ISSN: 1907-7912 EISSN: 2622-8092

Analisa Komparasi Algoritma Data Mining untuk Klasifikasi Heregistrasi Calon Mahasiswa STMIK Widya Pratama

Dadang Aribowo, Aris Ekyanto Heru Setiadi
STMIK Widya Pratama Pekalongan
E-mail: dadang.stmik.wp@gmail.com, aris_eky@stmik-wp.ac.id

RINGKASAN

Mahasiswa merupakan aset yang paling berharga dalam sebuah perguruan tinggi swasta (PTS). Karena sebagian besar pendapatan serta biaya operasional PTS didapatkan dari mahasiswa. Banyaknya mahasiswa yang melakukan heregistrasi jelas akan menjadi angin segar bagi lembaga. Dalam 5 tahun terakhir tercatat sekitar 20% mahasiswa STMIK Widya Pratama tidak melakukan heregistrasi. Pengetahuan dini terhadap calon mahasiswa yang mungkin tidak akan melakukan heregistrasi dapat menjadi acuan lembaga untuk melakukan tindakan guna mempertahankan mahasiswa. Pencatatan data mahasiswa yang tersusun rapi dapat digunakan pihak manajemen untuk melakukan analisa terhadap karakteristik serta penyebab mahasiswa tidak melakukan heregistrasi. Data mining dapat mengolah data lampau menjadi sebuah informasi atau pengetahuan baru. Dalam data mining terdapat satu fungsi mayor yaitu klasifikasi yang mengolah data training untuk menghitung data baru / data testing. Metode atau algoritma yang dapat digunakan dalam proses klasifikasi sangat banyak dengan berbagai macam karakteristik masing-masing. Beberapa algoritma klasifikasi terbaik antara lain naive bayes, knn, serta C4.5. Hasil penelitian menunjukkan bahwa ketiga algoritma yaitu knn, naive bayes serta decission tree C45 dapat digunakan untuk melakukan klasifikasi heregistrasi calon mahasiswa. Tingkat akurasi algoritma decission tree C45 merupakan yang terbaik yaitu 80,72% diikuti algoritma knn dengan tingkat akurasi 80,46%. Sedangkan tingkat akurasi naive bayes merupakan yang terendah dengan 74,49%.

Kata Kunci : Klasifikasi heregistrasi mahasiswa, KNN, Naive Bayes, Decission Tree C4.5

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Mahasiswa merupakan aset dalam sebuah perguruan tinggi. Kemajuan perguruan tinggi dapat dilihat dari jumlah mahasiswa yang melakukan pendaftaran pada setiap pembukaan tahun ajaran baru. Perguruan tinggi favorit tentunya akan lebih diminati calon mahasiswa dibandingkan dengan perguruan tinggi yang lain. Artinya banyaknya calon mahasiswa baru menjadi salah satu indikasi kemajuan perguruan tinggi.

Penerimaan mahasiswa baru di STMIK Widya Pratama dilakukan setiap tahun untuk keempat program studi yang ditawarkan. Tahapan dalam penerimaan Mahasiswa baru adalah sebagai berikut: isi formulir pendaftaran, tes ujian masuk, pengumuman hasil tes dan heregistrasi. Dalam 5 tahun trakhir selalu ada selisih yang cukup besar dari jumlah pendaftar dengan jumlah mahasiswa yang melakukan heregistrasi. Tabel 1 berikut menunjukkan data jumlah pendaftar pada 5 tahun terakhir.

