KOMPARASI METODE DECISION TREE, NAIVE BAYES DAN K-NEAREST NEIGHBOR 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—*In education, student performance is an important part. To achieve good and quality student performance requires analysis or evaluation of factors that influence student performance. The method still using an evaluation based only on the educator's assessment of information on the progress of student learning. This method is not effective because information such as student learning progress is not enough to form indicators in evaluating student performance and helping students and educators to make improvements in learning and teaching. Previous studies have been conducted but it is not yet known which method is best in classifying student performance. In this study, the Decision Tree, Naive Bayes and K-Nearest Neighbor methods were compared using student performance datasets. By using the Decision Tree method, the accuracy is 78.85, using the Naive Bayes method, the accuracy is 77.69 and by using the K-Nearest Neighbor method, the accuracy is 79.31. After comparison the results show, by using the K-Nearest Neighbor method, the highest accuracy is obtained. It concluded that the K-Nearest Neighbor method had better performance than the Decision Tree and Naive Bayes methods*

Keywords: Decision Tree, Naive Bayes, K-Nearest Neighbor, Student Performance

Intisari— Dalam pendidikan, kinerja siswa merupakan bagian yang penting. Untuk mencapai kinerja siswa yang baik dan berkualitas dibutuhkan analisa atau evaluasi terhadap faktor-faktor yang mempengaruhi kinerja siswa. Metode yang dilakukan masih menggunakan cara evaluasi berdasarkan hanya penilaian pendidik terhadap informasi kemajuan pembelajaran siswa. Cara tersebut tidak efektif karena informasi kemajuan pembelajaran siswa semacam itu tidak cukup untuk membentuk indikator dalam mengevaluasi kinerja siswa serta membantu para siswa dan pendidik untuk melakukan perbaikan dalam pembelajaran dan pengajaran. Penelitian-penelititan terdahulu telah dilakukan tetapi belum diketahui metode mana yang terbaik dalam mengklasifikasikan kinerja siswa. Pada penelitian ini dilakukan komparasi metode Decision Tree, Naive Bayes dan K-Nearest Neighbor dengan menggunakan dataset *student performance*. Dengan menggunakan metode Decision Tree didapatkan akurasi sebesar 78,85, dengan menggunakan metode Naive Bayes didapatkan akurasi sebesar 77,69 dan dengan menggunakan metode K-Nearest Neighbor didapatkan akurasi sebesar 79,31. Setelah dikomparasi hasil tersebut menunjukkan bahwa dengan menggunakan metode K-Nearest Neighbor didapatkan akurasi tertinggi. Hal tersebut menyimpulkan bahwa metode K-Nearest Neighbor memiliki kinerja yang lebih baik dibanding metode Decision Tree dan Naive Bayes.

Kata Kunci: Decision Tree, Naive Bayes, K-Nearest Neighbor, Kinerja Siswa

PENDAHULUAN

Kinerja siswa merupakan bagian penting dalam lembaga pendidikan (Shahiri, Husain, & Rashid, 2015). Tujuan utama dari lembaga pendidikan adalah untuk menyajikan pendidikan yang

berkualitas kepada siswanya sehingga dapat meningkatkan kinerja siswa (Hamsa, Indiradevi, & Kizhakkethottam, 2016). Untuk mencapai hal tersebut perlu dilakukan analisa atau evaluasi faktor-faktor yang mempengaruhi kinerja siswa. Evaluasi sangat penting untuk mempertahankan kinerja siswa dan efektivitas proses pembelajaran. Dengan menganalisis kinerja siswa, program strategis dapat direncanakan dengan baik selama masa studi mereka di sebuah lembaga pendidikan (Ibrahim & Rusli, 2007). Nilai akhir siswa didasarkan pada struktur mata pelajaran, tanda penilaian, nilai ujian akhir dan juga kegiatan ekstrakurikuler (Bin Mat, Buniyamin, Arsad, & Kassim, 2014). Prestasi siswa, kemajuan siswa dan potensi siswa sangat penting untuk mengukur hasil belajar, memilih bahan belajar dan kegiatan belajar (Yang & Li, 2018).

