XML yang ada pada suatu web.

Berikut untuk contoh untuk formula dalam pengambilan data diantaranya ada untuk pengambilan URL :

Formula Google Spreadsheet (Pengambilan URL)

```
=IMPORTXML("https://news.detik.com/berita/d-4547868/jadi-aktor-mutilasi-di-mana-prada-deri-bersembunyi","//a/@href")
```

Formula tersebut sudah tersedia pada menu help yang telah disediakan pada google spreadsheet.

2.2. Teks *Pre-Processing*

Dalam hal teks *pre-processing* merupakan suatu cara atau proses didalam sebelumnya adalah bentuk tidak terstruktur menjadi terstruktur, seperti contoh (merubah teks menjadi nominal *term index*). Tujuannya adalah untuk memperkecil dimensi data sehingga proses komputasi lebih menjadi efisien dan diharapkan lebih presisi.

Preprocessing terdiri dari beberapa tahapan. Adapun tahapan preprocessing berdasarkan, yaitu : *case folding, tokenizing/ parsing, filtering* dan *stemming* [7].

Dalam proses pertama ada *case folding*, yang merupakan suatu proses merubah teks menjadi seragam, bisa dalam bentuk huruf kecil semua, atau menjadi huruf besar semua. Selanjutnya adalah proses *tokenizing*, yaitu proses isi teks menjadi beberapa satuan kata.

Filtering merupakan tahap proses setelah proses *tokenizing*, yaitu proses membuang kata-kata tidak memiliki makna, seperti contohnya adalah "yang", "apa", "dengan", dan lain sebagainya. Selanjutnya dalam proses filtering terdapat menyimpan kata yang dianggap penting atau bisa masuk dalam data yang digunakan sebagai perhitungan.

Stemming adalah proses untuk mengurangi kata-kata yang berimbuhan menjadi bentuk dasarnya, dan hasil dari *stemming* ini merupakan hal yang akan menjadi dasar akan dirubahnya dari teks menjadi nominal[8].

2.3. Seleksi Fitur *Information Gain*

Seleksi Fitur *Information Gain* merupakan suatu teknik dalam mengurangi jumlah fitur yang sesuai atau relevan, lalu mengurangi dimensi fitur pada data yang akan digunakan. Untuk menghitung *infomation gain* menggunakan hitungan sebagai berikut[9] :

$$info(D) = -\sum_{i=1}^{c} p_i log_2(p_i) \quad (1)$$

Keterangan dari rumus tersebut adalah:

c : jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi)
p_i : jumlah sampe untuk kelas i

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ info(D_j) \quad (2)$$

Keterangan dari rumus tersebut adalah:

A : atribut
$|D|$: jumlah seluruh sampel data
$|D_j|$: jumlah sampel untuk nilai j
v : suatu nilai yang mungkin untuk atribut A

Selanjutnya nilai *information gain* yang akan dipakai dengan dihitung menggunakan rumus dibawah ini[10] :

$$Gain\ (A) = |info(D) - info_A(D)| \quad (3)$$

2.4. Algoritma *Naive Bayes Classifier*

Naive Bayes adalah metode algoritma yang bekerja atas bagaimana menghitung frekuensi atas setiap term yang ada dalam dokumen [5]. Dokumen dengan urutan kejadian yang muncul atas kata terhadap dokumen akan diabaikan, menyebabkan pengolahan kata menggunakan distiribusi yang multinomial[11].

Berikut persamaan rumus atas Naive Bayes Classifier [12]:

$$P(c|d) = P(c) \prod_{i=1}^{n} P(w_i|c) \quad (4)$$

d : besaran dokumen
n : jumlah semua kata yang ada pada dokumen

Selanjutnya nilai atas variabel $P(c)$ diperoleh dengan rumus berikut

$$P(c) = \frac{N_c}{N} \quad (5)$$

$P(c)$: peluang kelas c
N : jumlah seluruh dokumen

Selanjutnya untuk menghitung peluang kata ke- i pada kelas c menggunakan rumus berikut :

$$P(w_i|c) = \frac{count\ (w_i,c)+1}{count\ (c)+|V|} \quad (6)$$

$P(w_i|c)$: Peluang kata ke- i pada kelas c
$count\ (w_i, c)$: Jumlah kata ke- i pada kelas c
$count\ (c)$: Jumlah semua kata pada kelas c
$|V|$: Jumlah kata unik terhadap semua Kelas

