DATA MINING UNTUK KLASIFIKASI PENENTUAN PEMINATAN SISWA SMA NEGERI 2 TENGGARONG SEBERANG DENGAN MENGGUNAKAN ALGORITMA C4.5

Bambang Cahyono[1], Islamiyah[2]
Program Studi Teknik Informatika Universitas Mulawarman
cbambang87@gmail.com, islamiyah1601@yahoo.co.id

Abstract

Currently the majority of students choose a specialization majors following the choices made by the majority of his friends, without considering the factor of academic achievement of students. This resulted in a mismatch specialization interests and skills of the student, as a result many students who have difficulty in catch-up lessons. Application of C4.5 algorithm in the choice of subject specialization will assist in the classification of the variables that influence the selection of the field of specialization majors. C4.5 algorithm is an algorithm that is effective enough to help form a decision tree, the decision tree will then generate a new knowledge. Data collection techniques used interviews, observation, library research, and documentation. Research and evaluation results showed that the analysis of the determination of specialization using 150 data sets consisting of 70% of the training data and testing data is 30% resulting in a level of accuracy of 84,44% Confusion Matrix.

Keywords: majority of student, algoritma C4.5, confusion matrix, decision tree, classification

1. Pendahuluan

Kemajuan teknologi informasi telah menyebabkan banyak orang dapat memperoleh datadengan mudah bahkan cenderung berlebihan. Data tersebut semakin lama semakin banyak dan terakumulasi, akibatnya pemanfaatan data yang terakumulasi tersebut menjadi tidak optimal. Banyaknya data yang dimiliki oleh sebuah organisasi bisa menyebabkan kesulitan dalam pengklasifikasian data tersebut untuk kepentingan organisasi. Kegiatan pengklasifikasian yang dilakukan oleh manusia masih memiliki keterbatasan, terutama pada kemampuan manusia dalam menampung jumlah data yang ingin diklasifikasikan. Selain itu, bisa juga terjadi kesalahan dalam pengklasifikasian yang dilakukan. Salah satu cara mengatasi masalah ini adalah dengan menggunakan Data Mining (DM) dengan teknik klasifikasi. Data mining dapat membantu sebuah organisasi yang memiliki data melimpah untuk memberikan informasi yang dapat mendukung pengambilan keputusan.[7]

Dalam dunia pendidikan, data yang berlimpah dan berkesinambungan mengenai siswa yang dibina dan alumni terus dihasilkan. Data yang berlimpah membuka peluang diterapkannya data mining untuk pengelolaan pendidikan yang lebih baik dan data mining dalam pelaksanaan pembelajaran berbantuan komputer yang lebih efektif [2]. Sementara itu, Selain itu, data mining dapat digunakan untuk menyelesaikan siswa yang bermasalah dan membantu institusi menjadi lebih proaktif dalam mengidentifikasi dan merespon siswa tersebut. Luan menerapkan data mining sebagai cara untuk memprediksi ciri-ciri siswa yang akan dikeluarkan oleh sekolah dan kemudian kembali ke sekolah tersebut pada tahun berikutnya.[11]

SMA Negeri 2 Tenggerong Seberang adalah salah satu SMA yang ada di Kecamatan Tenggarong Seberang. Kurikulum yang diterapkan adalah kurikulum 2013 dimana kurikulum ini berbeda dengan kurikulum sebelumnya. Salah satu perbedaan dengan kurikulum sebelumnya adalah dalam hal penentuan peminatan siswa. Siswa kelas X sudah menentukan peminatannya terlebih lagi proses dalam penentuan peminatan ini masih bersifat manual sehingga memerlukan waktu yang lama dan hasilnya pun belum tentu akurat. Faktor utama dalam penentuan peminatan siswa ini ditentukan oleh beberapa faktor yaitu nilai ujian nasional yang terdiri dari matematika, bahasa inggris, bahasa Indonesia, IPA sedangkan nilai tes yaitu IPA, IPS, Bahasa Indonesia, Bahasa inggris dan wawancara.

