PENERAPAN METODE K-NEAREST NEIGHBOR DAN INFORMATION GAIN PADA KLASIFIKASI KINERJA SISWA

Tyas Setiyorini[1]; Rizky Tri Asmono[2]

Teknik Informatika[1]
STMIK Nusa Mandiri Jakarta[1];
http://nusamandiri.ac.id[1]
tyas.setiyorini@gmail.com[1]

Teknik Informatika[2]
STMIK Swadharma[2]
http://swadharma.ac.id[2]
rtriasmono@gmail.com[2]

Abstract— *Education is a very important problem in the development of a country. One way to reach the level of quality of education is to predict student academic performance. The method used is still using an ineffective way because evaluation is based solely on the educator's assessment of information on the progress of student learning. Information on the progress of student learning is not enough to form indicators in evaluating student performance and helping students and educators to make improvements in learning and teaching. K-Nearest Neighbor is an effective method for classifying student performance, but K-Nearest Neighbor has problems in terms of large vector dimensions. To solve this problem, the Information Gain feature selection method is needed to reduce vector dimensions. Several experiments were conducted to obtain an optimal architecture and produce accurate classifications. The results of 10 experiments with a value of k (1 to 10) on the student performance dataset using the K-Nearest Neighbor method showed the highest average accuracy of 74.068 whereas the K-Nearest Neighbor and Information Gain methods obtained the highest average accuracy of 76.553. From the results of these tests it can be concluded that Information Gain is able to reduce the vector dimensions, so that the application of K-Nearest Neighbor and Information Gain can improve the accuracy of student performance classification better than using the K-Nearest Neighbor method.*

Keywords**:** *K-Nearest Neighbor, Information Gain, Student Performance*

Intisari— Pendidikan merupakan masalah yang sangat penting dalam perkembangan suatu negara. Salah satu cara untuk mencapai tingkat kualitas pendidikan adalah dengan memprediksi kinerja akademik siswa. Metode yang dilakukan masih menggunakan cara yang tidak efektif karena evaluasi hanya berdasarkan pada penilaian pendidik terhadap informasi kemajuan pembelajaran siswa. Informasi kemajuan pembelajaran siswa tidak cukup untuk membentuk indikator dalam mengevaluasi kinerja siswa serta membantu para siswa dan pendidik untuk melakukan perbaikan dalam pembelajaran dan pengajaran. K-Nearest Neighbor merupakan metode yang efektif untuk klasifikasi kinerja siswa, namun K-Nearest Neighbor memiliki masalah dalam hal dimensi vektor yang besar. Untuk menyelesaikan masalah tersebut diperlukan metode seleksi fitur Information Gain untuk mengurangi dimensi vektor. Beberapa percobaan dilakukan untuk mendapatkan arsitektur yang optimal dan menghasilkan klasifikasi yang akurat. Hasil dari 10 percobaan dengan nilai k (1 sampai dengan 10) pada dataset *student performance* dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Information Gain didapatkan rata-rata akurasi terbesar yaitu 76,553. Dari hasil pengujian tersebut maka dapat disimpulkan bahwa Information Gain mampu mengurangi dimensi vektor, sehingga penerapan K-Nearest Neighbor dan Information Gain dapat meningkatkan akurasi klasifikasi kinerja siswa yang lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja.
. .

Kata Kunci*:* K-Nearest Neighbor, Information Gain, Kinerja Siswa

PENDAHULUAN

Pendidikan merupakan masalah yang sangat penting dalam perkembangan suatu negara. Pendidikan yang berkualitas merupakan tujuan

