Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 3, Maret 2019, hlm. 2154-2159 http://j-ptiik.ub.ac.id

Prediksi Tingkat Pemahaman Siswa Dalam Materi Pelajaran Bahasa Indonesia Menggunakan *Naive Bayes* Dengan Seleksi Fitur *Information Gain*

Siti Utami Fhylayli[1], Budi Darma Setiawan[2], Sutrisno[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]utamifhylayli@student.ub.ac.id, [2]s.budidarma@ub.ac.id, [3]trisno@ub.ac.id

Abstrak

Mata pelajaran Bahasa Indonesia secara umum dianggap sebagai pelajaran yang mudah dan tidak perlu dipelajari secara serius oleh kebanyakan siswa dan masyarakat. Berdasarkan hal tersebut, muncul berbagai masalah pembelajaran yang melibatkan pengajar, mata pelajaran Bahasa Indonesia, siswa yang menerima pelajaran, metode mengajar, sarana-prasarana, cara mengevaluasi, serta tujuan pelajaran Bahasa Indonesia (Moeljono, 1989). Kemampuan setiap siswa dalam memahami pelajaran tersebut berbeda-beda. Hal ini menyebabkan pengajar memiliki keterbatasan untuk mengukur tingkat pemahaman siswa. Maka diperlukan sistem untuk untuk memprediksi tingkat pemahaman siswa. Prediksi ini menggunakan metode klasifikasi dengan algoritme *Naïve Bayes*. Kelas yang akan digunakan pada penelitian ini diantaranya siswa sangat paham, cukup paham dan kurang paham. Pada penelitian ini penulis menggunakan seleksi fitur *Information Gain* (IG). Fitur yang terpilih akan diproses dengan algoritme klasifikasi naive bayes kemudian dilihat akurasinya jika belum maksimal maka dilakukan kembali proses seleksi fitur tadi hingga mendapatkan akurasi yang diinginkan. Dari pengujian yang sudah dilakukan, didapatkan hasil bahwa fitur yang memiliki nilai *Gain* lebih dari 0.2 memiliki akurasi terbesar yaitu mencapai 90%. Fitur yang terpilih dari 17 fitur yaitu diantaranya fitur jumlah anggota keluarga, status tempat tinggal, pekerjaan ibu, pengasuh, dukungan keluarga, ikut ekstrakurikuler, mengulang pelajaran di rumah, lama belajar di rumah, jenis bacaan di rumah, lama membaca di rumah.

Kata kunci: *klasifikasi, naive bayes, seleksi fitur, information gain, pemahaman, bahasa Indonesia.*

Abstract

Indonesian Language Subjects are generally regarded as easy lessons and do not need to be studied by students and society. Based on this, various learning problems arose involving instructors, Indonesian language subjects, students who received lessons, teaching methods, facilities, ways to obtain, and the objectives of Indonesian language learning (Moeljono, 1989). The difference between each student in different learning differences. This causes the teacher to have limitations in measuring the level of understanding of students. Then a system is needed to predict the level of understanding of students. This prediction uses the classification method with the Naïve Bayes algorithm. The class that will be used in this study is that students understand, are quite understanding and lack understanding. In this study, the authors used the Information Gain (IG) feature selection. The selected feature will be processed with the Naive Bayes classification algorithm, then the accuracy will be seen if it is not maximized, then the previous feature selection process will be done again to get the desired verification. From the tests that have been conducted, the results obtained which have a Gain value of more than 0.2 have the largest rating, reaching 90%. The features chosen from 17 included features of family members, residence status, mother's work, caregivers, family support, joining extracurricular activities, repeating lessons at home, length of study at home, reading at home, reading time at home.

Keywords*: classification, naive bayes, feature selection, information gain, understanding, Indonesian language.*

1. PENDAHULUAN

Tujuan pendidikan, secara umum yaitu untuk memberikan peningkatan terhadap kecerdasan bangsa, dengan cara meningkatan pemahaman siswa terhadap mata pelajaran yang

diajarkan. Kemampuan setiap siswa dalam memahami pelajaran berbeda-beda, sehingga proses belajar siswa pun akan berbeda pula. Agar pengajar dapat menentukan cara pembelajaran yang dapat di terima oleh semua siswa, maka diperlukan metode prediksi kemampuan siswa.