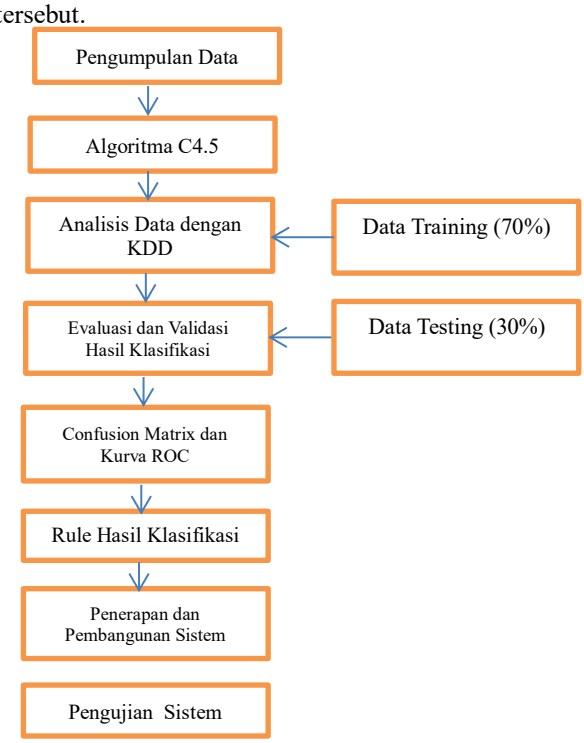
Peminatan merupakan suatu keputusan yang dilakukan oleh peserta didik untuk memilih kelompok matapelajaran sesuai dengan minat, bakat, dan kemampuan selama mengikuti pelajaran di SMA. Pemilihan peminatan dilakukan atas dasar kebutuhan untuk melanjutkan keperguruan tinggi. Ketepatan dalam menentukan peminatan dapat menentukan keberhasilan belajar siswa. Sebaliknya, kesempatan yang sangat baik bagi siswa akan hilang karena kekurangtepatan dalam penentuan peminatan.

Seiring dengan perkembangan teknologi hal tersebut dapat diatasi dengan teknik pengelompokan data dengan data mining. Sementara itu, data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan

mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [10]. Teknik pengelompokan atau pengklasifikasian yang dimaksud adalah teknik klasifikasi dengan menggunkan algoritma C4.5.

2. Metode Penelitian

Penelitian ini termasuk penelitian eksperimen dengan menggunakan data siswa SMA Negeri Tenggarong Seberang dengan menggunakan sebanyak 150 data siswa yang terdiri dari data training 70% dan data testing 30%, sehingga diperoleh data testing 105 data sedangkan data testing sebanyak 45 data. Data training digunakan untuk memperoleh hasil klasifikasi penentuan peminatan dalam bentuk *decision tree*, sedangkan data *testing* digunakan untuk mengukur tingkat akurasi dari hasil klasifikasi tersebut.



Gambar 1. Alur Penelitian

3. Hasil dan Pembahasan

3.1 Data Uji

Data uji yang digunakan dalam penelitian ini adalah data siswa SMA Negeri 2 Tenggarong Seberang Kelas X Tahun Akademik 2015/2016. Jumlah data siswa yang digunakan sebanyak 150 siswa yang terdiri dari siswa IPS sebanyak 30 orang dan siswa IPA senyak 130 orang. Data ini memiliki data atribut peminatan sebanyak 10 atribut diantaranya: Nilai Ujian Nasional Matematika, IPA, Bahasa Indonesia, Bahasa Inggris, Nilai Ujian Tes Akademik terdiri dari nilai Bahasa Indonesia, Bahasa Inggris, IPA, dan IPS. Selain itu ada juga wawancara dalam hal ini ketika proses wawancara ini calon siswa ditanyakan tentang minatnya. Atribut terakhir adalah atribut hasil yang merupakan target yang diingin dicapat yaitu peminatan IPA dan IPS.