3. Hasil dan Pembahasan

Pada pembahasan kali ini untuk proses penelitian akan dilakukan secara bertahap, mulai dari proses pengambilan data, tahap *pre-processing*, tahap seleksi fitur, dan tahap proses validasi. Dengan menggunakan data berita dari situs berita online yang digunakan untuk pengambilan data. Tahapan proses yang dilakukan :

3.1. Pengambilan Dataset

Proses pengambilan data menggunakan tools yang telah disediakan oleh *google*, dengan *google spreadsheet*. Serta sumber data mengambil dari situs detik.com. Berikut merupakan proses ekstraksi link yang akan menjadi dataset untuk digunakan:

Formula Google Spreadsheet (ekstraksi URL)

```
=IMPORTXML("https://news.detik.com/berita/d-4547868/jadi-aktor-mutilasi-di-mana-prada-deri-bersembunyi","//a/@href")
```

Dalam hal ini untuk labeling berdasrkan link dari hasil ekstraksi yang sebelumnya dilakukan pencarian pada situs detik.com, dengan kata kunci pencarian Pendidikan, Politik, dan Budaya. Selanjutnya adalah ekstraksi proses isi berita dengan menggunakan formula berikut :

Formula Google Spreadsheet (ekstraksi konten berita)

```
=IMPORTXML("https://news.detik.com/berita/d-4547868/jadi-aktor-mutilasi-di-mana-prada-deri-bersembunyi",
"//*[@id='detikdetailtext']"")
```

Setelah data terkumpul maka, seluruh data dikumpulkan pada suatu file untuk menjadi kesatuan dengan contoh format berikut, Taabel 1.

Tabel 1.Contoh Susunan Dataset Terkumpul

TEMA_BERITA	KONTEN_BERITA
PENDIDIKAN	Jakarta - Pernah terbayang ada .....
PENDIDIKAN	Jakarta - Tujuh hari jelang .....
POLITIK	Jakarta - menyebut pertemuan .....
POLITIK	Jakarta – Setan Gundul yang di.....
LINGKUNGAN	Jakarta-Terhadap penelitian sehingga..
LINGKUNGAN	Jakarta- Pengajaran lingkungan ....

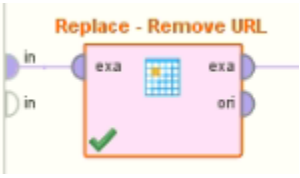
3.2. Proses *Pre-Processing*

Setelah dataset terkumpul, maka selanjutnya adalah proses untuk memulai pengolahan data, yaitu proses *pre-processing*. Tahapan pertama adalah proses menghilangkan link yang masih ada pada dataset pada konten berita tersebut diantaranya dengan menggunakan operator yang telah tersedia pada *tools Rapidminer*, yaitu dengan operator *Replace* dengan

menggunakan *Regex*, berikut untuk regex yang terpakai dan gambar operator yang terpakai pada Rapidminer beserta:

```
Contoh Regex untuk menghilangkan link URL

  "http\S+|\S+co\S+"
```



Gambar 1.Operator *Replace – Remove URL*

Atas proses operator tersebut maka akan menghasilkan contoh pada tabel 2 berikut ini :

Tabel 2.Contoh *Replace URL*

No	Konten Awal	Konten Akhir
1	Dengan siapa **dia.com**	Dengan siapa
2	Siapa sih **temannya.com**	Siapa sih
3	Aku telah melihat situs https:\\menang.com	Aku telah melihat situs

Setelah melalui proses tersebut maka berlanjut pada proses berikut ini :

Gambar 2.Operator *Replace – Remove Special Character*

Yang menghasilkan contoh proses sebagai berikut, Tabel 3.

