

Penerapan Algoritma C4.5 Untuk Klasifikasi Pola Penjurusan di Sekolah Menengah Kejuruan (SMK) Kota Tasikmalaya

Application Of C4.5 Algorithm For Classification Of Majors In Vocational High School (SMK) In City Of Tasikmalaya

Egi Badar Sambani¹, Fitri Nuraeni²

STMIK Tasikmalaya, Jalan RE Martadinata 272 A Tasikmalaya
e-mail: ¹egibadar@gmail.com, ²nufi3@stmik-tasikmalaya.ac.id

Abstrak

Penentuan jurusan di SMK merupakan hal yang sangat penting, karena siswa akan dibekali pembelajaran sesuai dengan jurusan yang telah dipilih sehingga memiliki keterampilan sebagai bekal memasuki dunia kerja. Namun dalam prosesnya masih sering mengalami masalah, karena pihak sekolah masih sulit dalam menentukan kebijakan untuk penentuan jurusan calon siswanya, diperlukan banyak factor penilaian dan semua data diolah dalam waktu yang terbatas. Maka dari itu pihak sekolah perlu mengetahui pola penjurusan siswa berdasarkan nilai rapor, tes kesehatan, dan olahraga. Nantinya pola tersebut memudahkan pihak sekolah dalam menentukan jurusan mana yang tepat untuk calon siswa tersebut. Pola penjurusan diperoleh dengan proses data mining yang mengikuti langkah-langkah CRIPS-DM. Algoritma C4.5 merupakan salah satu bagian dari Teknik klasifikasi dengan kemampuan model algoritma C4.5 yang dapat melakukan proses learning dan klasifikasi yang sederhana, cepat dan secara umum mempunyai tingkat akurasi yang tinggi. Dengan menggunakan dataset 1104 record data siswa baru, hasil dari pengklasifikasian data pemilihan jurusan SMK untuk jurusan AP, TKJ, RPL, TKR, dan TSM menghasilkan akurasi sebesar 92,30%. Dengan tingkat akurasi yang baik, pola klasifikasi penentuan jurusan ini dapat digunakan pihak sekolah dalam menentukan kebijakan dalam penentuan jurusan calon siswa pada proses penerimaan peserta didik baru.

Kata Kunci— C4.5, Jurusan, Klasifikasi, SMK

Abstract

The determination of majors in SMK is very important, because students will be equipped with learning in accordance with the majors that have been selected so as to have skills as stock enter the world of work. But in the process is still often experiencing problems, because the school is still difficult in determining the policy for the determination of the prospective student majors, it takes a lot of assessment factors and all data is processed in a limited time. Therefore, the school needs to know the pattern of student majors based on the value of report cards, health tests, and sports. Later the pattern allows the school in determining which majors are right for the prospective student. Pattern majors are obtained by data mining processes that follow CRIPS-DM steps. Algorithm C4.5 is one part of the classification technique with the ability of C4.5 algorithm model that can make the learning process and classification is simple, fast and generally have a high degree of accuracy. By using dataset 1104 record of new student data, result of classification of data of SMK majors for majors AP, TKJ, RPL, TKR, and TSM produce accuracy equal to 92,30%. With a good level of accuracy, the classification pattern of the determination of this department can be used by the school in determining the policy in the selection of prospective majors in the process of admission of new learners.

Keywords— C4.5, Majors, Classification, SMK

1. PENDAHULUAN

Sekolah menengah kejuruan (SMK) adalah salah satu bentuk satuan pendidikan formal yang menyelenggarakan pendidikan kejuruan pada jenjang pendidikan menengah sebagai lanjutan dari SMP/MTs atau bentuk lain yang sederajat atau lanjutan dari hasil belajar yang diakui sama/setara SMP/MTs. Mengacu pada isi Undang-Undang Sistem Pendidikan Nasional No. 20 Tahun 2003 pasal 3 mengenai tujuan pendidikan nasional dan penjelasan pasal 15 yang menyebutkan bahwa pendidikan kejuruan merupakan pendidikan menengah yang mempersiapkan peserta didik terutama untuk bekerja di bidang tertentu.