Untuk menghasilkan data yang akurat maka nilai dari setiap siswa baik nilai Ujian Nasional maupun nilai tes akademik akan diklasifikasikan dalam rentang nilai yang dapat dilihat pada tabel di bawah ini

Tabel 1. Klasifikasi Nilai

Nilai	Klasifikasi
80-100	A
70-79	B
55-69	C
54-40	D
<40	E

3.2 Analisis Data

Analisis data yang digunakan untuk proses klasifikasi dapat menggunakan KDD (*Knowledge Discovery in databases*) yang terdiri dari Sembilan langkah yang dimulai dari tahap pemahaman data yang akan digunakan hingga tahap terciptanya sebuah pengetahuan tentang klasifikasi penentuan peminatan. KDD sendiri diartikan sebagai proses terorganisir untuk mengidentifikasi pola dalam data yang besar dan kompleks dimana pola data tersebut ditemukan bersifat sah, baru dan dapat bermanfaat serta dapat dimengerti. Sembilan langkah dalam KDD yang digunakan dalam analisis data untuk proses klasifikasi.[12]

1. *Developing an understanding of the application domain* yang merupakan tahap untuk memahami apa yang akan dilakukan dalam penelitian
2. *Selecting and creating a data set on which discovery will be performed*, merupakan tahap untuk pemilihan set data dan mempersiapkannya untuk digunakan dalam penelitian.
3. *Preprocessing and cleansing*, merupakan tahap untuk meningkatkan kehandalan data dengan cara membersihkan data yang tidak lengkap (missing value) dan data yang tidak benar *(noise)*
4. *Data transformation*, merupakan tahap untuk menyusun dan mengembangkan set data menjadi lebih baik sehingga tahap ini membutuhkan prroses kreatif dan sangat bergantung ada jenis atau pola informasi yang akan dicari dalam basis data, seperti mengkategorikan data ke dalam beberapa kategori dan membagi data menjadi dua bagian yaitu data training dan testing.
5. *Choosing the appropriate Data Mining task*, tahap ini memilih teknik data mining yang digunakan yaitu klasifikasi.
6. *Choosing the Data Mining Algorithm*, merupakan tahap membuat klasifikasi beasiswa dengan menggunakanalgoritma yang telah dipilih dari proses

sebelumnya.

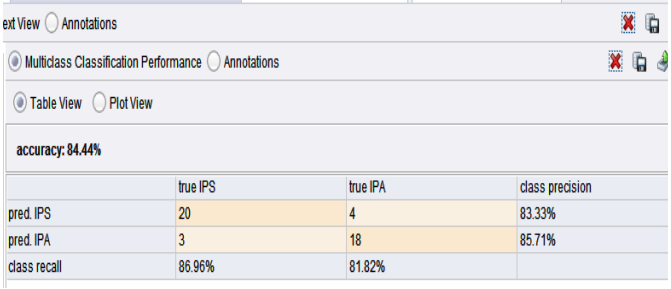
7. *Employing the data Data mining Algorithm*, merupakan tahap membuat klasifikasi beasiswa dengan menggunakan algoritma yang telah dipilih dari proses sebelumnya.
8. *Evaluation*, tahap evaluasi dilakukan dengan menggunakan data testing untuk mengukur tingkat akurasi pola data yang diperoleh data yang diperolehhasil klasifikasi dengan data training dengan menggunakan *confusion matrix* dan kurva ROC.
9. *Using the discovered knowledge*, merupakan tahap menggunakan *knowledge* yang diperoleh dari hasil klasifikasi dengan *Decision Tree* dan menerapkannya dalam klasifikasi penentuan peminatan

3.3 Evaluasi Hasil

Evaluasi dari hasil klasifikasi peminatan siswa dengan menggunakan C4.5 dapat menggunakan *confusion matrix* dan kurva ROC/AUC (*Area under Curve*).

1. *Confusion Matrix*

Confusion matrix merupakan sebuah metode untuk evaluasi yang mengunakan tabel matrix. Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy, precision,* dan *recall.*



	true IPS	true IPA	class precision
pred. IPS	20	4	83.33%
pred. IPA	3	18	85.71%
class recall	86.96%	81.82%	