utama dalam sebuah lembaga pendidikan. Salah satu cara untuk mencapai tingkat kualitas pendidikan adalah dengan memprediksi kinerja akademik siswa (Hamsa, Indiradevi, & Kizhakkethottam, 2016). Memprediksi secara akurat kinerja akademik siswa pada tahap awal pembelajaran membantu dalam mengidentifikasi siswa yang lemah dan memungkinkan manajemen untuk mengambil tindakan korektif untuk mencegah mereka dari kegagalan (Pandey & Taruna, 2016). Menyajikan pendidikan yang berkualitas dalam meningkatkan kinerja siswa merupakan tujuan utama dari lembaga pendidikan (Hamsa et al., 2016). Untuk mencapai tujuan tersebut, yang perlu dilakukan adalah menganalisis faktor-faktor apa saja yang mempengaruhi kinerja siswa. Dengan menganalisis kinerja siswa, program strategis dapat direncanakan dengan baik selama masa studi mereka di sebuah institusi (Ibrahim & Rusli, 2007). Namun, pekerjaan yang ada tidak menyediakan alat analisis yang cukup untuk menganalisis bagaimana siswa melakukan, faktor mana yang akan mempengaruhi kinerjanya, dengan cara mana siswa dapat membuat kemajuan, dan apakah siswa memiliki potensi untuk melakukan yang lebih baik (Yang & Li, 2018). Informasi kemajuan pembelajaran siswa menjadi salah satu faktor penilaian seorang pendidik dalam menganalisis kinerja siswa. Namun cara tersebut tidaklah efektif karena informasi kemajuan pembelajaran siswa tidak cukup sebagai indikator para siswa dan pendidik untuk melakukan perbaikan dalam pengajaran dan pembelajaran (Yang & Li, 2018).

Salah satu teknik yang paling populer untuk menganalisis kinerja siswa adalah data mining (Shahiri, Husain, & Rashid, 2015). Pendekatan data mining diusulkan untuk memprediksi kinerja siswa (Hamsa et al., 2016). Beberapa penelitian telah dilakukan dalam memprediksi kinerja siswa dengan teknik klasifikasi, seperti K-Nearest Neighbor (Pandey & Taruna, 2016), Regression (Conijn, Snijders, Kleingeld, & Matzat, 2017), Support Vector Machine (Al-Shehri et al., 2017), Decision Tree (Lopez Guarin, Guzman, & Gonzalez, 2015), Naive Bayes (Lopez Guarin et al., 2015), dan Artificial Neural Networks (Alkhasawneh & Hobson, 2011).

K-Nearest Neighbor bersifat efektif, intuitif dan sederhana sehingga K-Nearest Neighbor telah menarik minat luas dalam komunitas penelitian (Gou et al., 2014)(Lin, Li, Lin, & Chen, 2014)(Lin et al., 2014). K-Nearest Neighbor adalah salah satu metode yang mampu memecahkan masalah klasifikasi, sering menghasilkan hasil yang kompetitif dan memiliki keuntungan yang signifikan atas beberapa metode penambangan data lainnya (Adeniyi, Wei, & Yongquan, 2016).

K-Nearest Neighbor merupakan metode yang efektif namun memiliki beberapa kekurangan yaitu kompleksitas komputasi kemiripan datanya besar, kinerjanya mudah dipengaruhi oleh data noise, K-Nearest Neighbor merupakan metode lazy learning sehingga tidak membangun model klasifikasi. Untuk mengurangi kompleksitas K-Nearest Neighbor dapat dilakukan dengan tiga metode umum, yaitu mengurangi jumlah data pelatihan (Lu & Fa, 2004), mempercepat proses menemukan k tetangga terdekat (Aghbari, 2005), atau mengurangi dimensi vektor (de Vries, Mamoulis, Nes, & Kersten, 2003).

Seleksi fitur (*feature selection*) dapat digunakan untuk mengurangi dimensi vektor pada dataset *student performance*. Seleksi fitur merupakan salah satu bagian terpenting dalam mengoptimalkan performa *classifier* (Wang, Li, Song, Wei, & Li, 2011). Seleksi fitur berdasarkan pada pengurangan fitur yang besar, yaitu dengan menghapus atribut yang tidak relevan (Koncz & Paralic, 2011). Menggunakan algoritma seleksi fitur yang tepat dapat meningkatkan akurasi (Xu, Peng, & Cheng, 2012)(George Gorman, 2003).