Pendidikan kejuruan adalah bagian dari sistem pendidikan yang mempersiapkan seseorang agar lebih mampu bekerja pada suatu kelompok pekerjaan atau satu bidang pekerjaan dari pada bidang-bidang pekerjaan lainnya. Dengan pengertian bahwa setiap bidang studi adalah pendidikan kejuruan sepanjang bidang studi tersebut dipelajari lebih mendalam dan kedalaman tersebut dimaksudkan sebagai bekal memasuki dunia kerja. Maka pada setiap tahun ajaran baru, siswa yang ingin mendaftar di SMK mengajukan pilihan jurusan yang dipilih selama belajar di SMK tersebut. Tes yang biasanya dilakukan pada saat penyeleksian siswa baru di SMK diantaranya psikotest, wawancara, test kesehatan dan test olahraga yang meliputi sit up, push up dan lari. Selain itu, nilai rapor sewaktu SMP juga berperan penting dalam proses penyeleksian calon siswa SMK, khusus nya bagi yang menggunakan jalur prestasi.

Penentuan jurusan di SMK biasanya berdasarkan 2 program keahlian yang telah dipilih ketika mendaftar. Contoh-contoh pilihan jurusan yang ada di SMK seperti Teknik Kendaraan Ringan, Teknik Sepeda Motor, Rekayasa Perangkat Lunak, Administrasi Perkantoran dan Teknik Komputer dan Jaringan, dan lain sebagainya. Kemudian dari hasil tes dan nilai rapor diolah lagi untuk mendapatkan nilai yang memenuhi syarat untuk pilihan jurusan pertama. Jika standar nilai pada jurusan pertama tidak terpenuhi, maka dicocokkan untuk pilihan jurusan yang kedua. Jika sesuai dengan standar nilai jurusan yang ke 2 maka di masukan ke pilihan jurusan yang kedua. Dan jika tidak sesuai dengan kedua pilihan jurusan maka siswa di nyatakan tidak diterima oleh SMK tersebut.

Perlu diketahui bahwa proses penentuan jurusan di SMK itu merupakan hal yang sangat penting, karena nantinya siswa dibekali pembelajaran sesuai dengan jurusan yang telah dipilih secara lebih mendalam dan kedalaman tersebut dimaksudkan sebagai bekal memasuki dunia kerja. Maka dari itu pihak sekolah perlu mengetahui pola minat siswa yang sesuai jurusan yang tersedia untuk menghindari kesalahan dalam pemilihan jurusan. Pola tersebut juga nantinya memudahkan pihak sekolah untuk menentukan jurusan mana yang tepat untuk calon siswa tersebut.

Seiring dengan perkembangan teknologi, penggalan pola dapat dilakukan dengan teknik klasifikasi data yang didapat dari proses data mining. Data mining biasanya dilakukan karena adanya sejumlah besar data yang dapat digunakan untuk menghasilkan pola dan *knowledge* yang berguna untuk proses bisnis suatu perusahaan. Salah satu teknik yang dibuat dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak ada dalam basis data yang tersimpan[1].

Begitupun di SMK, pola yang dibutuhkan untuk proses penentuan jurusan ini dapat dihasilkan dari proses klasifikasi. Klasifikasi merupakan salah satu metode pada data mining yang termasuk kedalam model deskriptif. Teknik klasifikasi, sebagai metode deskriptif, adalah sebuah contoh dari teknik *supervised machine*, dengan asumsi adanya sekelompok contoh diberi label untuk setiap kategori objek.

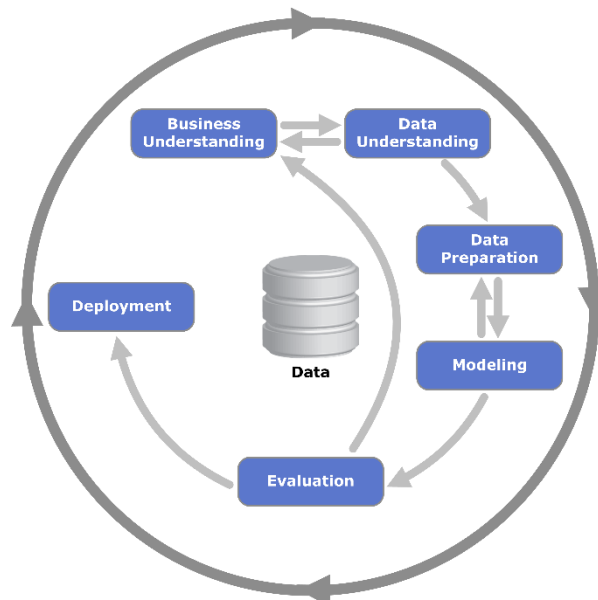
Proses klasifikasi jurusan yang telah ada sebelumnya menggunakan algoritma ID3, dengan menggunakan 371 dataset dan 20 data uji yang diinputkan terdapat 4 kasus yang meleset dan 16 kasus berhasil, sehingga didapat akurasi sebesar 80%[2]. Pada tahun 2013, klasifikasi pola jurusan dilakukan dengan algoritma Bayesian namun hanya dapat mendapatkan nilai akurasi 66% [3]. Klasifikasi penjurusan pada tahun 2015 dilakukan dengan membandingkan hasil dari 3 algoritma yaitu *Decission Tree* (C4.5), *Naïve Bayes* dan *K-Means*. Dari penelitian tersebut didapat nilai akurasi *Decission Tree* paling tinggi yaitu 79,14%[4]. Selanjutnya klasifikasi penjurusan sma