Saat ini sistem belum tercapai dengan baik untuk menganalisa kemampuan siswa memahami materi pembelajaran. Ada dua alasan mengapa hal ini terjadi, pertama karena penelitian tentang metode prediksi masih belum cukup mengidentifikasi metode yang paling tepat untuk memprediksi kemampuan pemahaman siswa. Kedua karena kurangnya investigasi terhadap faktor-faktor yang mempengaruhi kemampuan pemahaman siswa. (Shahiria, et al., 2015)

Salah satu teknik prediksi dapat menggunakan metode klasifikasi. Kelas yang akan digunakan pada penelitian ini diantaranya siswa sangat paham, cukup paham dan kurang paham. Model prediksi tersebut mencakup seluruh faktor personal, sosial, lingkungan, psikologis yang digunakan agar prediksi tersebut efektif bagi kemampuan memahami materi pelajaran Bahasa Indonesia. Klasifikasi dilakukan dengan menggunakan algoritme *Naïve Bayes*. *Naïve Bayes* merupakan salah satu metode *machine learning* yang menggunakan konsep dasar dari sebuah teori bernama Teorema Bayes, teori ini melakukan klasifikasi dengan menghitung nilai probabilitas (Trisedya & Jais, 2009). *Naïve Bayes* dalam beberapa penelitian secara empiris terbukti sangat mudah di implementasikan ke dalam berbagai studi kasus. Selain itu algoritme ini juga memiliki performa pengklasifikasian yang cukup tinggi (Larose, 2005).

Selain bertujuan mendapatkan nilai akurasi yang baik juga bertujuan mendapatkan model fitur dengan cara menerapkan seleksi fitur. Seleksi fitur adalah salah satu cara untuk menentukan fitur yang paling memiliki pengaruh di dalam dataset. Seleksi fitur berperan dengan memilih subset yang tepat dari set fitur asli, dikarenakan bahwa tidak semua fitur dapat relevan atau memiliki hubungan dengan studi kasus (Maimon & Rokach, 2010). *Noisy Features* atau fitur yang tidak terpakai, harus dihapus untuk meningkatkan akurasi. Selain itu, pada pengklasifikasian memiliki fitur yang banyak akan memperlambat proses komputasi.

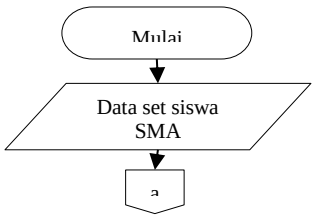
Penelitian terkait dengan penelitian ini adalah penilitan yang ditulis oleh Ade Ricky Rozzaqi pada tahun 2015 berjudul *Naïve Bayes* dan Filtering Feature Selection *Information Gain* untuk Prediksi Ketepatan Kelulusan Mahasiswa**.** Dalam penelitian ini dilakukan menggunakan dua metode yaitu metode yang hanya menggunakan algoritme klasifikasi *Naïve Bayes*, serta metode yang menggabungkan antara dua algoritme yaitu algoritme *Naïve Bayes* dan algoritme seleksi fitur *Information Gain*. Hasil penelitian menunjukan bahwa nilai akurasi tertinggi diperoleh dengan metode yang menggabungkan antara algoritme *Naïve Bayes* dan algoritme seleksi Fitur *Information Gain* yaitu mencapai nilai hingga 89,79 % fitur yang terpilih adalah sebayak 3 fitur, dan peningkatan AUC meningkat dengan 3 fitur tersebut. (Rozzaqi, 2015)

