e-ISSN: 2655-142X
p-ISSN: 2655-190X

PERBANDINGAN ALGORITME KLASIFIKASI UNTUK PREDIKSI KINERJA SISWA DI KELAS

Anna Baita [1] , Yoga Pristyanto [2] , Irfan Pratama [3]

[1] Informatika Universitas AMIKOM Yogyakarta
[2] Sistem Informasi Universitas AMIKOM Yogyakarta
[3] Sistem Informasi Universitas Mercubuana Yogyakarta
email : anna@amikom.ac.id[1], yoga.pristyanto@amikom.ac.id[2], irfanp@mercubuana-yogya.ac.id[3]

Abstraksi

Pendidikan memiliki peran penting dalam meningkatkan kualitas hidup masyarakat di suatu negara. Pendidikan terbagi menjadi tiga yaitu formal, non formal dan informal. Mayoritas pendidikan formal diselenggarakan melalui kelas konvensional. Dalam kelas konvensional, jumlah siswa yang banyak dapat menyebabkan materi tidak dapat tersampaikan dengan baik. Oleh karena itu diperlukan adanya pengelompokan siswa berdasarkan kemampuan belajarnya. Teknik *data mining* dengan metode klasifikasi diusulkan untuk memprediksi kinerja siswa di kelas. Hasil klasifikasi siswa dapat digunakan sebagai acuan dalam memberikan materi sesuai dengan kemampuan belajarnya. Dalam penelitian ini diusulkan perbandingan algoritme klasifikasi yang cukup populer yaitu k-Nearest Neighbor, Support Vector Machine, dan Naive Bayes. Berdasarkan hasil yang didapatkan, algoritme Support Vector Machine memiliki performa yang lebih baik dibandingkan dua algoritme lainnya.

***Kata Kunci* :**
Educational Data Mining, Classification, Students Classification, Students Performance.

Abstract

Education has an important role to improve the people's life quality in a country. In Indonesia, majority of formal education was held through a conventional classroom. In a conventional classroom, the large number of students cause learning material cannot be conveyed properly. Therefore, it is necessary to group students based on learning abilities. Classification method is proposed to handling these problems. In this study, we propose a comparison of three popular educational data mining methods. They are k-Nearest Neighbor (k-NN), Naive Bayes (NB), and Support Vector Machine (SVM). Those algorithms are very good at handling binary classification problems with multivariate kind of data. The experimental results show that Support Vector Machine (SVM) has the highest with 85.71%, compared to Naive Bayes (NB) and k-Nearest Neighbor (k-NN) with 81.91% and 83.81% respectively. Based on the result, Support Vector Machine (SVM) is expected to be used for grouping students appropriately and it can be used as a reference in providing student learning materials based on their group..

***Keywords* :**
Educational Data Mining, Classification, Students Classification, Students Performance.

Pendahuluan

Pendidikan merupakan salah satu faktor penting pada setiap negara dalam meningkatkan kualitas hidup masyarakatnya. Di beberapa negara, pemerintah memiliki peran penting dalam meningkatkan kualitas pendidikan [1]. Di Indonesia, terdapat tiga satuan proses penyelenggaraan pendidikan yang telah diatur didalam Undang-Undang yaitu formal, non formal, dan informal [2]. Pendidikan formal diselenggarakan melalui kelas kovensional di sekolah. Salah satu permasalahan yang terjadi dalam kelas konvensional ialah jumlah siswa yang sangat banyak. Hal ini menyebabkan materi pembelajaran tidak dapat tersampaikan dengan baik, karena setiap siswa memiliki metode belajar masing-masing [3]. Jika para pengajar memahami metode belajar yang dimiliki oleh siswa, maka para pengajar dapat mempersiapkan materi pembelajaran yang tepat untuk masing-masing siswa berdasarkan tingkat pengetahuan mereka. Hal ini dapat meningkatkan kemampuan serta menjadikan proses pembelajaran lebih efisien [1]. Oleh karena itu diperlukan adanya pengelompokan siswa berdasarkan metode belajarnya untuk memprediksi kinerjanya selama belajar di kelas [3].

Data mining merupakan salah satu metode yang dapat digunakan dalam pengelompokan siswa. Data mining merupakan sebuah konsep untuk mengenali pola yang tersembunyi dan menemukan relasi antar parameter didalam data dengan jumlah yang besar [4]. Data mining banyak digunakan dalam berbagai bidang seperti medis, teknik, ekonomi, keuangan, dan pendidikan. Hal tersebut menunjukan bahwa data mining dapat memberikan solusi alternatif bagi para pengambil keputusan dalam memecahkan

masalah tertentu. Eksplorasi data yang dilakukan pada bidang pendidikan sering dikenal dengan istilah *Educational Data Mining* (EDM) [5]. EDM merupakan proses yang digunakan untuk mengekstrak informasi yang berguna dan mengenali pola dari database pendidikan dengan jumlah yang besar [6].