Untuk mengevaluasi kemajuan belajar siswa, penelitian yang ada telah mengembangkan banyak metode untuk memodelkan pengetahuan dan keterampilan siswa secara komprehensif (Yang & Li, 2018). Misalnya (Chen, Lin, & Chang, 2001), menerapkan konsep peta yang dikaitkan untuk mengekspresikan kedua pengetahuan yang diperoleh oleh seorang siswa setelah belajar suatu kegiatan pembelajaran dan pengetahuan prototipikal seorang pendidik. (Stecker, Fuchs, & Fuchs, 2005) mengusulkan pengukuran berbasis kurikulum untuk secara langsung memantau kemajuan belajar siswa yaitu pengetahuan atau keterampilan siswa selama periode waktu tertentu dan apakah kemajuan tersebut dapat memenuhi harapan guru. Hal tersebut menunjukkan bahwa cara mengevaluasi kinerja siswa masih berdasarkan hanya penilaian pendidik terhadap informasi kemajuan pembelajaran siswa. Cara tersebut tidak efektif karena informasi kemajuan pembelajaran siswa semacam itu tidak cukup untuk membentuk indikator yang membantu para siswa dan pendidik untuk melakukan perbaikan dalam pembelajaran dan pengajaran (Yang & Li, 2018).

Data mining adalah salah satu teknik yang paling populer untuk menganalisis kinerja siswa (Shahiri et al., 2015). Data mining telah banyak diterapkan dalam dunia pendidikan. Data mining pada dunia pendidikan adalah proses yang digunakan untuk mengekstraksi informasi dan pola yang berguna dari database pendidikan yang sangat besar (D. Magdalene Delighta Angeline, 2013). Sebagai hasilnya, itu akan membantu para pendidik dalam menyediakan suatu pendekatan pengajaran yang efektif (Shahiri et al., 2015). Selain itu, pendidik juga dapat memantau prestasi siswa mereka. Dalam memprediksi kinerja siswa, banyak penelitian telah dilakukan dengan teknik klasifikasi, seperti Decision Tree (Lopez Guarin, Guzman, & Gonzalez, 2015), Artificial Neural Networks (Alkhasawneh & Hobson, 2011), Support Vector Machine (Al-Shehri et al., 2017), Regression (Conijn, Snijders, Kleingeld, & Matzat, 2017), Naive Bayes (Lopez Guarin et al., 2015).

Di antara banyaknya metode klasifikasi, Decision Tree banyak digunakan karena sederhana secara teori dan hasilnya mudah dibaca (Lolli, Ishizaka, Gamberini, Balugani, & Rimini, 2017). Decision Tree merupakan algoritma yang kuat, populer, mudah ditafsirkan dan banyak diterapkan untuk beberapa masalah dalam data mining. Algoritma ini memberikan kinerja yang sangat baik dan mudah dimengerti. (Lakshmi, Indumathi, & Ravi, 2016).

Algoritma klasifikasi lain selain Decision Tree seperti Naive Bayes juga memberikan hasil yang menjanjikan untuk prediksi penyakit pada dataset medis tertentu seperti dataset penyakit jantung (Kumar & Sahoo, 2015). Naive Bayes dan Decision Tree juga telah dievaluasi oleh banyak peneliti dan ditemukan cocok untuk tingkat prediksi yang lebih baik pada data medis (Deverapalli, 2016).

Selain Decision Tree dan Naive Bayes, K-Nearest Neighbor adalah model nonparametrik intuitif dan efektif yang digunakan untuk tujuan klasifikasi dan regresi (Cover & Hart, 1967). K-Nearest Neighbor diklaim sebagai salah satu dari sepuluh algoritma data mining yang paling berpengaruh (X. Wu & Kumar, 2009). Karena keefektififannya, intuitif dan kesederhanaannya, K-Nearest Neighbor telah menarik minat luas dalam komunitas penelitian (Gou et al., 2014)(Lin, Li, Lin, & Chen, 2014)(Lin et al., 2014). Algoritma K-Nearest Neighbor adalah salah satu metode paling sederhana untuk memecahkan masalah klasifikasi, sering menghasilkan hasil yang kompetitif dan memiliki keuntungan yang signifikan atas beberapa metode penambangan data lainnya (Adeniyi, Wei, & Yongquan, 2016).