Tabel 3.Contoh *Replace – Remove Special Character*

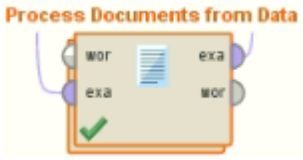
No	Konten Awal	Konten Akhir
1	Dengan siapa **?!!!!**	Dengan siapa
2	Siapa sih **^&*(@(#*!^**	Siapa sih
3	Aku telah melihat situs)*)@*#)!	Aku telah melihat situs

Proses selanjutnya adalah menentukan label yang digunakan dan merubah nilai yang nominal menjadi dalam bentuk text sebelum masuk proses selanjutnya :

Gambar 3.Operator *Set Role*

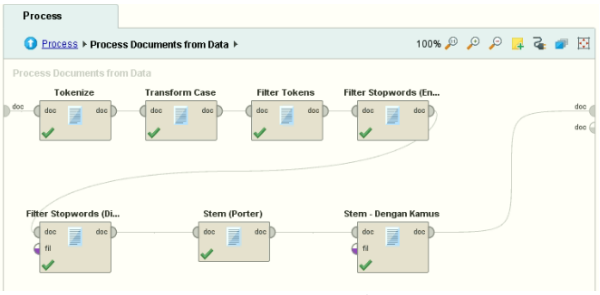
Gambar 4.Operator *Nominal to Text*

Jika data sudah sesuai maka data akan diteruskan kepada operator Process Documents from Data :



Gambar 5.Operator *Process Documents from Data*

Operator tersebut terdapat proses tahapan diantaranya mulai dari *Tokenize, Transform Case, Filter Tokens, Filter Stopwords,* dan *Stemming*. Berikut untuk gambar detail operatornya :



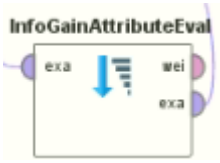
Gambar 6.Operator *Process Documents from Data* Secara Lengkap

3.3. Proses Seleksi Fitur *Information Gain*

Setelah proses *pre-processing* atas dataset yang ada, maka selanjutnya diteruskan kepada proses menghilangkan attribut yang yang tidak terpakai diantaranya yaitu menggunakan operator *Remove Useless Attributes*, berikut operator yang tersedia:

Gambar 7.Operator *Remove Useless Attributes*

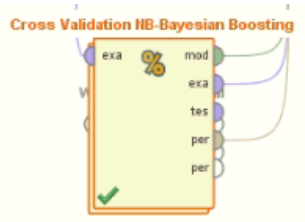
Proses selanjutnya adalah pembobotan dengan menggunakan operator Information Gain dengan menggunakan operator yang telah tersedia pada Rapidminer, berikut untuk gambar operator yang digunakan :



Gambar 8. Operator *Information Gain*

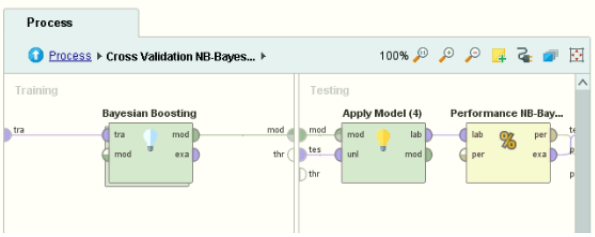
3.4. Proses Validasi Algoritma

Pada proses ini menggunakan beberapa operator, sebelumnya menggunakan operator *cross validation* dengan *k-10 fold cross validation*. Berikut untuk operator yang digunakan :



Gambar 9. Operator *Cross Validation K-Fold*

Yang didalamnya terdapat berbagai operasi diantaranya sebagai berikut :



Gambar 10. Kumpulan *Operator Cross Validation*

Dengan proses pertama yaitu proses *bayesian boosting* untuk memaksimalkan hasil dari perhitungan algoritma *Naive Bayes Classifier*, berikut untuk gambar operatornya :

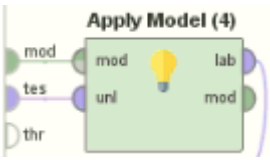
Gambar 11. Operator Bayesian Boosting

Pada *Bayesian Boosting*, proses sederhana yang dapat dijelaskan adalah melakukan boost untuk pengulangan atau iterasi atas penggunaan algoritma Naive Bayes Classifier untuk mendapat hasil yang lebih baik. Dengam melakukan iterasi *default* 10, tetapi pada penelitian ini menggunakan iterasi 30.