Yang dan Perdersen (Vercellis, 2009) melakukan perbandingan 5 algoritma seleksi fitur pada klasifikasi. Lima algoritma tersebut antara lain mutual information, term strength, chi-square, Information Gain, dan document frequency. Hasil penelitian tersebut membuktikan bahwa chi-square dan Information Gain paling efisien. Tan dan Zang (Zhang & Tan, 2008) menggunakan algoritma seleksi fitur yang menunjukkan Information Gain mendapatkan hasil yang paling baik. Hal tersebut menunjukkan bahwa Information Gain memiliki potensi yang lebih baik dalam proses menghilangkan fitur yang kurang relevan sehingga dapat mengurangi dimensi vektor sebelum dilakukannya klasifikasi. Untuk itu penelitian ini akan menggunakan kombinasi kedua metode yaitu K-Nearest Neighbor dan Information Gain untuk melakukan klasifikasi kinerja siswa sehingga mendapatkan tingkat akurasi yang optimal.

BAHAN DAN METODE

Bahan

Dataset *student performance* digunakan pada penelitian ini. Dataset tersebut didapat dari UCI Machine Learning Repository. Dataset *student performance* terdiri dari 30 atribut dan 1 kelas. Tabel 1 menunjukkan atribut dan keterangannya. Tabel 2 menunjukkan atribut, data, dan keterangan datanya.

Tabel 1. Atribut dan Keterangan pada Dataset *Student Performance*

No	Atribut	Keterangan
1	Result	Hasil kelulusan. (Merupakan atribut class)
2	School	Nama Sekolah
3	Sex	Jenis Kelamin
4	Age	Umur
5	Address	Alamat
6	Famsize	Jumlah anggota keluarga
7	Pstatus	Status tinggal dengan orang tua atau tidak
8	Medu	Pendidikan ibu
9	Fedu	Pendidikan ayah
10	Mjob	Pekerjaan ibu
11	Fjob	Pekerjaan ayah
12	Reason	Alasan memilih sekolah
13	Guardian	Wali siswa
14	Traveltime	Waktu tempuh dari rumah ke sekolah
15	Studytime	Waktu belajar dalam seminggu
16	Failures	Jumlah ketidaklulusan
17	Schoolsup	Dukungan pendidikan tambahan
18	Famsup	Dukungan pendidikan keluarga
19	Paid	Les tambahan
20	Activities	Kegiatan ekstrakurikuler
21	Nursery	
22	Higher	Ingin mengambil pendidikan tinggi
23	Internet	Akses internet di rumah
24	Romantic	Mempunyai pacar atau tidak
25	Famrel	Kualitas hubungan keluarga
26	Freetime	Waktu luang setelah sekolah
27	Goout	Pergi bersama teman-teman
28	Dalc	Mengkonsumsi alkohol pada hari kerja
29	Walc	Mengkonsumsi alkohol pada akhir pekan
30	Health	Status kesehatan saat ini
31	Absences	Jumlah ketidakhadiran

Sumber: (Cortez & Silva, 2008)