Metode Decision Tree, Naive Bayes dan K-Nearest Neighbor yang telah diterapkan pada penelitian-penelitian terdahulu mempunyai kelebihan masing-masing tetapi belum diketahui metode mana yang memiliki kinerja terbaik untuk mengklasifikasikan kinerja siswa. Untuk itu pada penelitan ini dilakukan komparasi antara metode Decision Tree, Naive Bayes dan K-Nearest Neighbor untuk klasifikasi kinerja siswa.

BAHAN DAN METODE

Bahan

Dalam penelitian ini digunakan dataset *student performance* yang didapat dari UCI Machine Learning Repository. Dataset tersebut 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)

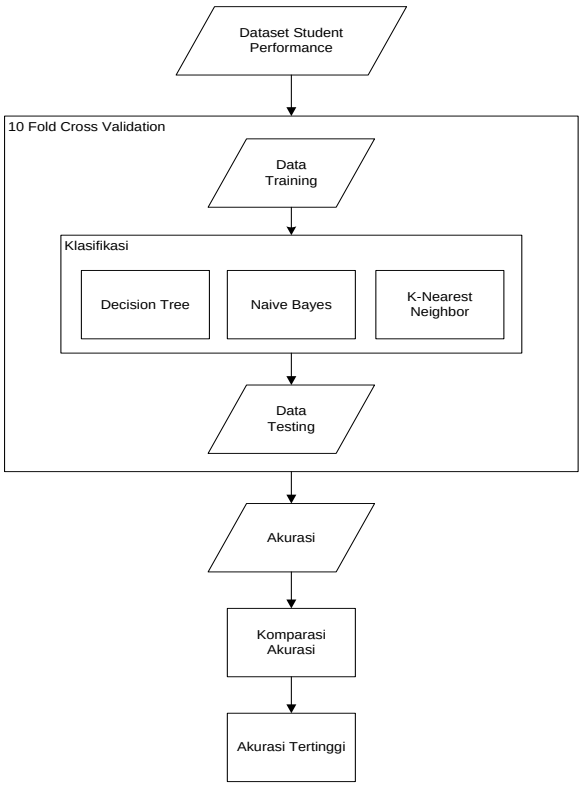
Tabel 2. Atribut, Data dan Keterangan Data pada Dataset *Student Performance*

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, 2018)

Gambar 1. Komparasi metode Decision Tree, Naive Bayes dan K-Nearest Neighbor

Pada penelitian ini dilakukan dengan mengkomparasi 3 metode yaitu Decision Tree, Naive Bayes dan K-Nearest Neighbor seperti pada Gambar 1. Proses yang dilakukan adalah *training* dataset *student performance* dengan menggunakan metode Decision Tree, Naive Bayes dan K-Nearest Neighbor untuk menghasilkan akurasi. Akurasi yang dihasilkan oleh ketiga metode tersebut kemudian dikomparasi untuk didapatkan akurasi tertinggi.

Decision Tree (DT)

Decision tree merupakan algoritma *supervised learning*, di mana algoritma tersebut membutuhkan data yang memiliki atribut kelas (Larose & Larose, 2014). Data tersebut harus beragam, jika datanya kurang maka proses klasifikasi tidak dapat dilakukan.

