Integrasi Metode Naive Bayes dengan K-Means dan K-Means-Smote untuk Klasifikasi Jurusan SMAN 3 Mataram

Hairani Hairani[1], Muhammad Ridho Hansyah[2], Lalu Zazuli Azhar Mardedi[3]
Universitas Bumigora
e-mail: [1]hairani@universitasbumigora.ac.id, [2]hansyahridho815@gmail.com, [3]zazuli@stmikbumigora.com

Diajukan: 22 Mei 2020; Direvisi: 21 Juli 2020; Diterima: 23 September 2020

Abstrak

Pihak SMAN 3 Mataram memiliki permasalahan yaitu kesulitan untuk memilihkan jurusan yang tepat bagi siswanya, karena tidak ada sistem yang memberi keputusan jurusan yang sesuai dengan minat dan bakat siswa, serta dibatasi dengan jumlah kuota di tiap kelasnya. Tujuan dari penelitian ini adalah integrasi metode Naive Bayes dengan K-Means dan K-Means-Smote untuk klasifikasi penjurusan SMAN 3 Mataram. Metodologi penelitian ini terdiri dari pengumpulan data siswa, pengolahan data, pengujian metode, dan evaluasi kinerja metode yang diusulkan. Berdasarkan hasil pengujian yang telah dilakukan, metode yang diusulkan memperoleh kinerja terbaik dibandingkan penelitian sebelumnya menggunakan metode C.45 dengan akurasi sebesar 99,16%, sensitivitas 99,58%, spesifisitas 98,77%, dan f-measure 99,16%. Dengan demikian metode yang diusulkan dapat digunakan untuk klasifikasi jurusan SMAN 3 Mataram karena memiliki kinerja paling baik.

Kata kunci: *Naive Bayes, Jurusan SMA, K-Means, K-Means-Smote, Data Mining.*

Abstract

SMAN 3 Mataram has a problem which is the difficulty in choosing the right majors for their students, because there is no system that determines majors in accordance with the interests and talents of students, and is limited by the amount of quota in each class. The purpose of this study is the integration of the Naive Bayes method with K-Means and K-Means-Smote for the classification of SMAN 3 Mataram majors. The research methodology consisted of collecting student data, processing data, testing methods, and evaluating the performance of the proposed method. Based on the results of tests that have been done, the proposed method obtained the best performance compared to previous studies using C4.5 method with an accuracy of 99.16%, sensitivity 99,58%, specificity 98,77%, and f-measure 99,16%. Thus the proposed method can be used for SMAN 3 Mataram majors classification because it has the best performance.

Keywords*: Naive Bayes, Jurusan SMA, K-Means, K-Means-Smote, Data Mining.*

1. Pendahuluan

Pemilihan jurusan di SMAN 3 Mataram masih menggunakan *check list* yang diisi langsung oleh siswa dalam pemilihan jurusan yang diminati. Pemilihan jurusan mempunyai masalah baik pada siswa dan sekolah. Permasalahan pada siswa terjadi kebingungan untuk memilih jurusan mana yang dipilih. Kebingungan memilih jurusan ini juga akan menyebabkan dilema bagi siswa di kemudian hari, apakah jurusan yang mereka pilih telah tepat sesuai dengan minat dan bakat mereka. Siswa yang salah memilih jurusan juga akan mengalami penurunan nilai karena tidak sesuai dengan bakat. Pihak sekolah sendiri mempunyai masalah untuk memilihkan jurusan yang tepat, karena tidak ada sistem yang memberi keputusan jurusan yang sesuai dengan minat dan bakat siswa, serta dibatasi dengan jumlah kuota di tiap kelasnya. Pemilihan jurusan yang salah juga akan berpengaruh pada saat siswa akan melanjutkan ke jenjang Pendidikan berikutnya. Silang jurusan antara siswa jurusan IPA dan IPS sering terjadi karena salah dalam memilih jurusan di awal masuk SMA, yang harusnya siswa dengan jurusan IPA menempuh perkuliahan sesuai dengan jurusannya, tetapi berpindah ke jurusan yang seharusnya di ambil oleh siswa jurusan IPS, dan begitu pula sebaliknya. Solusi yang digunakan pada penelitian ini untuk menyelesaikan permasalahan tersebut adalah konsep *data mining*. *Data mining* merupakan proses penemuan pola dan tren yang berharga dari data yang besar [1]. Beberapa penelitian tentang penjurusan siswa di SMA sudah dilakukan berbasis sistem pendukung keputusan [2]–[6] dan *data mining* [7]–[9] di antaranya adalah penelitian tentang penjurusan siswa SMA Negeri 5 Kediri menggunakan metode k-NN (k-Nearest Neighbor) [10]. Parameter