dilakukan pada tahun 2016 menggunakan Algoritma C4.5 dengan menggunakan data tes akademik didapatkan nilai akurasi 89,74%[5].

Secara keseluruhan, hasil klasifikasi dari algoritma C4.5 dapat diterima dan memiliki nilai akurasi yang cukup tinggi. Mengingat kemampuan model algoritma C4.5 yang dapat melakukan proses learning dan klasifikasi yang sederhana, cepat dan secara umum mempunyai tingkat akurasi yang tinggi[6]. Maka dari itu, padapenelitian ini digunakan algoritma klasifikasi data mining C4.5 untuk mengklasifikasi pola minat siswa-siswi dalam menentukan pilihan jurusan yang akan mereka ambil di SMK.

2. METODE PENELITIAN

Ada beberapa tahap yang dilakukan untuk penelitian ini, penulis mengikuti *model Cross-Standart Industry for Data Mining (CRIPS-DM)* [7], seperti yang digambarkan pada gambar 1 dibawah ini.



Gambar 1. Proses Data Mining Menurut CRISP-DM

Model tersebut terdiri dari 6 tahap, diantaranya :

1. Tahap *Business Understanding*

Penelitian pendahuluan ini dilakukan dengan mengadakan wawancara dengan pihak yang bersangkutan atau yang bertanggung jawab langsung di bidang penjurusan SMK untuk mendapatkan data yang diperlukan dalam penulisan Skripsi ini. Terdapat beberapa kesalahan dalam proses penentuan jurusan di SMK karena belum adanya klasifikasi pola minat siswa dalam menentukan pilihan jurusan, sehingga pihak sekolah masih sulit dalam menentukan kebijakan untuk penentuan jurusan calon siswa-siswi nya.

2. Tahap *Data Understanding*

Data diperoleh dari salah satu Sekolah Menengah Kejuruan di Kota Tasikmalaya dari data angkatan 2013 sampai dengan 2016. Atribut yang dipakai meliputi pilihan jurusan I, pilihan jurusan II, nilai rapor matematika, bahasa indonesia, bahasa inggris, IPA, tes kesehatan, dan tes olahraga.

2.3. Tahap *Data Preparation*

Setelah memasuki tahap *Data Understanding*, kemudian data mulai memasuki tahap preprocessing, yaitu :a. *Data Cleaning*, yaitu proses untuk membersihkan nilai yang kosong, tidak konsisten atau mungkin tupel yang kosong (*missing values* dan *noisy*); b. *Data integration*, yaitu

proses untuk menyatukan tempat penyimpanan (arsip) yang berbeda-beda kedalam satu data.;c. Data reduction, tupel dalam data set mungkin terjadi duplikasi atau terdapat tupel yang sama, sehingga untuk memperkecil jumlah tupel, tupel yang sama dijadikan dalam satu tupel untuk mewakili tupel tersebut.

4. Tahap *Modelling*

Computing approach pada penelitian ini dipilih berdasarkan studi literatur tentang algoritma C4.5 yang mampu mengklasifikasikan pola penjurusan SMK.

5. Analisa dan *Evaluation Pattern*

Algoritma yang telah diterapkan dalam penelitian ini akan diterapkan pada data penjurusan siswa SMK melalui suatu model simulasi. Evaluasi dilakukan dengan mengamati hasil klasifikasi dari penerapan algoritma C4.5. Pengukuran tingkat akurasi dilakukan dengan menggunakan evaluasi model confusion matrix. Dengan demikian dapat diketahui berapa tinggi akurasi dari algoritma tersebut.