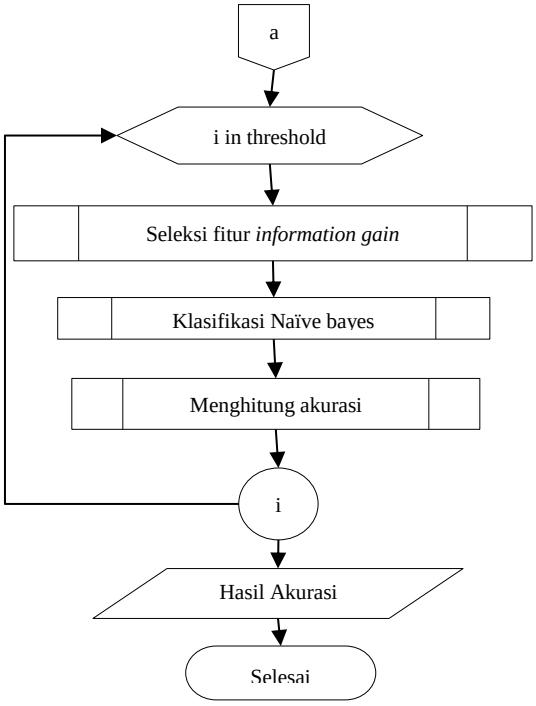
Berdasarkan uraian diatas, penulis melakukan penelitian untuk memprediksi kemampuan siswa dalam memahami materi pelajaran Bahasa Indonesia. Prediksi ini dapat digunakan untuk mengidentifikasi fitur penting dalam data siswa karena mengoptimalkan algoritme *Naïve Bayes* dengan menambah fitur yang sudah digunakan sebelumnya.

2. MODEL SISTEM

Berdasarkan

Gambar **1** secara umum sistem yang akan diimplementasikan adalah Sistem menerima masukan berupa data set siswa SMA yang akan diolah. Perulangan dilakukan saat i pada treshold kemudian dilakukan seleksi fitur dengan algoritme *Information Gain* lalu melakukan *training* dan *testing* dengan algoritme *Naïve Bayes* setelah itu melakukan perhitungan akurasi dengan menghitung nilai presisi, *recall* dan akurasi. Jika sudah memenuhi i maka akan memberikan keluaran hasil akurasi.





Gambar 1 Diagram Alir Kesuluruhan Sistem

Data yang digunakan terdiri dari 17 fitur dan 3 kelas, dapat dilihat pada tabel 1. Dan kelas terdiri dari kelas paham, cukup paham dan kurang paham.

Tabel 1 Fitur

Fitur	Kategori
Jumlah Anggota Keluarga	>3, <=3
Status tempat tinggal	Tinggal bersama dengan orang tua , Tinggal berpisah dengan orang tua
Pendidikan terakhir Ibu	Tidak sekolah, SD, SMP, SMA, >=S1
Pendidikan terakhir Ayah	Tidak sekolah, SD, SMP, SMA, >=S1
Pekerjaan Ibu	Tidak bekerja, PNS, Pegawai swasta, Pengajar, Wirausaha, Lainnya
Pekerjaan Ayah	Tidak bekerja, PNS, Pegawai swasta, Pengajar, Wirausaha, Lainnya
Pengasuh	Orang tua, Lainnya
Jarak Rumah	Sangat Dekat, Dekat, Sedang, Jauh
Dukungan Keluarga	Ya, Tidak
Ikut Ekstrakurikuler	Ya, Tidak
Mengulang Pelajaran Di rumah	Ya, Tidak
Posisi Duduk Di kelas	1,2,3,4,5
Internet di Rumah	Ada, Tidak
Mengikuti Bimbingan Belajar	Ya, Tidak
Lama Belajar di Rumah	< 1 jam, 1-2 jam , > 2 jam
Jenis bacaan di rumah	Buku pelajaran, Novel, Majalah, Koran, Lainnya
Lama membaca di rumah	< 1 jam, 1-2 jam , > 2 jam

2.1 Naive Bayes

Berdasarkan teorema Bayes, *naïve bayes* merupakan yang mengalami perkembangan sebagai pendeketan dalam melakukan klasifikasi terhadap kelas untuk suatu dokumen. Jika pada teorema bayes fitur-fitur yang ada terkait satu sama lain, maka pada *naïve bayes* ini memiliki asumsi bahwa setiap fitur yang ada adalah tidak memiliki kaitan satu sama lain. Walau pada kenyataan nya mungkin terdapat kaitan antara fitur-fitur ini. Berikut ini merupakan gambaran dari *naïve bayes* dalam melakukan klasifikasi:

$$p(c|d) = \frac{p(f_j|c) x\, p(c)}{p(d)} \qquad (1)$$

Probabilitas p(d|c) digantikan dengan perkalian probabilitas $p(f_j|c)$ dari f|buah fitur independen yang merepresntasikan d. Proses pembelajaran untuk topik yang telah didapatkan dari hasil pemodelan permasalahan, adalah dengan menghitung nilai $p(f_j|c)$ yang diperoleh dari data training.