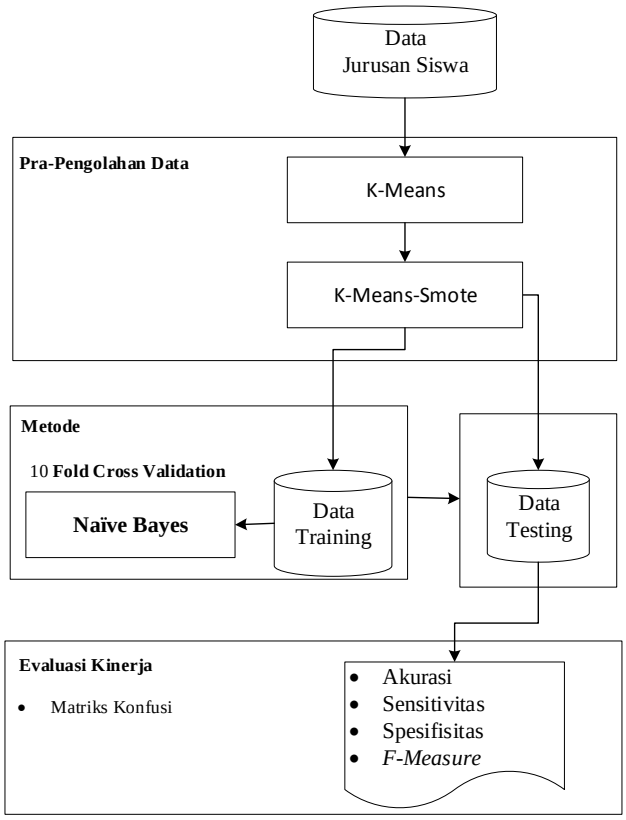
yang digunakan pada penelitian tersebut adalah Matematika, Biologi, Kimia, Fisika, Ekonomi, Sejarah, Geografi, dan Sosiologi. Kelemahan penelitian tersebut tidak disebutkan kinerja metode yang digunakan. Penelitian tentang implementasi metode K-Means untuk penjurusan siswa SMA Negeri 1 Pangkalan Kerinci [11]. Parameter yang digunakan penelitian tersebut adalah Matematika, Fisika, Kimia, Biologi, Geografi, Sosiologi, Ekonomi, dan Sejarah. Kelemahan penelitian tersebut tidak disebutkan kinerja metode yang digunakan.

Penelitian tentang implementasi metode k-Neares Neihghbor (k-NN) untuk penjurusan siswa SMA Negeri 1 Praya [12]. Parameter yang digunakan penelitian tersebut adalah jenis kelamin, asal sekolah, nilai UN Bahasa Indonesia, nilai UN Bahasa Inggris, nilai UN Matematika, nilai UN IPA, dan Bahasa Indonesia, Bahasa Inggris, Matematika, IPA, IPS, dan kelas. Kelemahan penelitian tersebut tidak disebutkan kinerja metode yang digunakan. Penelitian lain menggunakan metode C4.5 untuk penjurusan siswa SMA Negeri 1 Pontianak [13]. Parameter yang digunakan penelitian tersebut adalah nilai tes akademik Matematika, IPA, IPS, nilai rata-rata rapor SMP untuk mata pelajaran Matematika, IPA, IPS, nilai UN SMP untuk mata pelajaran Matematika, IPA, dan minat siswa. Metode C4.5 memiliki akurasi sebesar 89.79% untuk klasifikasi jurusan siswa. Kelemahan penelitian tersebut tidak menyelesaikan permasalahan ketidakseimbangan kelas pada data yang digunakan.