6. *Deployment phase*

Dari bentuk pohon keputusan yang dihasilkan pada tahapan modeling kemudian diekstraksi sejumlah aturan untuk diterapkan sebagai pengajuan kebijakan bagi pihak sekolah.

3. HASIL DAN PEMBAHASAN

Tahap ini biasanya disebut juga dengan tahap *learning*, karena pada tahap ini data training diklasifikasikan oleh model dan kemudian menghasilkan sejumlah pola. Model yang digunakan dalam tahap ini menggunakan algoritma C4.5. Seperti yang telah dijelaskan sebelumnya, ada beberapa tahap yang harus dilalui dalam membentuk pohon keputusan atau *decission tree*, tentunya algoritma C4.5 digunakan untuk membuat pohon keputusan tersebut.

Langkah pertama adalah menghitung jumlah class untuk hasil DITERIMA dan TIDAK DITERIMA. Kemudian menghitung nilai entropy dari semua class yang dibagi berdasarkan atribut Bahasa Inggris, Bahasa Indonesia, Matematika, IPA, Tes Kesehatan dan Tes Olahraga. Kemudian dilakukan perhitungan nilai information gain untuk setiap atribut, karena nilai gain tertinggi lah yang akan digunakan sebagai akar. Karena untuk menghitung nilai information gain membutuhkan entropy dari setiap kasus pada atribut tersebut maka entropy harus diketahui terlebih dahulu. Baris kolom *Entrop* dihitung dengan persamaan 1 sebagai berikut:

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i \quad (1)$$

Keterangan :

S : himpunan kasus

Pi : proporsi dari Si terhadap S

N : jumlah partisi S

Selanjutnya mencari Gain untuk tiap-tiap atribut dengan menggunakan persamaan 2.

$$\text{Gain(S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy (Si)} \quad (2)$$

Keterangan :

S :himpunankasus

A : atribut

N : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

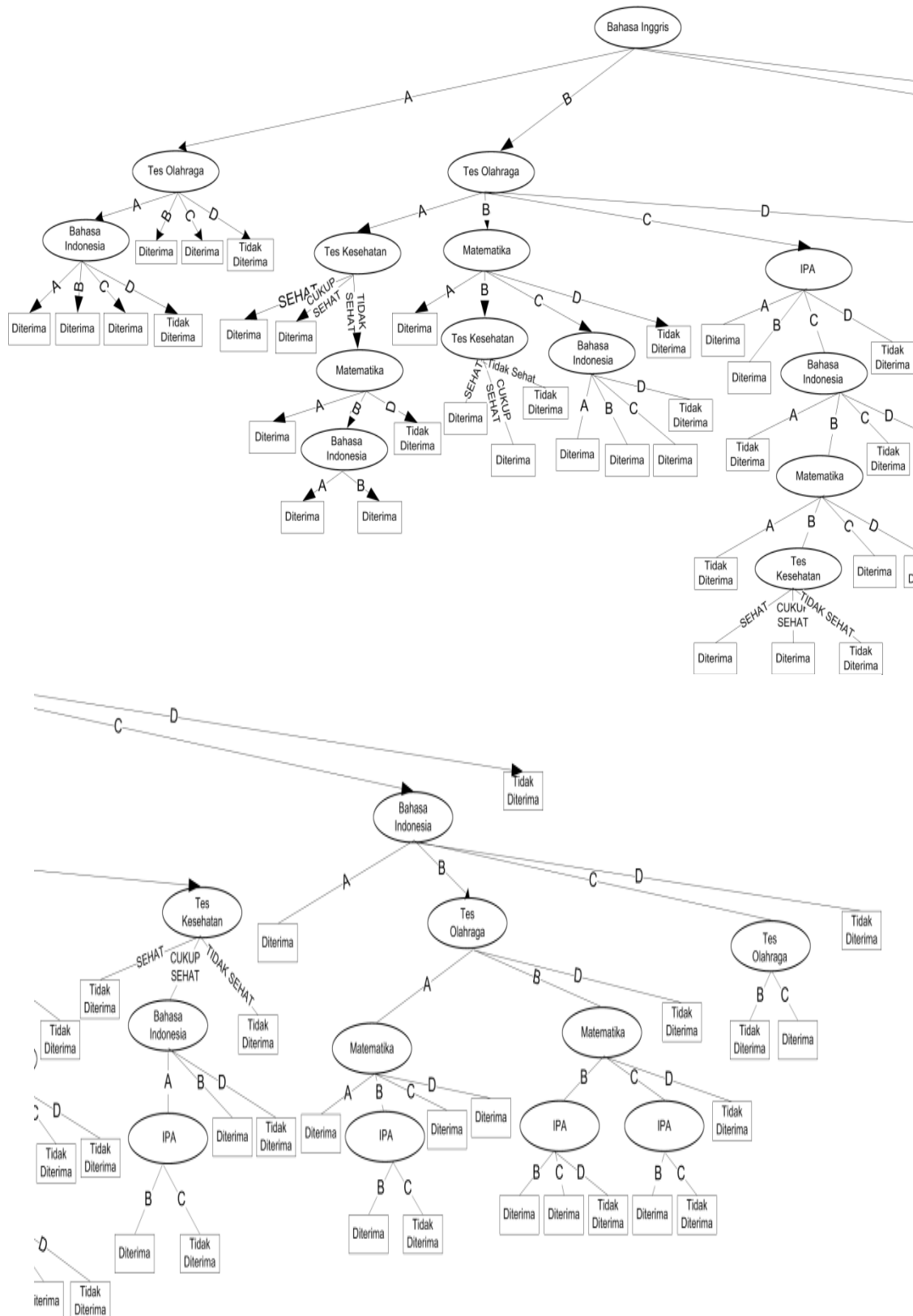
Contoh untuk atribut pilihan jurusan 1, dengan menggunakan persamaan 1, dihitung Entropy untuk tiap kategori dari atribut pilihan jurusan 1. Kemudian hasil perhitungan Entropy tiap atribut jurusan 1 dimasukkan pada persamaan 2 untuk menghitung Gain Information pilihan jurusan 1. Hasil perhitungana dapat dilihat pada table 1.