Salah satu aspek yang paling menarik dari Decision Tree adalah cara penyajian aturan yang terbentuk. Decision Tree adalah aturan keputusan khusus yang diatur ke dalam struktur pohon (Gries & Schneider, 2010). Aturan keputusan dapat dibangun dari pohon keputusan hanya dengan melintasi setiap jalur yang diberikan dari node root ke daun apa saja. Set lengkap aturan keputusan yang dihasilkan Pohon keputusan membagi ruang dokumen menjadi daerah yang tidak tumpang tindih di daunnya, dan prediksi dibuat di setiap daun (Gries & Schneider, 2010).

Algoritma Decision Tree dibentuk berdasarkan dataset. Pendekatan *divide and conquer* digunakan untuk membuat model Decision Tree dengan menggunakan IG untuk memilih atribut dari dataset dalam bentuk pohon (Lakshmi et al., 2016). Pada setiap langkah dalam membentuk Decision Tree, satu dari semua atribut dipilih untuk memisahkan data. Berdasarkan atribut yang terpilih, nilai pemisah ditentukan dengan menggunakan nilai atribut. IG dan entropi banyak digunakan untuk pohon klasifikasi. Untuk menghitung entropi didefinisikan (Shannon, 1948) sebagai berikut:

$$H_e(S) = 1 - \sum_{y \in C} p(y)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Di mana S adalah dataset, C adalah kelas dan $p(y)$ adalah perbandingan jumlah data terhadap kelas C. Baik entropi dan IG akan bernilai 0 apabila hanya ada 1 kelas dan mencapai nilai maksimum ketika semua kelas memiliki kemungkinan yang sama. Sedangkan IG dapat didefinisikan (Breiman, 2001) sebagai brikut:

$$G(r,S) = H(S) - \sum_{t} p(t)H(t) \dots\dots\dots\dots\dots\dots\dots (2)$$

Di mana r adalah aturan pemisahan dan t melambangkan node anak yang dipengaruhi oleh r pada dataset S di node induk. P(t) adalah perbandingan jumlah data yang berkaitan dengan t.

Naive Bayes (NB)

Naive Bayes adalah Bayesian network yang sederhana. Naive Bayes banyak digunakan untuk masalah klasifikasi (Zhang & Sheng, 2004). Klasifikasi merupakan aspek penting dari penambangan data. Dalam klasifikasi, pengklasifikasi dibuat dari contoh yang diberikan oleh label kelas. Setiap sampel E diwakili oleh vektor (a_1, a_2,..., a_n), di mana a_i adalah nilai atribut A_i dan A_1, A_2,..., A_n melambangkan n atribut. Pengklasifikasi memprediksi kemungkinan label pada data baru yang tidak berlabel.

Naive Bayes mudah dirancang, karena memiliki struktur yang sangat sederhana (X. Wu & Kumar, 2009). Prose *learning* pada Naive Bayes hanya menghitung probabilitas, secara spesifik, probabilitas bersyarat pada setiap atribut, dari data *training*. Ini berarti, nilai probabilitas p (a_i | c) harus ditentukan dari contoh data pelatihan, untuk setiap nilai a_i atribut A_i mempertimbangkan nilai variabel c pada kelas C.

Dalam Naive Bayes, diasumsikan bahwa atribut-atributnya bersifat independen satu sama lain terhadap *class* (J. Wu & Cai, 2011)(Turhan & Bener, 2009)(L. Jiang, Cai, & Wang, 2010). Setiap atribut hanya memiliki variabel *class* sebagai induknya (J. Wu & Cai, 2011)(Liangxiao Jiang, Wang, Cai, & Yan, 2007), P (E | c) dihitung oleh:

$$p(E|c) = p(a_1, a_2, \dots, a_n|c) = \prod_{i=1}^{n} p(a_i|c) \dots\dots (3)$$

Di mana p (a_i | c) mengacu pada probabilitas A_i, dan contohnya E = (a_1, a_2,..., a_n).