Gambar 2. Hasil *Confusion Matrix*

Tingkat akurasi dalam *confusion matrix* dapat dihitung menggunakan rumus sebagai berikut [6]

$$Sencitivity = \frac{t_{pos}}{t_{pos}+t_{neg}} x\ 100\% \quad \ldots\ldots(6)$$

$$Sencitivity = \frac{t\_neg}{neg} x\ 100\% \quad \ldots\ldots(7)$$

$$Precision = \frac{t\_pos}{t\_pos + f\_pos} x\ 100\% \quad \ldots..(8)$$

$$Accuracy \ \frac{t_{pos}+t\_neg}{t_{pos}+f_{neg}+f\_pos+t\_neg} x\ 100\% \ \ldots..(9)$$

Keterangan:

t_pos	: jumlah *true positif*
t_neg	: jumlah *true negative*
p	: jumlah *record positif*
n	: jumlah *tupel negative*
f_pos	: jumlah *false positif*
f_negatif	: jumlah *false negative*

Hasil evaluasi *confusion matrix* dalam klasifikasi menunjukkan tingkat akurasi hasil klasifikasi peminatan siswa sebesar 84,44%. Untuk nilai *precision* sebesar 85,71%, sedangkan nilai *recal* sebesar 81,82%.

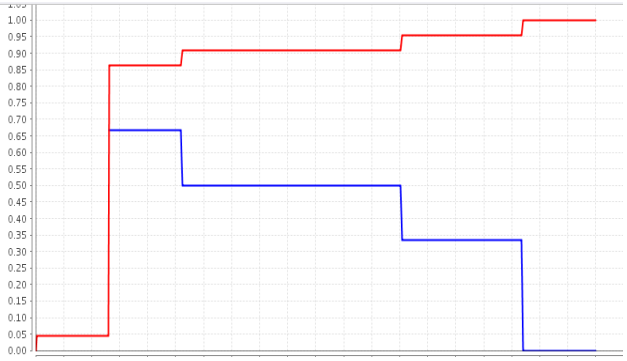
2. Kurva ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC merupakan grafik dua dimensi dengan *false positive* sebagai garis horizontal dan *true positive* sebagai garis *vertical* [15]

ROC memiliki tingkat nilai diagnose yaitu: [4]

a. Akurasi bernilai 0.90 – 1.00 = *excellent classification*
b. Akurasi bernilai 0.80 – 0.90 = *good classification*
c. Akurasi bernilai 0.70 – 0.80 = *fair classification*
d. Akurasi bernilai 0.60 – 0.70 = *poor classification*
e. Akurasi brnilai 0.50 – 0.60 = *failure*

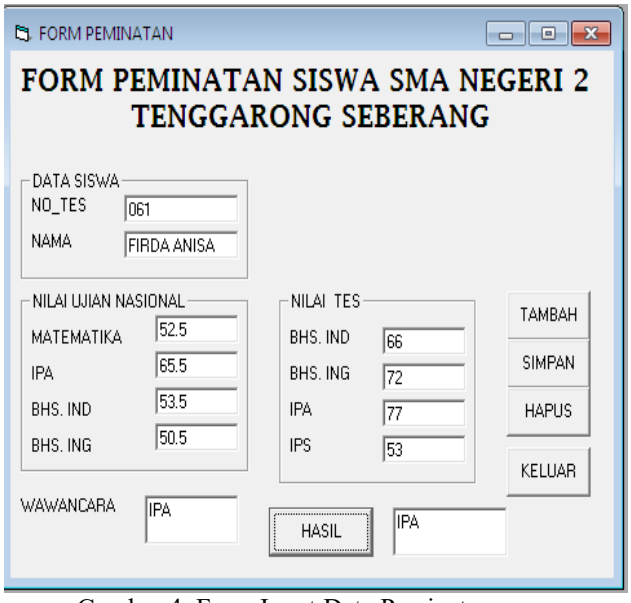
Hasil ROC dari penelitian ini dapat dilihat pada gambar dibawah ini



Gambar 3. Hasil Evaluasi Kurva ROC

3.3 Pembangunan Sistem

Pembangunan sistem klasifikasi peminatan siswa menggunakan *Microsoft Visual Basic 6.0* dengan menerapkan rule hasil klasifikasi dengan menggunakan algoritma C4.5. Hasil dari pembangunan sistem ini dapat dilihat pada gambar berikut ini:



Gambar 4. Form Input Data Peminatan

Dari gambar di atas dapat menjelaskan hasil dari proses data mining, yaitu memprediksi pohon keputusan yang terbentuk dari proses data mining yang menentukan bahwa inputan yang dimasukkan berdasarkan 10 atribut apakah target tersebut adalah IPA atau IPS. Untuk membandingkan data hasil uji yang sebenarnya dengan hasil uji menggunakan data mining maka diperoleh hasil seperti di bawah ini:

NO	No_ Test	NAMA	UN MAT	UN IPA	UN BHS. IND	UN BHS. ING	TES. BHS. IND	TES. BHS. ING	TES. IPA	TES. IPS	WAWAN CARA	HASIL	HASIL MINING	KET
1	061	Firda Anisa	D	C	D	D	C	B	B	D	IPA	IPA	IPA	Sesuai
2	123	Nurhalifah Tusadiyah	C	C	C	C	C	B	E	D	IPS	IPA	IPS	Tidak
3	143	Suciana Maharani	E	D	D	E	D	A	D	E	IPS	IPA	IPA	Sesuai
4	023	Novia Syahtri	E	D	D	E	E	C	D	D	IPA	IPA	IPA	Sesuai
5	037	Indar Paramitha	C	B	B	C	C	B	C	E	IPA	IPA	IPA	Sesuai
6	094	Vita Wulandari	C	C	C	C	E	C	E	E	IPA	IPA	IPA	Sesuai
7	012	Mareta Tri Utami	E	D	D	D	C	A	B	E	IPA	IPA	IPA	Sesuai
8	015	Mariyana	D	D	D	D	C	A	B	E	IPA	IPA	IPA	Sesuai
9	018	Muhammad Miftachul Rozikl	D	D	D	D	D	C	C	E	IPA	IPA	IPS	Tidak
10	023	Rendra Nur Hidayat	D	D	D	D	C	B	B	D	IPS	IPA	IPS	Tidak
11	055	Rilo Bayu Permadi	E	D	E	D	D	A	D	E	IPS	IPA	IPA	Sesuai
12	067	Oky Nurih Trinato	D	C	D	D	C	B	B	D	IPS	IPA	IPA	Sesuai
13	036	Riandi Erlangga	D	C	C	C	D	A	D	E	IPS	IPA	IPA	Sesuai
14	051	Sandra Kirana Mirella Manja	E	E	D	D	C	B	B	D	IPS	IPA	IPA	Sesuai
15	032	Andrea Jelita Putri	E	D	E	D	D	A	D	E	IPS	IPA	IPA	Sesuai
16	039	Ainul Latifah	D	C	D	C	E	C	B	E	IPA	IPA	IPA	Sesuai
17	016	Oktavia Devi Novrenda	C	C	C	C	E	A	D	E	IPS	IPA	IPA	Sesuai
18	087	Ersa Merliani	E	E	E	D	E	B	E	E	IPA	IPS	IPS	Sesuai
19	093	Anita Miftahul J	D	D	D	D	D	C	C	E	IPA	IPA	IPA	Sesuai
20	077	Dony Firmansyah	E	D	D	D	C	B	B	D	IPS	IPA	IPA	Sesuai
21	001	Agita Erzha W	D	C	D	C	D	A	D	E	IPS	IPA	IPA	Sesuai
22	009	Riki Al Fajri	D	D	D	D	E	C	B	E	IPA	IPA	IPA	Sesuai
23	007	Febby Paramitha Mis Palah	D	D	D	D	D	B	E	E	IPA	IPS	IPS	Sesuai
24	005	Irawan Rohmayadi	D	D	D	D	C	B	B	D	IPS	IPA	IPA	Sesuai
25	022	Fara Zulfiani	E	D	E	D	C	C	C	C	IPA	IPA	IPA	Sesuai
26	066	Meiroshe Sianturi	D	D	D	D	E	E	D	D	IPS	IPA	IPA	Sesuai
27	063	Andi Yusuf Kurniawan	D	C	D	C	D	D	D	D	IPS	IPA	IPA	Sesuai
28	052	Kevin Revail B	D	D	D	D	E	B	E	E	IPA	IPS	IPS	Sesuai
29	081	Novi Kartika Indriani	E	D	E	D	E	A	E	B	IPA	IPS	IPS	Sesuai
30	125	Erlina safira	D	D	D	D	D	B	E	B	IPA	IPS	IPS	Sesuai