Pada Bayesian Boosting didalamnya terdapat Naive Bayes Classifier, berikut untuk contoh operatornya :

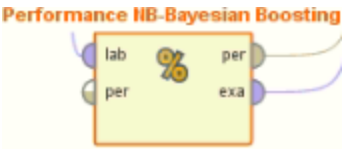
Gambar 12. Operator *Naive Bayes*

Setelah itu dilakukan penerapan model dengan operator *Apply Model* berikut ini :



Gambar 11. Operator *Apply Model*

Setelah itu menghitung performa yang muncul terhadap penerapan model dengan metode *k-10 fold*, yaitu dengan memecah 10 x 10 dataset yang dijadikan data *training* dan terhadap 1 bagian yang akan digunakan menjadi data *testing*.



Gambar 11. Operator *Performance – NB-Bayesian Boosting*

3.5. Hasil Penelitian

Atas percobaan tersebut, penelitian ini membuat beberapa skenario yang digunakan dengan 4 model, berikut hasil yang didapat, Tabel 3.

Tabel 3. Hasil *Cross Validation K-10*

Model	Akurasi	Presisi	*Recall*
NBC	72%	72.3%	72.3%
IG-NBC	69.5%	69.6%	69.6%
Bayesian Boosting-NBC	73.2%	73.2%	73.2%
IG-*Bayesian Boosting*-NBC	73.2%	73.2%	73.2%

4. Kesimpulan

Hasil penelitian menunjukan bahwa penerapan seleksi fitur information gain tidak memiliki pengaruh pada label yang bersifat *polynomial*. Selanjutnya dalam penerapan bayesian boosting untuk label yang bersifat polynomial memiliki pengaruh naiknya hasil evaluasi sebesar 4.3%. Berdasarkan hasil tersebut, saran terhadap penelitian selanjutnya dapat memilih seleksi fitur lain yang sesuai dengan karakteristik label dan data yang digunakan dalam pengolahan *text mining* dengan label yang bersifat *polynomial*.

Daftar Rujukan

[1] APJII, “Penetrasi dan perilaku pengguna internet Indonesia,” 2017.
[2] P. R. C. Xu, Shuo /(Research Center for Information Science

Theory and Methodology, Institute of Scientific and Technical Information of China, "Bayesian Naıve Bayes classifiers to text classification.pdf." .

[3] N. D. Pratama, Y. A. Sari, and P. P. Adikara, "Analisis Sentimen Pada Review Konsumen Menggunakan Metode Naive Bayes Dengan Seleksi Fitur Chi Square Untuk Rekomendasi Lokasi Makanan Tradisional," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 9, pp. 2982–2988, 2018.

[4] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.

[5] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[6] S. Moro *et al.*, "Leveraging national tourist offices through data analytics Leveraging national tourist of fi ces through data analytics," 2018.

[7] Y. Pramudita, U. T. Madura, S. S. Putro, and U. T. Madura, "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," *J. Teknol. Inf. dan Ilmu Komput.*, vol. Vol. 5, No, no. August 2018, p. hlm. 269-276, 2019.

[8] B. Aryoyudanta, T. B. Adji, and I. Hidayah, "Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm," *Proceeding - 2016 Int. Semin. Intell. Technol. Its Appl. ISITIA 2016 Recent Trends Intell. Comput. Technol. Sustain. Energy*, pp. 7–12, 2017.

[9] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, pp. 1880–1888, 2016.

[10] L. Dini Utami and R. S. Wahono, "Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[11] X. Feng, S. Li, C. Yuan, P. Zeng, and Y. Sun, "Prediction of Slope Stability using Naive Bayes Classifier," *KSCE J. Civ. Eng. 22(3)941-950, pISSN 1226-7988, eISSN 1976-3808*, vol. 22, pp. 941–950, 2018.

[12] A. S. Budiman, P. Studi, T. Komputer, X. A. Parandani, P. Studi, and M. Informatika, "Uji Akurasi Klasifikasi Dan Validasi Data Pada Penggunaan Metode Membership Function Dan Algoritma C4 . 5 Dalam," vol. 9, no. 1, pp. 565–578, 2018.