Tabel 1. Data Penerimaan Mahasiswa Baru 5 tahun terakhir

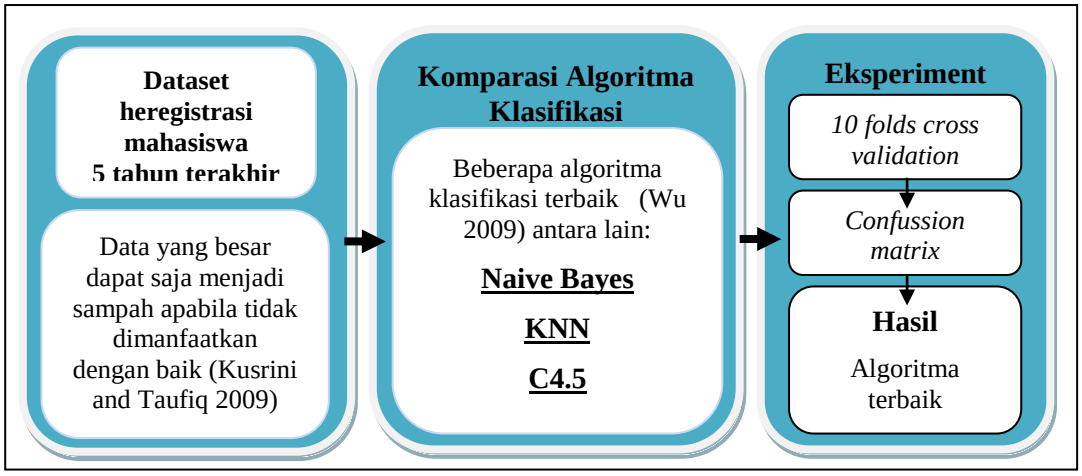
Tahun	Pendaftaran	Registrasi	Selisih	Prosentase	
				Registrasi	Tidak Registrasi
2013	705	513	192	73%	27%
2014	608	474	134	78%	22%
2015	667	500	167	75%	25%
2016	654	427	227	77%	23%
2017	506	428	78	84,6%	15,4%

Dari data tabel 1 terlihat adanya prosentase yang cukup besar yang terdapat pada calon mahasiswa yang tidak melakukan heregistrasi. Penurunan jumlah mahasiswa jelas akan berpengaruh dalam pendapatan lembaga, mengingat STMIK Widya Pratama merupakan kampus swasta. Adanya selisih jumlah yang cukup banyak antara jumlah pendaftar dan jumlah mahasiswa yang melakukan heregistrasi, membuat pihak yayasan berpikir keras agar jumlah mahasiswa STMIK Widya Pratama dapat terus dipertahankan dan ditingkatkan. Jika kemungkinan pengunduran diri mahasiswa dapat diketahui sejak dini, maka pihak manajemen dapat melakukan tindakan-tindakan preventif untuk mempertahankan calon mahasiswa tersebut (Kusrini and Taufiq 2009).

Klasifikasi adalah salah satu peran utama data mining (Witten, Frank, and Hall 2011). Proses klasifikasi dapat dilakukan dengan berbagai macam cara (Larose 2005). Berbagai macam algoritma juga dapat diaplikasikan dalam proses klasifikasi ini (Han and Kamber 2006). Berbagai masalah dalam dunia nyata banyak terselesaikan dengan menggunakan teknik klasifikasi data mining (Ashari, Paryudi, and Tjoa 2013). Penelitian klasifikasi lain juga banyak dilakukan oleh peneliti dengan menggunakan berbagai macam dataset serta algoritma yang berbeda (Ragab et al. 2014) (Patel, Vala, and Pandya 2014) (Amancio et al. 2013). Beberapa algoritma klasifikasi terbaik menurut Xindong Wu antara lain Naive Bayes, C4.5 serta K-Nearest Neighbour (Wu et al. 2007).

Klasifikasi heregistrasi mahasiswa menjadi bahan penelitian yang cukup menarik. Pada tahun 2014 Alkaromi (Alkaromi 2014) melakukan klasifikasi heregistrasi mahasiswa dengan menggunakan *K-NN*. Data yang digunakan adalah data PMB dengan *record* sebanyak 2389 dan atribut sebanyak 44. Kesemua atribut tersebut adalah atribut dasar yang didapatkan dari panitia PMB tanpa dikurangi atau dipilih sebelumnya. Untuk mengurangi jumlah atribut dalam penelitian tersebut dilakukan seleksi fitur dengan menggunakan *information gain*. Hasil dari penelitian tersebut akurasi *K-NN* yang telah dilakukan seleksi fitur menggunakan *information gain* naik hingga 83,93%. Sebelumnya algoritma *K-NN* tanpa menggunakan seleksi fitur hanya memeperoleh tingkat akurasi 78,15%. Dalam penelitian tersebut digunakan alat pengukuran *confussion matrix* dan validasi yaitu *10folds cross validation*.