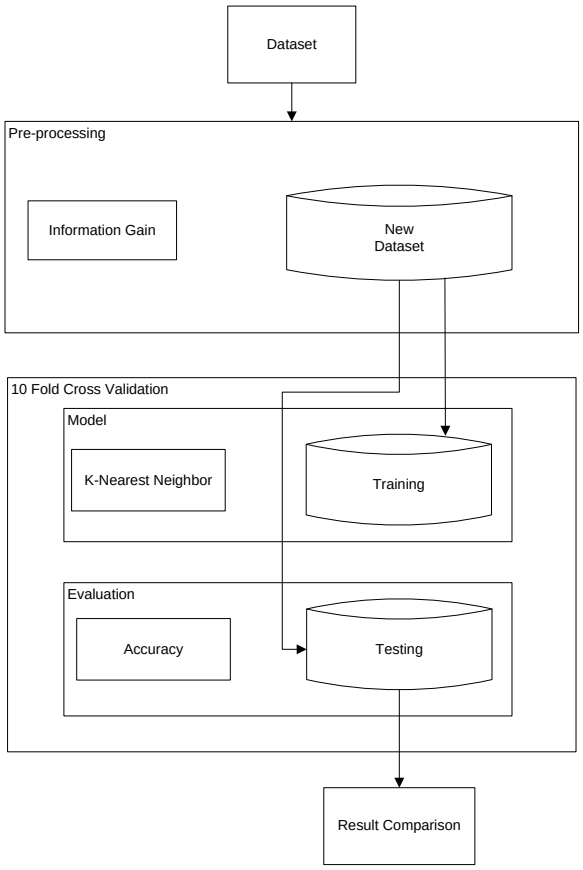
Tabel 2. Atribut, Data dan Keterangan Data pada Dataset *Student Performance*

No	Atribut	Data	Keterangan Data
1	Result	Fail/ pass	Gagal/ lulus
2	School	MS/ GP	MS: Mousinho da Silveira GP: Gabriel Pereira
3	Sex	M/ F	Laki-laki/ perempuan
4	Age	15-22	
5	Address	R/U	R: rural, U: urban
6	Famsize	LE3/GT3	LE3: <=3 GT: >3
7	Pstatus	A/T	A: terpisah T: bersama orang tua
8	Medu	0/ 1/ 2/ 3/ 4	0: tidak ada 1: SD 2: SMP 3: SMA 4: pendidikan yang lebih tinggi
9	Fedu	0/ 1/ 2/ 3/ 4	0: tidak ada 1: SD 2: SMP 3: SMA 4: pendidikan yang lebih tinggi
10	Mjob	Techer/ health/ services/ at home/ other	Teacher: guru Health: di bidang kesehatan Services: PNS At home: di rumah Other: lain-lain
11	Fjob	Techer/ health/ services/ at home/ other	Teacher: guru Health: di bidang kesehatan Services: PNS At home: di rumah

			Other: lain-lain
12	Reason	Home/ reputation/ course/ other	Home: dekat dengan rumah Reputation: reputasi sekolah Course: mata pelajaran
13	Guardian	Mother/ father/ other	Ayah/ Ibu/ Lain-lain
14	Traveltime	1/ 2/ 3/ 4	1: <15 menit 2: 15-30 menit 3: 30 menit-1 jam 4: > 1 jam
15	Studytime	1/ 2/ 3/ 4	1: < 2 jam 2: 2-5 jam 3: 5-10 jam jam 4: > 10 jam
16	Failures	1/ 2/ 3/ 4	1: 1 kali 2: 2 kali 3: 3 kali 4: > 3 kali
17	Schoolsup	Yes/ no	
18	Famsup	Yes/ no	
19	Paid	Yes/ no	
20	Activities	Yes/ no	
21	Nursery	Yes/ no	
22	Higher	Yes/ no	
23	Internet	Yes/ no	
24	Romantic	Yes/ no	
25	Famrel	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
26	Freetime	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
27	Goout	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
28	Dalc	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
29	Walc	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
30	Health	1/ 2/ 3/ 4/ 5	1: sangat buruk 2: buruk 3: normal 4: baik 5: sangat baik
31	Absences	0-75	

Sumber: (Cortez & Silva, 2008)

Metode



Sumber: (Setiyorini & Asmono, 2019)

Gambar 1. Penerapan Metode K-Nearest Neighbor dan Information Gain

Gambar 1 menggambarkan metode yang diusulkan dalam penelitian ini yaitu metode K-Nearest Neighbor dan Information Gain. Pada tahap *pre-processing*, dilakukan seleksi fitur dengan menggunakan metode Information Gain sehingga menghasilan *new dataset* dengan atribut-atribut yang paling optimal. Kemudian *new dataset* dibagi dengan metode 10 Fold Cross Validation yaitu *data training* dan *data testing.* Kemudian data training diklasifikasi dengan menggunakan metode K-Nearest Neighbor. Langkah terakhir data testing diuji dengan melihat performa akurasi.