Proses klasifikasi dokumen dilakukan dengan menentukan nila a yang akan memberikan nila p(a|b) yang paling besar dan dinyatakan sebagai berikut:

$$c^* = \arg\max_{c \epsilon A} p(c|d) = \arg\max \prod_{c \epsilon A} p(f_j|d) x\, p(c) \quad (2)$$

Kelas c* merupakan kelas yang memilii nilai p(c|d) terbesar. Nilai p(d) dapat diabaikan karena nilai p(d) akan bernilai sama untuk semua kelas sehingga tidak akan memberikan pengaruh apapun dalam proses perbandiangan p(c|d). (Mitchell, 1997)

2.2 Information Gain

Information Gain digunakan sebagai fitur pemilih ukuran. Fitur dengan *information Gain* tertinggi dipilih sebagai fitur pemisah untuk node N. Fitur ini meminimalisasi informasi yang dibutuhkan untuk mengklasifikasi tuple dalam memberikan hasil pembelahan dan perlu dielaborasi nilai acak yang paling sedikt "kesalahannya" pada kelas partisi tersebut. (Maimon & Rokach, 2010)

Perhitungan *Entropy* dapat dilakukan seperti pada rumus berikut:

$$info\ (D) = -\sum pi\ log_2(pi) \quad (3)$$

D adalah himpunan kasus pi : Proporsi dari Di terhadap D. Fungsi log dalam hal ini digunakan log berbasis 2 karena informasi dikodekan berbasis bit.

Melakukan perhitungan nilai setelah pemisahan dapat dilakukan dengan menggunakan rumus berikut:

$$info_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|}\ x\ I(Dj) \quad (4)$$

D adalah hmpunan kasus, A adalah fiturnya, v jumlah partisi fitur A. |Dj| adalah Jumlah kasus pada partisi ke j, |D| yaitu jumlah kasus dalam D , I (Dj) adalah total *entropy* dalam partisi.

Terakhir menetapkan *information Gain* untuk fitur A dengan rumus:

$$Gain(A) = I(D) - I(A) \quad (5)$$

Gain (A) adalah *Information* fitur A, I(D) adalah total *entropy* dan I (A) adalah *entropy* A

2.3 Confusion Matrix

Untuk melakukan perhitungan Akurasi, *Precision* dan *recall*, perlu dilakukan pencarian nilai TF (*True Positif*), TN (*True Negatif*) FP (*False Positif*) dan (*False Negatif*). TF adalah jumlah data yang bersifat benar yang ada dikelas itu sendiri. TN adalah jumlah data yang bersifat benar pada kelas lain selain kelasnya sendiri. FP adalah jumlah data yang salah pada kelasnya sendiri. Sedangkan FN adalah jumlah data salah pada kelas lain selain kelasnya sendiri. Klasifikasi ini bertipe *multiclass* maka pencarian confusion matrixnya dilakukan untuk setiap kelasnya.

Tabel 2 Confusion Matrix

Nilai Prediksi	True	False
True	TP	FP
False	FN	TN

Persamaan untuk menghitung nilai *Precision* adalah sebagai berikut:

$$Precision = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} TP_i + FP_i} \quad (6)$$

Nilai *Precision* adalah untuk pengukuran kemiripan atau kecocokan antara informasi yang ingin didapatkan dengan hasil yang didapatkan. Nilai *Precision* menggambarkan jumlah data dengan kategori positif yang terklasifikasi dengan benar dibagi dengan seluruh jumlah data yang diklasifikasi positif.

$$\boldsymbol{Recall} = \frac{\sum_{i=1}^{l} \boldsymbol{TP_i}}{\sum_{i=1}^{l} \boldsymbol{TP_i + FN_i}} \quad (7)$$

Nilai *Recall* adalah seberapa banyak data dengan kategori positif yang terklasifikasikan secara benar oleh sistem. Nilai *recall* menggambarkan jumlah data dengan kategori positif yang terklasifikasikan dibagi dengan seluruh dokumen yang relevan.