Tabel 1. Hasil Perhitungan Entropy dan Gain Informasi untuk Node 1

Node	Atribut	Kategori	JumlahKasus	Diterima	TidakDiterima	Entropy	Gain Information	Keterangan
				S_i	S_i			
1	Total		709	611	98	0,57956562		
	Bahasa Inggris							
		A	195	189	6	0,19823496	0,152868168	Gain Information Tertinggi
		B	448	402	46	0,47742941		
		C	52	20	32	0,9612366		
		D	14	0	14	0		
	Bahasa Indonesia							
		A	304	296	8	0,17556503	0,099394016	
		B	372	306	66	0,67439889		
		C	20	9	11	0,99277445		
		D	13	0	13	0		
	Matematika							
		A	107	105	2	0,1340298	0,099394016	
		B	452	412	40	0,43142821		
		C	133	91	42	0,89974376		
		D	17	3	14	0,67229482		
	IPA							
		A	132	130	2	0,1132743	0,078604766	
		B	421	378	43	0,47573576		
		C	148	101	47	0,90171664		
		D	8	2	6	0,81127812		
	TesKesehatan							
		SEHAT	499	440	59	0,52427033	0,053217908	
		CUKUP SEHAT	167	153	14	0,41554077		
		TIDAK SEHAT	43	18	25	0,98079836		
	TesOlahraga							
		A	228	213	15	0,35001059	0,080964074	
		B	310	279	31	0,46899559		
		C	138	110	28	0,72768745		
		D	33	9	24	0,84535094		

Dari table1 diatas, didapatkan bahwa Bahasa Inggris mendapatkan gain information tertinggi, maka pada atribut inilah yang dijadikan node pertama.

Hasil pengujian menghasilkan pohon keputusan sesuai dengan pengujian yang sudah dilakukan dan menunjukkan akurasi prediksi penjurusan siswa baru SMK khusus untuk jurusan Teknik Komputer Jaringan (TKJ) menggunakan algoritma C4.5 yang dimodelkan dengan pohon keputusan seperti gambar 2 dibawah ini.