Karena setiap sampel E nilai p (E) adalah konstan, kemungkinan pembentukan label *class* adalah sebagai berikut:

$$p(c|E) = p(c) \prod_{i=1}^{n} p(a_i|c) \dots\dots\dots\dots\dots\dots\dots\dots (4)$$

Contoh E diklasifikasikan ke dalam kelas C = c' jika dan hanya jika:

$$p(c'|E) = \arg\max_{c} p(c|E) \qquad (2.6)$$

Lebih tepatnya, klasifikasi yang diberikan oleh Naive Bayes, dilambangkan dengan V_{nb} (E), didefinisikan sebagai berikut: (Zhang & Sheng, 2004)

$$\boldsymbol{V_{nb}(E)} = \boldsymbol{\arg\max_{c} p(c) \prod_{i=1}^{n} p(a_i|c)} \dots\dots\dots\dots\dots (5)$$

K-Nearest Neighbor (KNN)

KNN adalah algoritma *lazy learning* yang efektif dan kuat, meskipun mudah diimplementasikan. Namun, kinerjanya sangat bergantung pada kualitas data *training*. Karena banyaknya aplikasi nyata yang kompleks sehingga noise yang berasal dari berbagai sumber yang mungkin sering lazim dalam *database* skala besar (Liu & Zhang, 2012).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. KNN 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:

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{1i} - x_{2i})^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (6)$$

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

HASIL DAN PEMBAHASAN

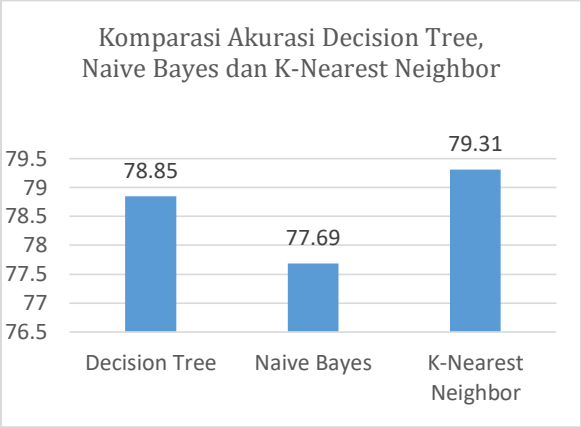
Tabel 3 merupakan komparasi metode NN, SVM dan KNN pada pengkasifikasian dataset *student performance*. Pada tabel 3 memperlihatkan dengan menggunakan metode DT didapatkan akurasi 78,85, dengan menggunakan metode NB didapatkan akurasi 77,69 dan dengan menggunakan metode KNN didapatkan akurasi 79,31. Komparasi akurasi dengan menggunakan metode DT, NB dan KNN dapat digambarkan dengan grafik pada Gambar 2.

Tabel 3. Komparasi Akurasi Metode NN dan LR

Metode	Akurasi
Decision Tree	78,85
Naive Bayes	77,69
K-Nearest Neighbor	79,31

Sumber: (Setiyorini & Asmono, 2018)

Dari hasil komparasi tersebut menunjukkan KNN memiliki tingkat akurasi yang paling tinggi. Hal tersebut menunjukkan bahwa kinerja K-Nearest Neighbor lebih baik dibanding dengan Decision Tree dan Naive Bayes. Hal ini membuktikan penelitan Adeniyi et al (Adeniyi et al., 2016) bahwa algortima klasifikasi KNN melakukan kinerja yang lebih baik daripada metode lain, termasuk Decision Tree dan Naive Bayes.



Sumber: (Setiyorini & Asmono, 2018)

Gambar 2. Komparasi Akurasi Metode Decision Tree, Naive Bayes dan K-Nearest Neighbor

KESIMPULAN

Dalam pendidikan, kinerja siswa merupakan bagian yang penting. Pada penelitian ini dilakukan komparasi metode Decision Tree, Naive Bayes dan K-Nearest Neighbor untuk mengklasifikasikan kinerja siswa dengan menggunakan dataset *student performance*. Hasil penelitian pada dataset *student performance* dengan menggunakan metode Decision Tree didapatkan akurasi 78,85, dengan menggunakan metode Naive Bayes didapatkan akurasi 77,69 dan dengan metode K-Nearest Neighbor didapatkan akurasi 79,31. Dari hasil tersebut dapat disimpulkan bahwa kinerja metode K-Nearest Neighbor lebih baik dibanding metode Decision Tree dan Naive Bayes.