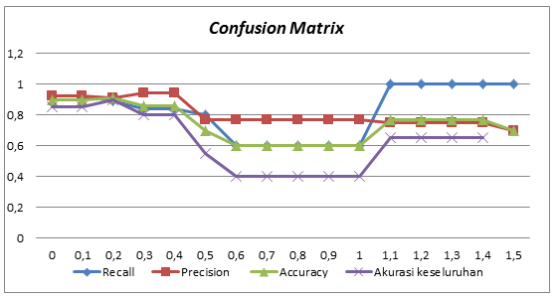
$$\boldsymbol{Accuracy} = \sum_{i=1}^{l} \frac{\frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}}{\boldsymbol{l}} \quad (8)$$

Nilai akurasi menunjukan bahwa sberapa akurat sistem dapat memberikan hasil klasifikasi dari data dengan benar. Artinya nilai akurasi ini adalah nilai yang membandingkan antara nilai hasil klasifikasi yang benar dengan seluruh hasil klasifikasi.

Selain itu, akan dilakukan pengujian fitur menggunakan nilai threshold dari nilai *Gain* masing-masing fitur dengan merubah nilai threshold mulai dari 0 hingga 1.5 kemudian akan dilakukan perhitungan satu persatu nilai dari *Confusion Matrix.*

3. HASIL DAN PEMBAHASAN

Pengujian dilakukan menggunakan data set sebanyak 100 yang terdiri dari 80 data latih dan 20 data uji. Berdasarkan skenario uji yaitu dengan merubah nilai threshold dari ≥ 0 hingga ≥1.5, didapatkan hasil akurasi, *recall, precision* dari setiap kelas. Hal tersebut terjadi karena pada penelitian ini mengguanakan klasifikasi bertipe *multi-class* yaitu kelas berjumlah lebih dari 2. Terdapat juga hasil akurasi sistem secara keseluruhan. Dapat dilihat dari Gambar 2. Angka secaravertikal yaitu nilai akurasinya, dan angka secarahorizontal adalah nilai gainnya.



Gambar 2 Diagram Confusion Matrix

3.1. Akurasi

Gambar 2 menyatakan bahwa akurasi secara keseluruhan tertinggi diperoleh dari nilai *Gain* lebih dari 0.2 yaitu sebesar 90%. Confusion Matrix dari hasil klasifikasi yang memiliki nilai gain lebih dari 0.2 dapat dilihat pada Tabel **3** *Confusion Matrix Gain* ≥ 0.2.

Tabel 3 *Confusion Matrix Gain* ≥ 0.2

Predicted *Actual*	cukup paham	kurang paham	paham	jumlah
Cukup paham	3	0	1	4
kurang paham	0	5	0	5
paham	0	1	10	11
Jumlah	3	6	11	20

Fitur yang memiliki nilai *Gain* lebih dari 0.2 terpilih fitur ke 0, 1, 4, 6, 8, 9, 10, 14, 15, 16 adalah fitur jumlah anggota keluarga, status tempat tinggal, pekerjaan ibu, pengasuh, dukungan keluarga, ikut ekstrakurikuler, mengulang pelajaran di rumah, lama belajar di rumah, jenis bacaan di rumah, lama membaca di rumah. Fitur yang tidak terpilih adalah fitur pendidikan terakhir Ibu, pendidikan terakhir ayah, pekerjaan ayah, jarak rumah, posisi duduk dikelas, internet di rumah dan mengikuti bimbingan belajar.

3.2. *Recall*

Recall tertinggi diperoleh dari niai *Gain* yang bernilai lebih dari 1.1 hingga 1.5 yaitu mencapai angka 1 dengan fitur yang terpilih adalah fitur ke 8 dan fitur 9 yaitu fitur dukungan keluarga dan ikut ekstrakurikuler.

Tabel 4 *Confusion Matrix ≥ 1.1*

Predicted *Actual*	cukup paham	kurang paham	paham	jumlah
Cukup paham	2	0	2	4
kurang paham	0	0	5	5
paham	0	0	11	11
Jumlah	2	0	18	20

Berdasarkan Tabel 4 *Confusion Matrix ≥ 1.1*, maka dilakukan perhitungan nilai *recall* seperti pada persamaan (7) yaitu dengan menghitung berapa banyak data dengan kategori positif yang terklasifikasikan secara benar oleh sistem. Seperti kita ketahui *recall* merupakan kualitas seberapa lengkap hasil relevan yang ditampilkan oleh sistem pencarian maka dapat dikatakan bahwa sistem ini memberikan hasil yang relevan.