Penelitian ini membandingkan algoritma KNN, Naive Bayes serta Decission Tree C45 untuk klasifikasi heregistrasi calon mahasiswa di STMIK Widya Pratama. Gambar 1 merupakan kerangka pemikiran dalam penelitian ini.



Gambar 1 Kerangka pemikiran penelitian

2. METODE PENELITIAN

Metode penelitian yang akan digunakan dalam penelitian ini adalah eksperimental. Tahapan penelitian antara lain konsep perencanaan dengan pengumpulan data, sampai dengan validasi dan evaluasi algoritma yang akan dibahas dalam sub bab berikut:

2.1 Pengumpulan Data

Tahapan pengumpulan data dalam penelitian ini dilakukan dengan mengambil data dari panitia penerimaan mahasiswa baru selama 5 tahun terakhir. Data tersebut dikonfersi dan dijadikan menjadi sebuah dataset baru. Data mentah berisikan 44 atribut data dan hanya akan

digunakan sejumlah 18 atribut untuk proses klasifikasi berikutnya. Pengurangan atribut ini dilakukan karena beberapa atribut tidak terkait dan memiliki tingkat varian yang sangat tinggi. Atribut yang tidak digunakan seperti halnya nama, alamat, nomor ktp, nomor telepon serta beberapa atribut lain.

Atribut data yang akan digunakan dalam penelitian ini antara lain: kode konsentrasi, biaya kuliah, gelombang grade, shift kelas, sesi, gelombang daftar, status daftar, kode daftar, pendidikan ortu, tahun lulus, status sipil, jenis kelamin, status pekerjaan, progdi, kelas, jenjang, kota asal, serta satu atribut label yaitu status registrasiDesain Eksperimen dan Pengujian

2.2 Desain Eksperimen Algoritma

Tahapan selanjutnya setelah pengumpulan data adalah desain eksperimental algoritma dilanjutkan dengan pengujian algoritma. Dalam tahapan ini nantinya akan dibandingkan ketiga algoritma untuk klasifikasi heregistrasi mahasiswa.

2.2.1 Tahap Eksperimen

Tahap eksperimen dilakukan dengan menggunakan *tools software* Rapid Miner. Dalam tahapan ini nantinya akan dilakukan perhitungan terhadap dataset yang telah terkumpul sebelumnya. Proses ini memungkinkan perulangan dan akan memilih algoritma dengan tingkat akurasi terbaik. Dalam tahap eksperimen ini dilakukan pula validasi dan evaluasi terhadap ketiga algoritma klasifikasi.

2.2.2 Validasi

Dalam proses validasi penelitian ini akan digunakan *10 folds cross validation*. Proses ini banyak digunakan oleh peneliti karena sudah terbukti baik dan menghasilkan tingkat akurasi yang stabil. Secara tori *10 folds cross validation* sudah dijelaskan secara lebih terinci dalam bab sebelumnya (Witten, Frank, and Hall 2011).