Penelitian ini bertujuan untuk integrasi metode Naive Bayes dengan K-Means dan K-Means-Smote untuk klasifikasi jurusan siswa. Metode K-Means digunakan untuk pelabelan jurusan IPA dan IPS. Sedangkan metode K-Means-Smote digunakan untuk menyeimbangkan kelas agar jumlah data jurusan IPA dan IPS menjadi seimbang, sehingga metode Naive Bayes dapat mengklasifikasikannya dengan baik berdasarkan akurasi, sensitivitas, spesifisitas, dan *f-measure*.

2. Metode Penelitian

Penelitian ini terdiri dari beberapa tahapan seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian.

2.1. Pengumpulan Data

Data siswa yang digunakan dalam penelitian ini adalah data siswa SMAN 3 Mataram kelas 1 yang berjumlah 432 *instances*. Atribut data yang digunakan penelitian ini adalah mata pelajaran Bahasa

Indonesia, Bahasa Inggris, Matematika, IPA, dan IPS. Adapun contoh data siswa SMA Negeri 3 Mataram yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Contoh data siswa SMA Negeri 3 Mataram kelas 1.

No	Bahasa Indonesia	Bahasa Inggris	Matematika	IPA	IPS
1	87	90	85	87	85
2	93	86	87	88	88
3	89	92	92	94	88
4	82	84	82	82	81
5	90	93	88	89	88
6	90	85	86	84	85
7	93	90	91	92	90
8	87	84	81	82	83
9	91	90	89	90	90
10	91	86	90	89	90
11	90	89	91	88	91
12	89	92	92	87	90
13	88	89	94	95	90
14	92	90	94	94	90
15	91	89	93	91	91

2.2. Pra Pengolahan Data

Pra pengolahan data digunakan untuk menghasilkan data yang berkualitas. Metode K-Means digunakan untuk pelabelan jurusan (IPA atau IPS) dari data siswa. Hasil pelabelan penjurusan siswa menggunakan metode K-Means, terdapat jurusan IPA sebanyak 191 *instances* dan IPS 241 *instances*. Terdapat ketidakseimbangan kelas antara jurusan IPA dengan IPS. Untuk menyeimbangkan distribusi kelas jurusannya, penelitian ini menggunakan K-Means-Smote. K-Means-Smote merupakan algoritma modifikasi metode Smote berbasis pengelompokan menggunakan metode K-Means [14][15].

2.3. Metode Klasifikasi

Metode klasifikasi yang digunakan untuk klasifikasi jurusan siswa SMA adalah metode Naive Bayes. Metode Naive Bayes merupakah sebuah metode klasifikasi berbasis statistik berdasarkan teorema Bayes. Metode Naive Bayes diasumsikan sangat kuat (naif) akan indepedensi antara atribut dengan kelasnya. Formula metode Naive Bayes ditunjukkan persamaan (1) [16], [17].

$$P(B|A) = \frac{P(A|B)*P(B)}{P(A)} \quad (1)$$

Dengan :
$P(B|A)$ = Probabilitas hipotesis B berdasarkan kejadian A.
$P(A|B)$ = Probabilitas kejadian A berdasarkan hipotesis B.
$P(B)$ = Probabilitas awal hipotesis B.
$P(A)$ = Probabilitas kejadian A.