3.3. *Precision*

Precision tertinggi diperoleh oleh nilai *Gain* yang lebih dari atau sama dengan 0.4 yaitu mencapai 0.942129667 dengan fitur yang terpilih adalah fitur 1, 4, 6, 8, 9, 14, 16 pekerjaan ibu, pengasuh, dukungan keluarga, ikut ekstrakurikuler, lama belajar di rumah, lama membaca di rumah.

Tabel 5 *Confusion Matrix ≥ 0.4*

Predicted *Actual*	cukup paham	kurang paham	paham	jumlah
Cukup paham	3	0	1	4
kurang paham	0	5	0	5
paham	1	2	8	11
Jumlah	4	7	9	20

Berdasarkan Tabel 5 *Confusion Matrix ≥ 0.4*, maka dilakukan perhitungan nilai *precision* seperti pada persamaan (6) yaitu menjumlahkan data dengan kategori positif yang terklasifikasi dengan benar dibagi dengan seluruh jumlah data yang diklasifikasi positif. *Precision* merupakan pengukuran kualitas seberapa bergunakah sistem pencarian tersebut maka dapat dikatakan bahwa sistem ini 94% berguna.

4. KESIMPULAN

Kesimpulan yang diperoleh oleh penulis adalah akurasi keseluruhan tertinggi pada sistem ini adalah sebesar 0.9 yaitu dengan fitur yang terpilih adalah fitur yang memiliki nilai *Gain* lebih dari atau sama dengan 0.2. Rata-rata akurasi dari setiap kelas adalah sebesar 0.914 fitur dengan nilai *Gain* yang sama yaitu 0.2. Rata-rata *recall* tertinggi adalah sebesar 1 yaitu fitur dengan nilai *Gain* yang bernilai 1.1 hingga 1.5. Sedangkan untuk *prcision* yang tertinggi adalah sebesar 0.942 yaitu fitur yang memiliki nilai *Gain* lebih dari atau sama dengan 0.4. Fitur yang memberikan pengaruh terhadap hasil klasifikasi adalah fitur yang memiliki nilai *Gain* lebih dari atau sama sengan 0.2 diantaranya yaitu fitur index ke 0, 1, 4, 6, 8, 9, 10, 14, 15, 16 antaralain fitur jumlah anggota keluarga, status tempat tinggal, pekerjaan ibu, pengasuh, dukungan keluarga, ikut ekstrakurikuler, mengulang pelajaran di rumah, mengikuti bimbingan belajar, lama belajar di rumah, jenis bacaan di rumah dan lama membaca di rumah.

Untuk penelitian selanjutnya penulis merasa perlu untuk menggunakan metode klasifikasi yang lain selain klasifikasi *naive bayes* dan menggunakan metode seleksi fitur yang lain

selain *information Gain* agar dapat menjadi perbandingan metode klasifikasi apa yang memiliki kualitas dan hasil yang terbaik. Serta lakukan analaisis pada pemahaman selain mata pelajaran Bahasa Indonesia dengan mengubah fitur-fitur dari penelitian ini dengan fitur yang memiliki kerterkaitan dengan mata pelajaran yang akan diteliti.

5. DAFTAR PUSTAKA

Ali Khan, S., 2005. *Filsafat Pendidikan Al-Ghazali.* Bandung: Pustaka Setia. Anon., n.d. s.l.:s.n.

Larose, D. T., 2005. *Discovering Knowledge in Data An Introduction to Data Mining.* 1st ed. Canada: A John Wiley & Sons, Inc..

Maimon, O. & Rokach, L., 2010. *Data Mining and Knowledge Discovery Handbook.* 2nd ed. Eropa: Springer.

Mitchell, T. M., 1997. *Machine Learning,*. United States of America: McGraw –Hill.

Moeljono, S., 1989. *Bahasa Indonesia dan Problematikanya.* 1st ed. Madiun: Widya Mandala.

Rozzaqi, A. R., 2015. *Naïve Bayes dan Filtering Feature Selection Information Gain untuk Prediksi Ketepatan Kelulusan Mahasiswa,* Semarang: Universitas PGRI Semarang.