K-Nearest Neighbor

K-Nearest Neighbor merupakan algoritma yang efektif dan kuat. Dalam pengenalan pola, algoritma K-Nearest Neighbor adalah salah satu metode non-parametrik yang paling terkenal dan berguna untuk mengelompokkan objek berdasarkan fitur-fitur yang dekat. K-Nearest Neighbor dirancang dengan konsep bahwa label atau kelas ditentukan oleh suara mayoritas tetangganya (Won Yoon & Friel, 2015). Prinsip kerja KNN adalah mencari jarak terdekat antara data yang dievaluasi dengan k tetangga terdekatnya dalam data pelatihan. Persamaan penghitungan untuk mencari Euclidean dengan d adalah jarak dan p adalah dimensi data dengan:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

di mana:
x1: sample data uji
x2: data uji
d: jarak
p: dimensi data

Information Gain

Information Gain sering digunakan untuk meranking atribut yang paling berpengaruh terhadap kelasnya. Nilai gain dari suatu atribut, diperoleh dari nilai entropi sebelum pemisahan dikurangi dengan nilai entropi setelah pemisahan. Tujuan pengurangan fitur pengukuran nilai informasi diterapkan sebagai tahap sebelum pengolahan awal. Hanya atribut memenuhi kriteria (treshold) yang ditentukan dipertahankan untuk digunakan oleh algoritma klasifikasi (Hand, 2007). Ada 3 tahapan dalam pemilihan fitur menggunakan Information Gain diantaranya adalah:

1. Hitung nilai gain informasi untuk setiap atribut dalam dataset asli.
2. Buang semua atribut yang tidak memenuhi kriteria yang ditentukan.
3. Dataset direvisi.

Pengukuran atribut ini dipelopori oleh Cloude Shannon pada teori informasi (Gallager, 2001), dituliskan sebagai (Han, Kamber, & Pei, 2012):

$$Info\ (D) = -\sum_{i=1}^{m}\ p_i log_2(p_i) \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

di mana:
D: Himpunan Kasus
m: Jumlah partisi D
p_i : Proporsi dari D_i terhadap D.

Dalam hal ini p_i adalah probabilitas sebuah *tuple* pada D masuk ke kelas C_i dan diestimasi dengan $|C_i,D|/|D|$. Fungsi log diambil berbasis 2 karena informasi dikodekan berbasis bit. Selanjutnya nilai entropi setelah pemisahan dengan cara sebagai berikut (Han et al., 2012).

$$Info_A\ (D) = -\sum_{j=1}^{v}\ \frac{|D_j|}{|D|} \times Info\ (D_j) \dots\dots\dots\dots (3)$$

di mana:
D: himpunan kasus
A: atribut
v: jumlah partisi atribut A
$|D_j|$: jumlah kasus pada partisi ke j
$|D|$: jumlah kasus dalam D
Info(D_j): total entropi dalam partisi

Untuk mencari nilai information gain atribut A diperoleh dengan persamaan berikut (Han et al., 2012):

$$Gain\ (A) = Info(D) -\ Info_A\ (D) \dots\dots\dots\dots\dots (4)$$

di mana:
Gain(A): Information atribut A
Info(D): Total entropi
$Info_A$(D): Entropi A

Dengan penjelasan lain, Gain (A) adalah reduksi yang diharapkan di dalam entropi yang disebabkan oleh pengenalan nilai atribut dari A. Atribut yang memiliki nilai Information Gain terbesar dipilih sebagai uji atribut untuk himpunan S. Selanjutnya suatu simpul dibuat dan diberi label dengan label atribut tersebut, dan cabang-cabang dibuat untuk masing masing nilai dari atribut