Dalam EDM terdapat beberapa metode yang digunakan dalam pengelompokan siswa, salah satunya adalah metode klasifikasi. Klasifikasi merupakan proses untuk menemukan sebuah model atau pola yang dapat menggambarkan serta membedakan kelas pada suatu dataset. Tujuannya agar model tersebut dapat digunakan untuk memprediksi obyek dengan label kelas yang tidak diketahui. Model tersebut didasarkan pada analisis data latih. Model dari hasil klasifikasi dapat dimanfaatkan untuk memprediksi tren data masa depan [7].

Dalam EDM terdapat beberapa algoritme klasifikasi yang populer. Dalam penelitian yang dilakukan oleh Ahmad et al. [5] membandingan algoritme Decision Tree, Rule Based dan Naive Bayes dalam mengklasifikasikan dataset siswa dengan jumlah yang kecil. Pada penelitian tersebut menyebutkan algoritme Naive Bayes memiliki tingkat akurasi yang paling baik dalam klasifikasi. Selain itu pada penelitian yang dilakukan oleh Jishan et al [8], menggunakan algoritme Naive Bayes dalam mengelompokan siswa untuk mengetahui dan membantu mahasiswa yang lemah dalam belajar.

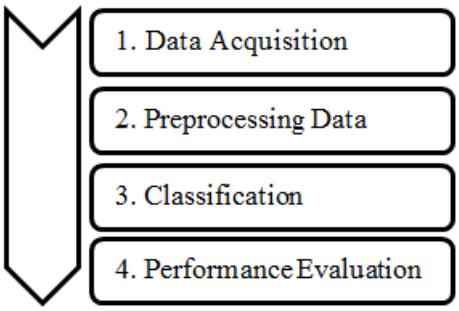
Mayilvaganan et al [9] menggunakan algoritme k-Nearest Neighbor (kNN) dalam mengelompokan siswa untuk memprediksi kinerjanya di kelas berdasarkan kemampuan kognitifnya. Sedangkan Gray et al [10] dalam penelitiannya menyatakan bahwa algoritme k-Nearest Neighbor (kNN) memiliki kinerja yang baik dalam menangani klasifikasi dengan dataset jumlah kecil.

Pada penelitian yang dilakukan Worapat et al [1] menggunakan algoritme Support Vector Machine (SVM) dalam mengelompokan siswa untuk memprediksi kemampuan belajarnya dikelas. Selain itu Sembiring et al [11] dalam penelitiannya menyatakan bahwa algoritme Support Vector Machine (SVM) memiliki kemampuan yang sangat baik dalam menangani klasifikasi dua kelas.

Berdasarkan uraian diatas, pada penelitian ini akan dilakukan perbandingan algoritme Naive Bayes, k-Nearest Neighbor (kNN) dan Support Vector Machine (SVM) untuk mengetahui kinerja dari ketiga algoritme tersebut dalam memprediksi kinerja siswa di kelas. Sehingga akan diperoleh satu algoritma klasifikasi yang memiliki kinerja paling baik untuk mnangani kasus ini.

Metode Penelitian

Gambar 1 merupakan merupakan diagram alur tahapan klasifikasi siswa.



Gambar 1. Diagram Alur Tahapan Klasifikasi Siswa

Data Acquisition

Data acquisition merupakan proses untuk mempersiapkan data yang akan digunakan pada penelitian, pada tahap ini data yang akan digunakan merupakan Dataset *private*. Data yang diambil dari hasil ujian mata kuliah struktur data pada sebuah Universitas di Indonesia. Dataset tersebut merupakan data numerik dengan jumlah *instance* sebanyak 105. Masing-masing *instance* terdiri dari 10 attribute dan 1 label atau kelas. *Attribute* menunjukkan karakteristik pada data. Sedangkan *label* merupakan target kelas pada data, dimana pada datasaet ini berisi *label* "Berhasil" sebanyak 27 *instance* dan label "Tidak Berhasil" sebanyak 78 *instance*

Preprocessing Data

Pada tahap *preproccesing data* dilakukan pembersihan data. Proses pembersihan meliputi pengisian data yang kosong, menghilangkan duplikasi data, memeriksa inkonsistensi data, dan memperbaiki kesalahan pada data. Biasanya data yang kosong disebabkan oleh adanya data baru yang belum ada informasinya [12].