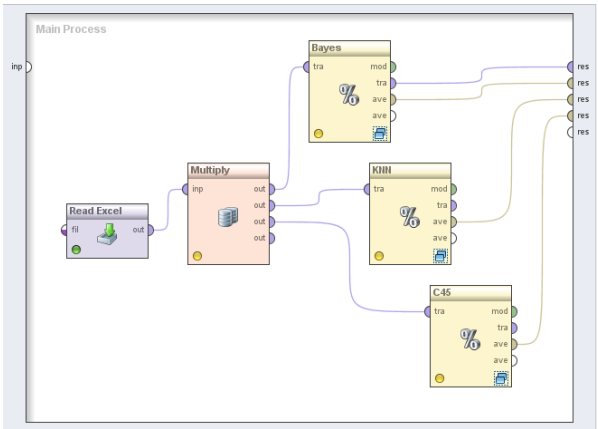
2.2.2.1 Pengukuran akurasi algoritma

Pengukuran dari suatu algoritma merupakan suatu pembuktian yang banyak dilakukan peneliti (Amancio et al. 2013). Dalam prosesnya banyak cara dapat digunakan untuk mengetahui performa suatu algoritma. Salah satu yang paling banyak digunakan adalah dengan menggunakan *confussion matrix* untuk menghitung akurasi algoritma. Perhitungan akurasi adalah presentase dari jumlah data *testing* dengan klasifikasi yang sesuai dengan aslinya dibagi keseluruhan data. Cara lain adalah dengan menghitung *Error rate. Error rate* adalah kebalikan dari tingkat akurasi, yaitu presentase kesalahan klasifikasi dibagi dengan keseluruhan dataset.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Eksperimen

Dalam tahapan eksperimen digunakan aplikasi rapid miner untuk mengolah data. Aplikasi ini dapat melakukan komparasi antara beberapa algoritma terpilih untuk melakukan perhitungan pada satu dataset yang sama. Dalam tahapan ini akan dilakukan validasi menggunakan *10 folds cross validation* serta pengukuran akurasi algoritma menggunakan *confussion matrix*. Gambar 2 merupakan desain eksperimen penelitian yang dilakukan.

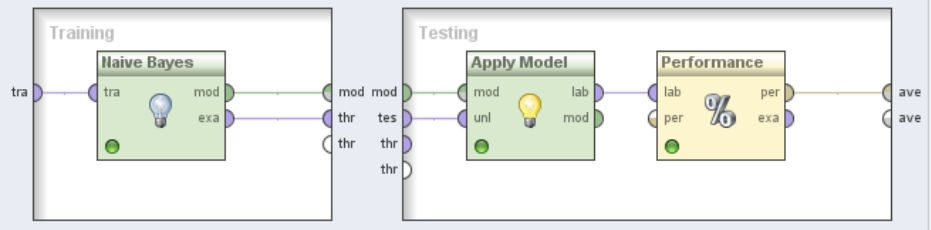


Gambar 2. Tampilan desain penelitian

Dalam gambar 2 tersebut terdapat beberapa proses yang juga telah ditandai menggunakan angka 1 sampai dengan 5. Berikut adalah penjelasan terkait proses yang ada:

1. **Read Excell,** bagian ini merupakan proses dimana dataset original diambil dari file excell. Proses ini memungkinkan pemilihan atribut dari dataset secara manual. Selain melakukan pemilihan secara manual, proses ini juga melakukan pemilihan tipe dari semua atribut termasuk atribut label atau atribut tujuan untuk klasifikasi.
2. **Multiply,** proses ini hanya memberikan lebih banyak output dari satu input yang sama. Multiply dalam penelitian ini digunakan untuk data heregistrasi mahasiswa agar dapat diakses untuk tiga jenis algoritma.

3. **Bayes,** dalam proses ini sebenarnya adalah proses validasi yaitu menggunakan *X-Validation*. Proses ini memungkinkan perulangan sebanyak 10 kali agar data yang ada dicampur secara merata. Didalam proses ini terdapat proses perhitungan algoritma naive bayes yang disertakan pula *confussion matrix* untuk melakukan perhitungan akurasi dari algoritma tersebut sebagaimana ada di gambar 2. Proses ini memungkinkan dataset dibagi menjadi dua bagian yaitu satu digunakan untuk data *training*, dan digunakan untuk data *testing.*
4. **KNN,** Dalam proses ini sebenarnya sama dengan proses pada bayes. Perbedaannya hanya ada pada bagian training yang diisi menggunakan algoritma KNN.
5. **C45,** Proses ini juga sama dengan proses algoritma lain seperti bayes dan KNN, proses training dalam hal ini diisi menggunakan algoritma Decission tree C45.