3. Hasil dan Pembahasan

Pada penelitian ini mencoba mengombinasikan metode Naive Bayes dengan metode K-Means dan K-Means-Smote untuk penjurusan siswa SMA Negeri 3 Mataram. Metode K-Means digunakan untuk pelabelan jurusan (IPA dan IPS) pada data siswa yang digunakan. Hasil pelabelan jurusan menggunakan K-Means yaitu IPA sebanyak 191 *instances* dan IPS 241 *instances* sehingga terdapat ketidakseimbangan kelas antara jurusan IPA dengan IPS. Penelitian ini menggunakan K-Means-Smote untuk menyeimbangkan kelasnya sehingga menjadi seimbang. Hasil penyeimbangan kelas menggunakan K-Means-Smote adalah jurusan IPA sebanyak 238 *instances* dan jurusan IPS 241 *instances*. Pengujian metode yang diusulkan menggunakan validasi 10 *fold cros validation* dan kinerjanya diukur menggunakan tabel matrik konfusi berdasarkan akurasi, sensitivitas, spesifisitas, dan *f-measure*. Hasil pengujian metode usulan dibandingkan dengan metode penelitian sebelumnya yang menggunakan metode C4.5 dengan data yang sama ditunjukkan pada Tabel 2.

Tabel 2. Matrik konfusi metode usulan.

Prediksi	Aktual	
	IPA	**IPS**
IPA	235 (TP)	3 (FP)
IPS	1 (FN)	240 (TN)

$$Akurasi = \frac{235\ + 240}{235 +\ 3\ + 240\ + 1} = \frac{475}{479} = \mathbf{99.16}\%$$

$$Sensitivitas\ (Recall) = \frac{235}{235\ +\ 1} = \frac{235}{236} = \mathbf{99.58}\%$$

$$Spesifisitas = \frac{240}{240\ +\ 3} = \frac{240}{243} = \mathbf{98.77}\%$$

$$Precision = \frac{235}{235\ +\ 3} = \frac{235}{238} = \mathbf{98.74}\%$$

$$F - measure = 2 * \left(\frac{98.74\ *\ 99.58}{98.74\ +\ 99.58}\right) = \mathbf{99.16}\%$$

Tabel 3. Hasil perbandingan kinerja metode usulan dengan metode C4.5.

Metode	Kinerja			
	Akurasi	**Senistivitas**	**Spesifisitas**	***F-Measure***
Metode C4.5	91.89%	91.48%	92.21%	90.76%
Metode Naive Bayes	98.84%	98.95%	98.76%	98.69%
Usulan Metode (*K-Means-Smote-K-Means- Naive Bayes*)	99.16%	99.58%	98.77%	99.16%

Berdasarkan Tabel 2 menunjukkan bahwa metode yang diusulkan memiliki kinerja paling baik dibandingkan dengan penelitian sebelumnya menggunakan metode C4.5 dan Naive Bayes berdasarkan akurasi, sensitivitas, spesifisitas, dan *f-measure*. Hasil pengujian yang telah dilakukan, mendapatkan akurasi sebesar 99.16%, sensitivitas 99.58%, spesifisitas 98.77%, dan *f-measure* 99.16%. Metode yang diusulkan memiliki kinerja paling baik dengan penelitian sebelumnya menggunakan metode C4.5, dikarenakan pada penelitian ini menggunakan metode K-Means-Smote untuk menyeimbangkan kelas pada data siswa yang digunakan. Oleh karena itu, kombinasi metode Naive Bayes dengan K-Means dan K-Means-Smote dapat digunakan untuk penjurusan siswa SMA Negeri 3 Mataram.

4. Kesimpulan

Integrasi metode Naive Bayes dengan K-Means dan K-Means-Smote untuk klasifikasi jurusan siswa SMA Negeri 3 Mataram menunjukkan peningkatan kinerjanya. Metode K-Means digunakan untuk pelabelan jurusan, metode K-Means-Smote untuk menyeimbangkan kelasnya, dan metode Naive Bayes digunakan sebagai metode klasifikasinya. Metode yang diusulkan memiliki kinerja lebih baik dibandingkan penelitian sebelumnya menggunakan metode C4.5 dan Naive Bayes, hal ini ditunjukkan berdasarkan hasil pengujian yang telah dilakukan yaitu akurasi 99.16%, sensitivitas 99.58%, spesifisitas 98.77%, dan *f-measure* 99.16%. Untuk penelitian selanjutnya dapat menggunakan metode Smote untuk menyeimbangkan kelasnya, dan metode klasifikasi yang lain seperti *Support Vector Machine* (SVM), Random Forest, dan Gradient Boosting Tree (GBT).