Pada data *pretest* mahasiswa yang dilakukan terdapat beberapa *missing value* atau data yang kosong. *Missing value* tersebut disebabkan oleh adanya beberapa mahasiswa yang hanya mengikuti *pretest* satu kali. Oleh karena itu data mahasiswa yang akan hanya mengikuti satu kali pretest diisi dengan nilai "0".

Classification

Pada tahap ini akan fokus untuk membandingkan 3 algoritme klasifikasi yaitu k-Nearest Neighbor, Support Vector Machine, dan Naive Bayes.

Evaluation

Pada penelitian metode *confusion matrix* akan digunakan dalam pengujian kinerja algoritme. Metode ini menggunakan tabel matriks seperti pada Tabel 1, jika dataset hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai kelas positif dan yang lainnya sebagai kelas negatif [13]. Evaluasi dengan *confusion matrix* akan menghasilkan nilai

e-ISSN: 2655-142X
p-ISSN: 2655-190X

accuracy, *precison*, dan *recall*. Akurasi dalam klasifikasi merupakan persentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi [14]. *Precision* atau *confidence* adalah proporsi kasus dengan hasil positif yang benar. *Recall* atau *sensitivity* merupakan proporsi kasus positif yang diidentifikasi dengan benar [15].

TABEL 1 CONFUSION MATRIX [16]

Actual Classification	*Prediction Classificatiom*	
	Positif	Negatif
Positif	True Positif	False Negatif
Negatif	False Positif	True Negatif

True Positive adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false* positif adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false* negatif adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true* negatif adalah jumlah *record* negatif yang diklasifikasikan sebagai negatif. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *precision*, dan accuracy. Untuk menghitung nilai-nilai tersebut digunakan persamaan dibawah ini [14].

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN+FP)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Sensitivity\ (recall) = \frac{TP}{(TP+FN)} \quad (3)$$

$$F - Measure = 2\ \frac{Precision\ x Recall}{Precision + Recall} \quad (4)$$

Hasil dan Pembahasan

Dalam penelitian ini, kami akan melakukan perbandingan kinerja dari tiga algoritme klasifikasi. Ketiga algoritme tersebut ialah k-Nearest Neighbor, Support Vector Machine, dan Naive Bayes. Alat bantu yang digunakan ialah WEKA v.3.8.

Untuk validasi digunakan metode 5-*fold cross validation*. *K-fold Cross validation* merupakan sebuah metode statistika dengan membagi data menjadi dua bagian yaitu data data *training* dan data *testing* [14].

Berikut ini tabel 2 *confusion matrix hasil* dari pengolahan data menggunakan algoritme k-Nearest Neighbor, Support Vector Machine dan Naive Bayes dengan medium sebagai positif *class* dan low sebagai negatif *class*.

TABEL 2 CONFUSION MATRIX K-NEAREST NEIGHBOR

Actual Classification	*Prediction Classificatiom*	
	Low	Medium
Low	69	9
Medium	8	19

Dari table 2 *confusion matrix* algoritme k-Nearest Neighbor diatas maka dapat dihitung nilai akurasi, presisi, *sensitivity* dan *f-measure* dengan menggunakan persamaan (3), (4) (5) dan (6). Tabel 3 menunjukan hasil kinerja klasifikasi algoritme kNN.

TABEL 3 KINERJA K-NEAREST NEIGHBOR

Akurasi (%)	Sensitifity (%)	Presisi (%)	F-Measure (%)
83,81	83,8	84,0	83,9

TABEL 4 CONFUSION MATRIX SUPPORT VECTOR MACHINE

Actual Classification	*Prediction Classificatiom*	
	Low	Medium
Low	74	4
Medium	11	16

Dari tabel 4 *confusion matrix* algoritme Support Vector Machine diatas maka dapat dihitung nilai akurasi, presisi, *sensitivity* dan *f-measure* dengan menggunakan persamaan (1), (2) (3) dan (4). Tabel 5 menunjukan hasil kinerja klasifikasi algoritme kNN.