Gambar 2 Proses X-Validation

3.2 Validasi

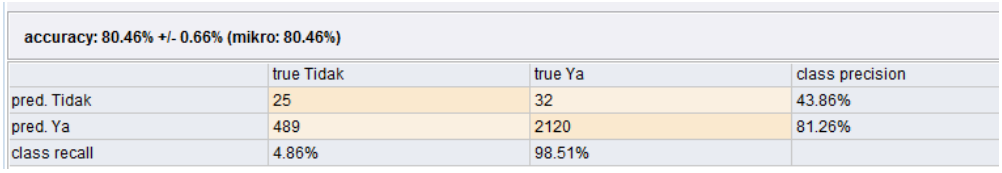
Proses validasi sebenarnya telah sedikit dibahas pada tahap eksperimen. Proses validasi yang digunakan dalam penelitian ini adalah *X-Validation* dengan perincian menggunakan *10 folds cross validation.* Proses ini memungkinkan dataset yang ada dibagi menjadi 10 bagian. Satu bagian diantaranya dijadikan *data testing* dengan 9 bagian lainnya menjadi *data training* untuk satu persatu algoritma. Proses ini berlanjut dengan menggunakan satu bagian yang lain untuk digunakan sebagai *data testing.* Proses ini akan berlanjut sampai dengan iterasi yang ke 10 agar kesemua bagian data mendapatkan proporsi menjadi *data testing.*

3.3 Pengukuran Akurasi Algoritma

Pengukuran akurasi dalam penelitian ini menggunakan *confussion matrix* atau matrix kebingungan untuk semua algoritma yang dibandingkan. Aplikasi rapid miner memiliki output yang sangat mudah dibaca dengan menggunakan matrix ini. Gambar 3 merupakan hasil *print screen* dari aplikasi rapid miner yang menunjukkan performa algoritma KNN untuk klasifikasi heregistrasi mahasiswa.

Dalam gambar 3 terlihat akurasi dari KNN sebesar 80,46%. Artinya dalam keseluruhan data *testing* ada 8046 yang sesuai dari 10000 percobaan *record*. Label dari data yang ada memiliki 2 varian yaitu Ya dan Tidak. Dalam matrix kebingungan ini terdapat tabel berwarna kuning dengan jumlah matrix sebesar 2 x 2. Matrix ini menunjukkan jumlah data *testing* yang sesuai dan tidak sesuai dengan label yang sebenarnya. Bagian atas merupakan bagian data atau label sebenarnya sedangkan bagian kiri merupakan bagian prediksi atau hasil perhitungan algoritma. Dalam gambar 3 true Tidak dengan pred. Tidak sebesar 25 dan true Ya dengan pred. Ya sebesar 2120. Artinya dataset hasil perhitungan algoritma yang memiliki hasil sama dengan data asli sebanyak 2120+25 yaitu 2145. Sedangkan keseluruhan jumlah *record* adalah 25+32+489+2120 yaitu 2666. Tingkat akurasi algoritma KNN dapat dihitung dengan 2145 dibagi 2666 dikalikan 100% yaitu 80,4576%.

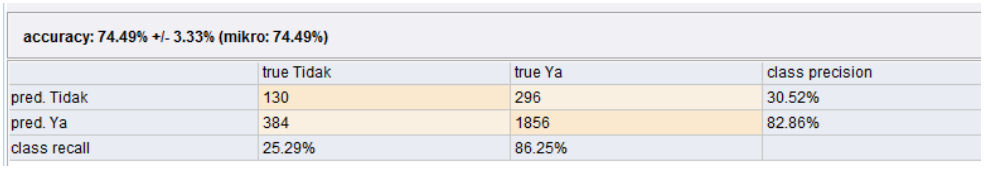
Untuk algoritma naive bayes dan *decission tree* C45 dilakukan proses yang sama dengan proses dari KNN diatas. Gambar 4 merupakan hasil akurasi dari algoritma naive bayes dengan nilai akurasi sebesar 74,49%. Sedangkan gambar 5 merupakan hasil akurasi dari *decission tree* C45 dengan tingkat akurasi tertinggi dari kesemuanya yaitu 80,72%.