Daftar Pustaka

[1] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data An Introduction to Data Mining*, Second Edi. Hoboken: Jhon Wiley & Sons, 2014.

[2] Y. S. Nugroho, "Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 1, no. 1, p. 1, 2015, doi: 10.23917/khif.v1i1.1175.

[3] D. A. Pertiwi, B. Daniawan, and Y. Gunawan, "Analysis And Design of Decision Support System in Major Assignment at Buddhi High School Using AHP and SAW Methods," *J. Tech-e*, vol. 3, no. 1, pp. 14–21, 2019.

[4] F. Frieyadie and S. M. Ramadhan, "Penerapan Metode AHP Untuk Membantu Siswa Memilih Jurusan Yang Tepat Di SMK," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 3, pp. 662–667, 2018, doi: 10.29207/resti.v2i3.396.

[5] B. S. Prayoga and W. M. Pradnya, "Sistem Pendukung Keputusan Jurusan Di Man Ii Yogyakarta

Menggunakan Algoritma Topsis," in *Semnasteknomedia Online*, 2017, pp. 55–60.
[6] D. A. Anju, F. Agustian, and K. I. Walid, "Sistem Pendukung Keputusan Pemilihan Jurusan di SMA dengan Analytic Hierarchy Process (AHP)," *J. Multinetics*, vol. 4, no. 1, pp. 27–33, 2018, doi: 10.32722/multinetics.v4i1.1075.
[7] Basri, M. Siddiq, R. Tamin, and S. Azis, "Data Mining Technique as Majors Support System Management with Classification Approach," *J. Phys. Conf. Ser.*, vol. 1244, no. 3, pp. 1–9, 2019, doi: 10.1088/1742-6596/1244/1/012004.
[8] E. Yudi Hidayat *et al.*, "Implementation of Weighted Naive Bayes Algorithm for Major Determination in Indonesian High School," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 580–584, doi: 10.1109/ISEMANTIC.2018.8549761.
[9] D. Gustian, A. F. Rahmawati, Titin, R. R. Putra, and P. Anisa, "Comparison of Classification Data Mining in Process Majors Students," in *2018 International Conference on Computing, Engineering, and Design (ICCED)*, 2018, pp. 125–130, doi: 10.1109/ICCED.2018.00033.
[10] C. R. Dedy Satrio Winarso, "Implementasi Algoritma k-Nearest Neighbor untuk Penjurusan Siswa SMA," *Cahayatech*, vol. 6, no. ISSN : 2302 – 2426 ISSN Online : 2580-2399, p. 50, 2017.
[11] Yuda Irawan, "Implementation Of Data Mining For Determining Majors Using K-Means Algorithm In Students Of SMA Negeri 1 Pangkalan Kerinci," *J. Appl. Eng. Technol. Sci.*, vol. 1, no. 1, pp. 17–29, 2019, doi: 10.37385/jaets.v1i1.18.
[12] M. G. Zataliny, "Sistem Pendukung Keputusan Pemilihan Jurusan pada Siswa SMA Negeri 1 Praya dengan Metode K-NN (K-NEAREST NEIGHBOR)," *Jurnal Mahasiswa Teknik Informatika*., vol. 1, no. 1, pp. 617–624, 2017.
[13] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak)," *J. Coding, Sist. Komput. Untan*, vol. 04, no. 3, pp. 75–84, 2016.
[14] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018.
[15] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi dan Sistem Komputer.*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: https://doi.org/10.14710/jtsiskom.8.2.2020.89-93.
[16] M. J. Zaki and J. Meira Wagner, "Probabilistic Classification," in *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed., Cambridge: Cambridge University Press, 2020, pp. 469–482.
[17] X.-S. Yang, *Introduction to Algorithms for Data Mining and Machine Learning*. London: Candice Janco, 2019.[18] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *J. ELTIKOM*, vol. 2, no. 2, pp. 67–77, 2018, doi: 10.31961/eltikom.v2i2.86.
[18] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *J. ELTIKOM*, vol. 2, no. 2, pp. 67–77, 2018, doi: 10.31961/eltikom.v2i2.86