Gambar 2. Pohon Keputusan Klasifikasi Pola Penjurusan Siswa SMK

Langkah selanjutnya dilakukan validasi serta pengukuran keakuratan hasil yang dicapai oleh model menggunakan beberapa teknik yang terdapat dalam *framework* Rapid Miner yaitu, Confusion Matrix untuk pengukuran tingkat akurasi model, dan *split validation* dengan mengimport data. Selanjutnya pengujian dilakukan menggunakan Algoritma C4.5 untuk melihat akurasi, pohon keputusan dan *rule*. Berdasarkan hasil dari proses Rapid Miner didapatkan nilai akurasi dan *confussion matrix* yang dapat dilihat pada gambar 3.

accuracy: 92.81% +/- 1.71% (mikro: 92.81%)			
	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	594	34	94.59%
pred. TIDAK DITERIMA	17	64	79.01%
class recall	97.22%	65.31%	

Gambar 3. Nilai Akurasi dan *Confussion Matrix*

Dari *confussion matrix* diatas didapat nilai : *True positive* (TP) sebanyak 594 tupel b) *False negative* (FN) sebanyak 17 tupel c) *False positive* (FP) sebanyak 34 tupel; d) *True negative* (TN) sebanyak 64 tupel. Akurasi adalah hasil perhitungan dari jumlah kasus yang diprediksi DITERIMA yang benar DITERIMA (*true positive*) dan diprediksi TIDAK DITERIMA dibandingkan dengan seluruh kasus yang ada.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{594 + 64}{594 + 64 + 17 + 34} = 92,81$$

Selain akurasi, untuk mengevaluasi kinerja model klasifikasi dapat digunakan sensitivitas dan spesifitas sebagai ukuran statistic kinerja klasifikasi biner. Sensitivitas, disebut juga *true positive rate* (TP rate) atau *recall*. Mengukur proporsi kasus “DITERIMA” yang diidentifikasi dengan tepat. Sedangkan spesifikasi mengukur kasus “TIDAK DITERIMA” yang diidentifikasi dengan tepat. Secara teori *classifier* secara optimal dapat mencapai nilai keduanya 100%, tapi secara prakteknya ini sangat tidka mungkin.

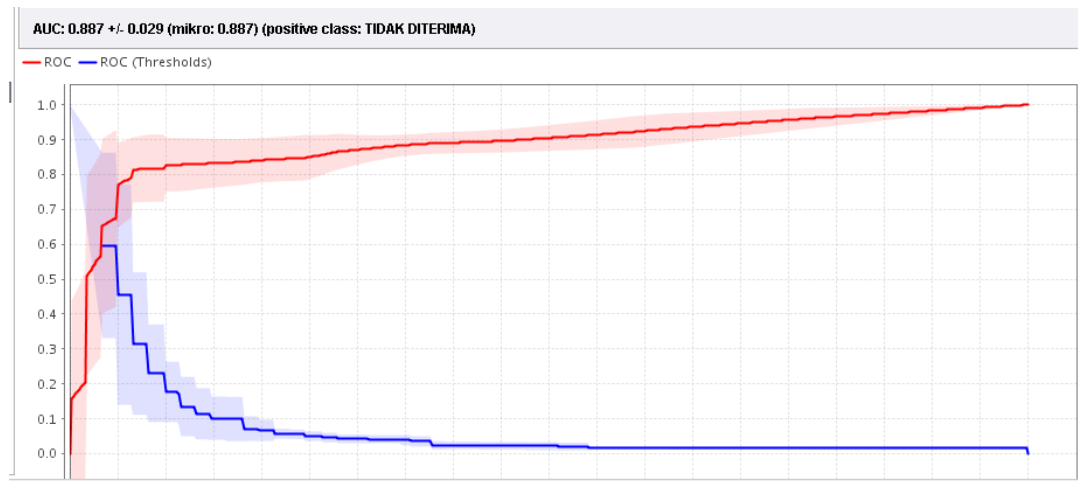
$$\text{Sensitivity or Recall} = \frac{TP}{TP + FN} = \frac{594}{594 + 17} = 97,22$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{64}{64 + 34} = 65,31$$

Tabel 2. Tabel nilai *Precision*, *Accuracy* dan *Recall* untuk Algoritma C4.5

	Nilai (%)
Precision	65,31 %
Accuracy	92,81 %
Recall	97,22 %

Kurva ROC menunjukkan *trade-off* antara *true positive rate* (proporsi tuple positif yang teridentifikasi salah sebagai positif) dalam suatu model. untuk mengukur area dibawah kurva ROC. Gambar 4. menunjukkan grafik ROC dengan nilai AUC (*Area Under Curve*) dengan C4.5 sebesar 0,893. Akurasi AUC dikatakan sempurna apabila nilai AUC mencapai 1000 dan akurasinya buruk jika nilai AUC dibawah 0,500.