TABEL 5 KINERJA SUPPORT VECTOR MACHINE

Akurasi (%)	Sensitifity (%)	Presisi (%)	F-Measure (%)
85,71	85,7	85,2	85

TABEL 6 CONFUSION MATRIX NAIVE BAYES

Actual Classification	*Prediction Classificatiom*	
	Low	Medium
Low	65	13
Medium	6	21

Dari tabel 6 *confusion matrix* algoritme Naive Bayes diatas maka dapat dihitung nilai akurasi, presisi, *sensitivity* dan *f-measure* dengan menggunakan persamaan (1), (2) (3) dan (4). Tabel 7 menunjukan hasil kinerja klasifikasi algoritme kNN.

TABEL 7 KINERJA NAIVE BAYES

Akurasi (%)	Sensitifity (%)	Presisi (%)	F-Measure (%)
81,91	81,9	83,9	82,5

Berikut ini Tabel 8 menunjukan perbandingan tingkat akurasi, presisi, *sensitivity* dan *f-measure* hasil klasifikasi dari ketiga algoritme. Algoritme SVM memiliki kinerja yang paling baik dari sisi akurasi, presisi, *sensitifity* dan *f-measure* dibandingkan algoritme kNN dan Naive Bayes.

TABEL 8 PERBANDINGAN KINERJA ALGORITME

Algoritme	Akurasi (%)	Sensitifity (%)	Presisi (%)	F-Measure (%)
kNN	83,81	83,8	84,0	83,9
Naive Bayes	81,91	81,9	83,9	82,5
SVM	85,71	85,7	85,2	85

Kesimpulan dan Saran

Berdasarkan hasil dari pembahasan, perbandingan antara algoritme kNN, SVM, dan Naive Bayes telah berhasil dilakukan, terlihat pada Tabel VIII. Secara keseluruhan kinerja algoritme SVM lebih baik dibandingkan dengan algoritme kNN dan Naive Bayes dalam memprediksi kinerja siswa di kelas. Selain itu algoritme SVM juga memiliki kinerja yang lebih baik dibandingkan algoritme k-NN dan Naive Bayes dalam menangani data set kecil dan klasifikasi dua kelas. Hasil klasifikasi dapat digunakan sebagai acuan dalam memberikan materi belajar siswa sesuai dengan pengelompokannya. Untuk penelitian selanjutnya dengan permasalahan yang sama, akan dilakukan penanganan terhadap imbalance pada dataset yang mungkin dapat memberikan hasil yang lebih baik.

Daftar Pustaka

[1] W. Paireekreng and T. Prexawanprasut, "An integrated model for learning style classification in university students using data mining techniques," *2015 12th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol.*, pp. 1–5, 2015.

[2] Presiden RI, *UU No 12 Thn 2012 ttg Pendidikan Tinggi*. 2012.

[3] I. Hidayah, A. E. Permanasari, and N. Ratwastuti, "Student classification for academic performance prediction using neuro fuzzy in a conventional classroom," *Inf. Technol. Electr. Eng. (ICITEE), 2013 Int. Conf.*, pp. 221–225, 2013.

[4] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[5] Fadhilah Ahmad, N. H. Ismail, and Azwa Abdul Aziz, "The prediction of students' academic performance using classification data mining techniques," *Appl. Math. Sci.*, vol. 9, no. 129, pp. 6415–6426, 2015.

[6] C. Bambrah, M. Bhandari, N. Maniar, and V. Munde, "Mining Association Rules in Student Assessment Data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 3, pp. 5340–5342, 2014.

[7] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," in *Proceedings of 2014 International Conference on Data and Software Engineering*, 2014, pp. 1–5.

[8] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015.

[9] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," in *Communication and Network Technologies (ICCNT), 2014 International Conference on Computational Intelligence and Computer Research*, 2014, pp. 113–118.

[10] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," *2014 4th IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 549–554, 2014.

[11] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance By an Application of Data Mining Techniques.," *Int. Conf. Manag. Artif. Intell.*, vol. 6, pp. 110–114, 2011.

[12] O. N. Pratiwi, "Predicting student placement class using data mining," in *Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2013*, 2013, no. August, pp. 618–621.

[13] B. Max, *Principles of Data Mining*. London: Springer, 2007.

[14] M. Han, J., & Kamber, *Data Mining: Concepts and Techniques Second*, Second Edi., vol. 12. San Fransisco: Morgan Kauffman, 2006.

[15] D. M. W. Powers, "Evaluation: From Precision, Recall And F-Measure To ROC, Informedness, Markedness & Correlation," vol. 2, no. 1, pp. 37–63, 2011.

[16] Jiawei Han and Micheline Kamber, *Jiawei Han & Micheline Kamber*, Second Edi. San Francisco: Morgan Kaufmann Publishers, 2006.