PENGUJIAN KINERJA ALGORITMA C4.5 UNTUK PREDIKSI PEMILIHAN MINAT LANJUTAN PADA SISWA SMK

Etika Wahyu Perdani[1)], Kusrini[2)], Hanif Al Fatta[3)]
[1)] "S2 Magister Teknik Informatika" Universitas Amikom Yogyakarta
Email : tkwahyu845@gmail.com[1)], kusrini@amikom.ac.id[2)], hanif.a@amikom.ac.id[3)]

Abstract

In an effort to improve the quality of education in SMK Negeri 1 Cilacap, schools need a tool used to predict student interest after graduating from vocational school. So through the tool the teacher counseling guidance can provide guidance to students according to their interests.
This research discusses the process of decision tree development using C4.5 algorithm and uses attributes such as academic value, parent's income, PKL value, organizational experience, achievement of student during school and students interest. The decision tree is then interpreted into the form of easy-to-understand decision rules and is used as a reference for predicting continued interest in students after graduating from vocational school.
The test results were obtained by conducting two experiments that used five attributes and six attributes and used the Forward Selection feature to find out the most influential variables in the study. From the test data it was concluded that experiments using six attributes (including student interest attributes) gave better results than experiments with only five attributes. The highest level of accuracy generated when the data reached 2050 data, the accuracy of 83.71%.The most influential variable in the experiment was PKL.

Keywords: decision tree, C4.5, student

Abstrak

Dalam upaya peningkatan mutu pendidikan di SMK Negeri 1 Cilacap, sekolah membutuhkan alat bantu yang digunakan untuk memprediksi minat siswa setelah lulus dari SMK. Sehingga melalui alat tersebut guru bimbingan konseling dapat memberikan bimbingan kepada siswa sesuai dengan minatnya.Penelitian ini membahas proses pembangunan pohon keputusan yang menggunakan algoritma C4.5 dan menggunakan atribut seperti nilai akademik, penghasilan orang tua, nilai PKL (Praktek Kerja Lapangan), pengalaman organisasi, prestasi yang diraih siswa selama di sekolah serta minat siswa. Pohon keputusan tersebut kemudian diinterpretasikan kedalam bentuk aturan-aturan keputusan yang mudah dipahami dan digunakan sebagai acuan untuk memprediksi minat lanjutan pada siswa setelah lulus SMK. Hasil pengujian diperoleh dengan melakukan dua percobaan yaitu percobaan yang menggunakan lima atribut dan enam atribut serta menggunakan fitur Forward Selection untuk mengetahui variabel yang paling berpengaruh dalam penelitian. Dari data pengujian tersebut diambil kesimpulan bahwa percobaan yang menggunakan enam atribut (termasuk atribut minat siswa) memberikan hasil yang lebih baik dibandingkan dengan percobaan yang hanya menggunakan lima atribut saja. Tingkat akurasi tertinggi dihasilkan saat jumlah datanya mencapai 2050 data dan menghasilkan akurasi sebesar 83,71%. Variabel yang berpengaruh dalam percobaan tersebut adalah PKL.

Kata kunci: decision tree, C4.5, siswa

1. PENDAHULUAN

Pendidikan sangat penting sebagai salah satu faktor pendorong pembangunan sebagai sumber daya manusia dengan tujuan meningkatkan kemampuan pada masyarakatnya dalam mengembangkan ilmu pengetahuan [1]. Siswa di SMK mengalami berbagai hambatan dalam menentukan minatnya setelah lulus dari sekolahnya. Permasalahan yang muncul adalah

bagaimana melakukan klasifikasi yang dapat digunakan untuk memprediksi minat siswa setelah lulus dari SMK agar prediksi tersebut dapat dimanfaatkan untuk memberikan bimbingan kepada siswa.

Oleh karena itu dalam upaya peningkatan mutu pendidikan dan mutu lulusan di SMK, sekolah membutuhkan alat bantu yang digunakan untuk memprediksi minat siswa setelah lulus dari SMK. Berdasarkan hal tersebut peneliti melakukan penelitian dengan menggunakan algoritma C4.5 agar mampu memberikan prediksi sehingga dapat digunakan guru BK dalam membimbing siswa untuk menentukan minatnya yaitu bekerja atau melanjutkan kuliah setelah lulus SMK.

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database yang besar[2]. Salah satu metode dalam data mining yaitu klasifikasi. Dalam pengelompokkan klasifikasi terdapat metode *decision tree* dengan algoritma C4.5.