HASIL DAN PEMBAHASAN

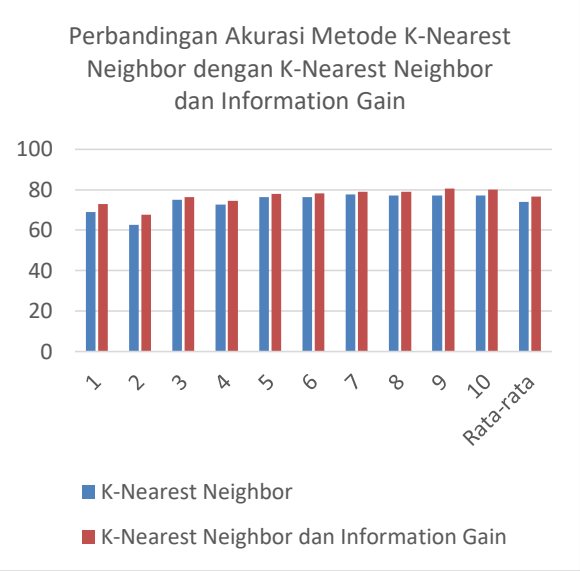
Tabel 3 merupakan perbandingan akurasi metode K-Nearest Neighbor dengan K-Nearest Neighbor dan Information Gain pada

klasifikasi kinerja siswa dengan menggunakan dataset *student performance*. Pada Tabel 3 menunjukkan dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Information Gain didapatkan rata-rata akurasi terbesar yaitu 76,553. Pada Gambar 2 juga menunjukkan kenaikan grafik pada penggunaan metode K-Nearest Neighbor dan Information Gain dibanding dengan metode K-Nearest Neighbor saja.

Tabel 3. Perbandingan Akurasi K-Nearest Neighbor dengan K-Nearest Neighbor dan Information Gain

	Akurasi	
Percobaan (k)	**K-Nearest Neighbor**	**K-Nearest Neighbor dan Information Gain**
1	68,96	72,98
2	62,55	67,53
3	75	76,24
4	72,6	74,52
5	76,34	77,78
6	76,34	78,07
7	77,58	78,84
8	77,11	78,93
9	77,1	80,56
10	77,1	80,08
Rata-rata	74,068	76,553

Sumber: (Setiyorini & Asmono, 2019)



Sumber: (Setiyorini & Asmono, 2019)

Gambar 2. Grafik Perbandingan Akurasi K-Nearest Neighbor dengan K-Nearest Neighbor dan Information Gain

Dari hasil pengujian tersebut menunjukkan bahwa Information Gain pada K-Nearest Neighbor mampu mengurangi dimensi vektor, sehingga menghasilkan tingkat akurasi klasifikasi kinerja siswa lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja. Hal ini membuktikan penelitan Gorman (George Gorman, 2003), Tan dan Zang(Zhang & Tan, 2008) bahwa Information Gain mampu mengurangi dimensi vektor. Selain itu hasil tersebut juga membuktikan penelitan Setiyorini dan Asmono (Setiyorini, 2017), bahwa klasifikasi tingkat kognitif soal pada taksonomi Bloom dengan menggunakan metode K-Nearest Neighbor dan Information Gain menghasilkan tingkat akurasi yang lebih baik dibanding menggunakan metode K-Nearest Neighbor saja.

KESIMPULAN

Hasil dari 10 percobaan dengan nilai k (1 sampai dengan 10) pada dataset *student performance* dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Information Gain didapatkan rata-rata akurasi terbesar yaitu 76,553. Dari hasil pengujian tersebut maka dapat disimpulkan bahwa seleksi fitur dengan Information Gain mampu mengurangi dimensi vektor, sehingga penerapan K-Nearest Neighbor dan Information Gain dapat meningkatkan akurasi klasifikasi kinerja siswa yang lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja.