accuracy: 80.46% +/- 0.66% (mikro: 80.46%)

	true Tidak	true Ya	class precision
pred. Tidak	25	32	43.86%
pred. Ya	489	2120	81.26%
class recall	4.86%	98.51%	

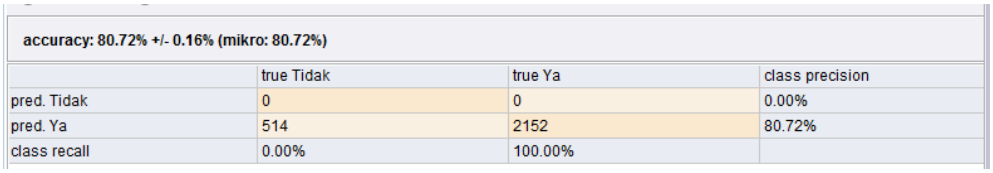
Gambar 3 Akurasi algoritma KNN



accuracy: 74.49% +/- 3.33% (mikro: 74.49%)

	true Tidak	true Ya	class precision
pred. Tidak	130	296	30.52%
pred. Ya	384	1856	82.86%
class recall	25.29%	86.25%	

Gambar 4 Akurasi algoritma naive bayes



accuracy: 80.72% +/- 0.16% (mikro: 80.72%)

	true Tidak	true Ya	class precision
pred. Tidak	0	0	0.00%
pred. Ya	514	2152	80.72%
class recall	0.00%	100.00%	

Gambar 5 Akurasi algoritma *decission tree* C45

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Hasil perhitungan menunjukkan bahwa ketiga algoritma yaitu KNN, naive bayes serta *decission tree* C45 dapat digunakan untuk melakukan klasifikasi heregistrasi calon mahasiswa. Hasil akurasi dari penelitian ini dapat dilihat pada tabel 2 berikut.

4.2 Saran

Penelitian ini melakukan komparasi antara algoritma KNN, naive bayes serta *decission tree* C45. Hasil dari penelitian ini adalah tingkat akurasi untuk setiap algoritma yang ada. Dataset yang digunakan adalah data gabungan dari tahun 2013 sampai dengan tahun 2017. Penelitian berikutnya dapat dilakukan perbandingan dengan dataset tahun setelahnya sebagai data *testing*.

Tabel 2 Hasil Penelitian

Algoritma	accuracy	precission	recall	AUC
KNN	80,46 %	81,26 %	98,51 %	0,578
Naive bayes	74,49 %	82,85 %	86,25 %	0,602
***Decission tree* C45**	80,72	80,72 %	100 %	0,500

5. DAFTAR PUSTAKA

Alkaromi, M Adib. 2014. "Information Gain Untuk Pemilihan Fitur Pada Klasifikasi Heregistrasi Calon Mahasiswa Dengan Menggunakan K-NN."

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. a. Rodrigues, and L. Da F. Costa. 2013. "A Systematic Comparison of Supervised Classifiers," October. http://arxiv.org/abs/1311.0202v1.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques Second Edition*. *Elsevier*. Elsevier.

Kusrini, and Luthfi Emha Taufiq. 2009. *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Patel, Kanu, Jay Vala, and Jaymit Pandya. 2014. "Comparison of Various Classification Algorithms on Iris Datasets Using WEKA" 1 (1): 1–7.

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14*. New York, New York, USA: ACM Press, 106–13. doi:10.1145/2643604.2643631.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.

Wu, Xindong. 2009. *The Top Ten Algorithms in Data Mining*. Edited by Vipin Kumar. New York: Taylor & Francis Group, LLC.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining*. *Knowledge and Information Systems*. Vol. 14. doi:10.1007/s10115-007-0114-2.