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

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

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. A. (2014). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. *2013 IEEE 5th International Conference on Engineering Education: Aligning Engineering Education with Industrial Needs for Nation Development, ICEED 2013*, 126–130. https://doi.org/10.1109/ICEED.2013.6908316

Breiman, L. (2001). *Classification and regression tree*.

Chen, S. W., Lin, S. C., & Chang, K. E. (2001). Attributed concept maps: Fuzzy integration and fuzzy matching. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *31*(5), 842–852. https://doi.org/10.1109/3477.956047

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.

Cover, T., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

D. Magdalene Delighta Angeline. (2013). Association Rule Generation for Student Performance Analysis using Apriori Algorithm. *The SIJ Transactions on Computer Science Engineering & Its Applications (CSEA)*, *1*(1), 12–16.

Deverapalli, P. S. D. (2016). A Critical Study of

Classification Algorithms Using Diabetes Diagnosis. *2016 IEEE 6th International Conference on Advanced Computing (IACC)*.

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

Gries, D., & Schneider, F. B. (2010). *Texts in Computer Science*. *Media* (Vol. 42). https://doi.org/10.1007/978-1-84882-256-6

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

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/Daniela_Rusli/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression/links/0deec51bb04e76ed93000000.pdf

Jiang, L., Cai, Z., & Wang, D. (2010). IMPROVING NAIVE BAYES FOR CLASSIFICATION. *International Journal of Computers and Applications*, *32*(3). https://doi.org/10.2316/Journal.202.2010.3.202-2747

Jiang, L., Wang, D., Cai, Z., & Yan, X. (2007). Survey of Improving Naive Bayes for Classification. *Proceedings of the Third International Conference of Advanced Data Mining and Applications*, *4632*, 134–145. https://doi.org/10.1007/978-3-540-73871-8_14

Kumar, S., & Sahoo, G. (2015). Classification of heart disease using Naïve Bayes and genetic algorithm. *Smart Innovation, Systems and Technologies*. https://doi.org/10.1007/978-81-322-2208-8_25

Lakshmi, B. N., Indumathi, T. S., & Ravi, N. (2016). A Study on C.5 Decision Tree Classification Algorithm for Risk Predictions During Pregnancy. *Procedia Technology*, *24*, 1542–1549. https://doi.org/10.1016/j.protcy.2016.05.128

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data*. https://doi.org/10.1002/9781118874059

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

Liu, H., & Zhang, S. (2012). Noisy data elimination using mutual k-nearest neighbor for classification mining. *Journal of Systems and Software*, *85*(5), 1067–1074. https://doi.org/10.1016/j.jss.2011.12.019

Lolli, F., Ishizaka, A., Gamberini, R., Balugani, E., & Rimini, B. (2017). Decision Trees for Supervised Multi-criteria Inventory Classification. *Procedia Manufacturing*, *11*(June), 1871–1881. https://doi.org/10.1016/j.promfg.2017.07.326

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

Setiyorini, T., & Asmono, R. T. (2018). Laporan Akhir Penelitian Mandiri. Jakarta: STMIK Nusa Mandiri

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

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Labs Technical Journal*, *27*(3), 379–423.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Turhan, B., & Bener, A. (2009). Analysis of Naive Bayes' assumptions on software fault data: An empirical study. *Data and Knowledge Engineering*, *68*(2), 278–290. https://doi.org/10.1016/j.datak.2008.10.005

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

Wu, J., & Cai, Z. (2011). Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB). *Journal of Computational Information Systems*, *5*(5), 1672–1679.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. *Physics of Fluids*. https://doi.org/10.1063/1.2756553

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, H., & Sheng, S. (2004). Learning weighted naive bayes with accurate ranking. In *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004* (pp. 567–570). https://doi.org/10.1109/ICDM.2004.10030