Gambar 5. Perbandingan Data uji

4 Kesimpulan

Berdasarkan hasil penelitian peminatan jurusan pada SMA Negeri 2 Tenggarong Seberang dengan menggunakan algoritma C4.5 dapat diambil beberapa kesimpulan antara lain:

1. Model *decision tree* dapat digunakan dalam membuat klasifikasi sebagai dasar dalam pembangunan sistem penentuan peminatan siswa IPA dan IPS.
2. Hasil evaluasi dan validasi dengan menggunakan *confusion matrix* menunjukkan tingkat akurasi algoritma C4.5 sebesar 84,44% dengan menggunakan data set sebanyak 150 data yang dibagi kedalam data training 70% dan data testing 30%.
3. Atribut yang digunakan untuk menentukan peminatan siswa terdiri dari 10 atribut antara lain: UN IPA, UN Bahasa Indonesia, UN Bahasa Inggris, UN Matematika, Tes IPA, Tes IPS, Tes Bahasa Indonesia, Tes Bahasa Inggris. Sebanyak 10 atribut tersebut yang memiliki Gain tertinggi adalah Tes IPA sehingga dijadikan sebagai *root* pada *decision tree*.
4. Penerapan *rules* dari algoritma C4.5 selanjutnya diterapkan menggunakan *Visual Basic 6.0* yang digunakan dalam klasifikasi peminatan siswa IPA dan IPS.

DAFTAR PUSTAKA

[1] Anik Andriani, 2012, "*Penerapan Algoritma C4.5 pada Program Klasifikasi Mahasiswa Dropout*".

[2] Ayub, Mewati, 2007, Proses Data Mining dalam Sistem Pembelajaran Berbantuan Komputer, *Jurnal Sistem Informasi*, Vol. 2 No. 1, Maret 2007, hal. 21-30.

[3] Eka Budi Rahayu, 2014. Algoritma C4.5 untuk Penjurusan Siswa SMA Negeri 3 Pati.

[4] Gorunnescu, F. 2011. Data mining Concept and Techniques. Berlin: Springer. ISBN 978-3-642-19720-8.

[5] Han, J., Kamber, M., Tung, A. K. H., 2001, *Spatial Clustering Methods in Data Mining: A Survey*, School of Computing Science Simon Fraser University Burnaby, Canada.

[6] Han, J., & Kamber, M. 2006. *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman. ISBN 13:978-1-55860-901-3.

[7] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M., 2012, *Analytics: The Widening Divide*. MIT Sloan Management Review, 53(2), 1-22.

[8] Kusrini, Luthfi, and Emha, 2009, *Algoritma Data Mining*. Andi Offset.

[9] Larose T. Daniel. 2006. *Data Mining Methods and Models*. John Wiley & Sons, Inc Publication

[10] Liliana Swastika, "*Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa*", Gema Aktualita, Juni 2013.

[11] Luan, J., 2002, *Data Mining and Knowledge Management in Higher Education Applications, Paper presented at the Annual Forum for the Association for Institutional Research,* Toronto, Ontario, Canada. http://eric.ed.gov/ERICWebPortal/detail?accno=ED474143.

[12] Maimon, Oded., & Rokach, Lior. 2010. *Data Mining and Knowledge Discovery Handbook*. 2nd Edition. New York: Springer. ISBN 978-0-387-09822-7.

[13] Obbie Kristianto, Penerapan Algoritma Klasifikasi Data Mining ID3 untuk Menentukan Penjurusan Siswa SMAN 6 Semarang", September 2014.

[14] Ratih Ariadni and Isye Arieshanti, "Imlementasi Metode Pohon Keputusan untuk Klasifikasi Data dengan Nilai Fitur yang tidak Pasti.

[15] Vercellis, Carlo. 2009. *Business Intelligence: Data Mining and Optimization for Decision Making*. United Kingdom: John Willey & Son.

[16] Yusuf Sulistyo Nugroho, Penerapan Algoritma C4.5 untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta, SNAST 2014.