REFERENSI

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108. https://doi.org/10.1016/j.aci.2014.10.001

Aghbari, Z. Al. (2005). Array-index: A plug&search K nearest neighbors method for high-dimensional data. *Data and Knowledge Engineering*, *52*(3), 333–352. https://doi.org/10.1016/j.datak.2004.06.015

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., … Olatunji, S. O. (2017). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *Canadian Conference on Electrical and Computer Engineering*, 17–20.

https://doi.org/10.1109/CCECE.2017.7946847

Alkhasawneh, R., & Hobson, R. (2011). Modeling student retention in science and engineering disciplines using neural networks. In *2011 IEEE Global Engineering Education Conference, EDUCON 2011* (pp. 660–663). https://doi.org/10.1109/EDUCON.2011.5773209

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 5–12.

de Vries, A. P., Mamoulis, N., Nes, N., & Kersten, M. (2003). Efficient k-NN search on vertically decomposed data (p. 322). https://doi.org/10.1145/564728.564729

Gallager, R. G. (2001). Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Transactions on Information Theory*, *47*(7), 2681–2695. https://doi.org/10.1109/18.959253

George Gorman. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Gou, J., Zhan, Y., Rao, Y., Shen, X., Wang, X., & He, W. (2014). Improved pseudo nearest neighbor classification. *Knowledge-Based Systems*, *70*, 361–375. https://doi.org/10.1016/j.knosys.2014.07.020

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Procedia Technology*, *25*, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. *Data Mining*. https://doi.org/10.1016/b978-0-12-381479-1.00001-0

Hand, D. J. (2007). Principles of data mining. *Drug Safety*, *30*(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010

Ibrahim, Z., & Rusli, D. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision tree And Linear Regression. *Proceedings of the 21st Annual SAS Malaysia Forum*, (September), 1–6. Retrieved from https://www.researchgate.net/profile/Daliela_Rusli/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression/links/0deec51bb04e76ed93000000.pdf

Koncz, P., & Paralic, J. (2011). An approach to feature selection for sentiment analysis. In *INES 2011 - 15th International Conference on Intelligent Engineering Systems, Proceedings* (pp. 357–362). https://doi.org/10.1109/INES.2011.5954773

Lin, Y., Li, J., Lin, M., & Chen, J. (2014). A new nearest neighbor classifier via fusing neighborhood information. *Neurocomputing*, *143*, 164–169. https://doi.org/10.1016/j.neucom.2014.06.009

Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *10*(3), 119–125. https://doi.org/10.1109/RITA.2015.2452632

Lu, L. R., & Fa, H. Y. (2004). A Density-Based Method for Reducing the Amount of Training Data in kNN Text Classification [J]. *Journal of Computer Research and Development*, *4*, 003.

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366. https://doi.org/10.1016/j.pisc.2016.04.076

Setiyorini, T. (2017). Penerapan Information Gain pada K-Nearest Neighbor untuk Klasifikasi

Tingkat Kognitif Soal pada Taksonomi Bloom. *Sistem Informasi STMIK Antar Bangsa*, *VI*, 57–62.

Setiyorini, T., & Asmono, R. T. (2019). *Laporan Akhir Penelitian Mandiri* (Vol. 1).

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Vercellis, C. (2009). *Data mining and optomization for decision making*. *Business Intelligence* (Vol. 1). https://doi.org/10.1017/CBO9781107415324.004

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, *38*(7), 8696–8702. https://doi.org/10.1016/j.eswa.2011.01.077

Won Yoon, J., & Friel, N. (2015). Efficient model selection for probabilistic K nearest neighbour classification. *Neurocomputing*, *149*(PB), 1098–1108. https://doi.org/10.1016/j.neucom.2014.07.023

Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems*, *35*, 279–289. https://doi.org/10.1016/j.knosys.2012.04.011

Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers and Education*, *123*(October 2017), 97–108. https://doi.org/10.1016/j.compedu.2018.04.006

Zhang, J., & Tan, S. (2008). An empirical study of sentiment analysis for chinese documents. *EXPERT SYSTEMS WITH APPLICATIONS*, *34*(4), 2622–2629. https://doi.org/10.1016/j.eswa.2007.05.028