Gambar 4. Kurva AUC

Dengan kurva ROC, kita dapat melihat trade off antara tingkat dimana suatu model dapat mengenali tuple positif secara akurat dan tingkat dimana model tersebut salah mengenali tuple negatif sebagai tuple positif. Kurva ROC terdiri atas sumbu vertikal yang menyatakan true positive rate, dan sumbu horizontal yang menyatakan false positive rate. Jika memiliki true positif (sebuah tupel positif yang benar diklasifikasikan) maka pada kurva ROC bergerak ke atas dan plot titik. Sebaliknya, jika tupel milik kelas “tidak” ketika memiliki false positif, maka kurva ROC bergerak kekanan dan plot titik. proses ini diulang untuk setiap tupel tes (setiap kali bergerak ke atas kurva untuk true positif atau terhadap hak untuk false positif).

4. KESIMPULAN

Dengan dihasilkannya klasifikasi pola penjurusan SMK dengan akurasi 97,22% merupakan tingkat akurasi yang baik, sehingga kelancaran proses penjurusan minat siswa dapat tercipta dan pola klasifikasi penentuan jurusan berdasarkan minat siswa sudah terprediksi dari awal untuk dapat diwaspadai agar tidak terjadi kesalahan dan kesulitan dalam menentukan kebijakan untuk penentuan jurusan calon siswa-siswinya. Tingkat akurasi algoritma C4.5 pada penelitian ini sudah mencapai tingkat baik, sehingga dapat meningkatkan ketelitian dalam proses klasifikasi penentuan jurusan berdasarkan minat siswa.

Namun, penelitian ini belum dapat menyelesaikan klasifikasi dengan banyak kelas, baru 2 kelas saja yaitu DITERIMA dan TIDAK DITERIMA pada jurusan tertentu. Sedangkan melihat proses dilapangan proses penjurusan dilakukan secara serempak untuk seluruh jurusan yang ada di SMK tersebut. Sehingga untuk penelitian selanjutnya dapat melakukan klasifikasi pola penjurusan siswa smk dengan multiclass.

5. SARAN

Saran-saran untuk untuk penelitian lebih lanjut untuk menutup kekurangan penelitian. Tidak memuat saran-saran diluar untuk penelitian lanjut.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kemenristekdikti yang telah memberi dukungan *financial* terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] Liliana Swastina, "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa," *Gema Aktualita*, Juni 2013
- [2] E. Prasetyo, *Data Mining : Konsep dan Aplikasi menggunakan MATLAB*. Yogyakarta: ANDI, 2012.
- [3] O. Kristanto, "PENERAPAN ALGORITMA KLASIFIKASI DATA MINING ID3 UNTUK MENENTUKAN PENJURUSAN SISWA SMAN 6," p. 9, 2013.
- [4] S. G. Winanta Send, Oslan Yetli, "IMPLEMENTASI METODE BAYESIAN DALAM PENJURUSAN DI SMA BRUDERAN PURWOREJO STUDI KASUS: SMA BRUDERAN PURWOREJO," vol. 6, pp. 21–28, 2013.
- [5] Y. S. Nugroho and S. N. Haryati, "Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali," *Khazanah Inform. J. Ilmu Komput. Inform.*, vol. 1, no. 1, pp. 3–8, 2015.
- [6] 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. 4, no. 3, pp. 75–84, 2016.
- [7] Liliana Swastina, "PENERAPAN ALGORITMA C4.5 UNTUK PENENTUAN JURUSAN MAHASISWA," VOL. 2 No. 1 Juni 2013
- [8] Farid Ali Prastya, "PENERAPAN ALGORITMA C4.5 UNTUK PREDIKSI JURUSAN SISWA SMAN 3 REMBANG", 2014
- [9] J. Han and M. Kamber, "Data Mining: concepts and techniques slides," *Techniques*. 2006.
- [10] D. T. Larose, *Discovering Knowledge in Data - An Introduction to Data Mining*. Canada: A JOHN WILEY & SONS, INC., PUBLICATION, 2005.