Sementara itu rumusan masalah dalam penelitian ini adalah mencari hasil akurasi tertinggi dan mengetahui atribut apakah yang berpengaruh dalam penelitian, kemudian diberikan penambahan atribut baru yaitu minat siswa dan dianalisis kembali hasil akurasinya beserta variabel yang mempengaruhi penelitian tersebut.

Untuk penentuan semacam itu, telah banyak dilakukan penelitian yang membandingkan teknik klasifikasi algoritma C4.5 dengan Naïve Bayes untuk memprediksi kemampuan kognitif siswa yang dievaluasi dengan melakukan tes online. Atribut prediksi yang digunakan yaitu *Logical Reasoning*, *Numerical Ability*, *Perceptual Speed*, *Hours spend to study*, dan *Resource* [3]. Hasil klasifikasi menunjukkan bahwa kinerja C4.5 memberikan tingkat akurasi yang lebih baik dari Naïve Bayes yaitu 0,938.

Katore [4] menggunakan algoritma C4.5 untuk merancang prediksi dan sistem rekomendasi pilihan karir yang akurat dengan menggunakan pendekatan data mining dan algoritma stastistik. Pendekatan algoritma yang digunakan adalah algoritma C 4.5 dan terdapat 12 atribut data meliputi Kerja Tim, Keterampilan Teknis, Kejujuran, Pengambilan Keputusan, Kepemimpinan, Kedisiplinan, Kemampuan Adaptasi, Komunikasi, Sikap, Verbal, Tanggung Jawab dan Kekuatan. Hasil eksperimen dan evaluasi menunjukkan bahwa keakuratan algoritma C 4.5 sebesar 86%, algoritma Naïve Bayes 84%, K-Star 82% dan Simple Cart 80% dengan sampel sebanyak 110 siswa.

Berdasarkan pertimbangan di atas, pendekatan data mining dengan penerapan algoritma Decision Tree C4.5 akan dilakukan untuk memprediksi minat lanjutan pada siswa setelah lulus dari SMK. Dengan demikian diharapkan penelitian ini mampu menjadi alat bantu yang digunakan oleh pihak sekolah dalam mengembangkan potensi siswa.

2. KAJIAN LITERATUR

Beberapa penelitian sebelumnya yang digunakan sebagai referensi dari penelitian yang akan dilakukan adalah sebagai berikut : Mayilvaganan, dalam penelitian yang berjudul "*Comparison Of Classification Techniques Predicting the Cognitive Skill of Students Education Environment*" [3]. Penelitian ini bertujuan membandingkan teknik klasifikasi algoritma C4.5 dengan Naïve Bayes untuk memprediksi kemampuan kognitif siswa yang dievaluasi dengan melakukan tes online. Atribut prediksi yang digunakan yaitu *Logical Reasoning*, *Numerical Ability*, *Perceptual Speed*, *Hours spend to study*, dan *Resource*. Hasil klasifikasi menunjukkan bahwa kinerja C4.5 memberikan tingkat akurasi

yang lebih baik dari Naïve Bayes yaitu 0,938.

Penelitian berikutnya yaitu Mishra dkk, dalam penelitiannya yang berjudul "*Mining Student's Data for Performance Prediction*"[5]. Penelitian ini bertujuan untuk membangun model prediksi kinerja berdasarkan integrasi sosial siswa, integrasi akademis dan berbagai keterampilan emosional yang belum pernah dipertimbangkan. Penelitian ini menggunakan perbandingan dua algoritma yaitu J48 (implementasi dari C4.5) dan Random Tree. Makalah ini berfokus pada identifikasi atribut yang mempengaruhi kinerja selama tiga semester pada siswa. Pengaruh parameter hasil kecerdasan emosi pada penempatan telah ditetapkan. Random Tree memberikan akurasi prediksi yang lebih tinggi daripada J 48 yaitu sebesar 94.41%.

Penelitian berikutnya yaitu Katore, dalam penelitiannya yang berjudul "*Novel Professional Career Prediction And Recommendation Method For Individual Through Analytics On Personal Traits Using C 4.5 Algorithm*"[4]. Penelitian ini bertujuan untuk merancang prediksi dan sistem rekomendasi pilihan karir yang akurat dengan menggunakan pendekatan data mining dan algoritma stastistik. Pendekatan algoritma yang digunakan adalah algoritma C 4.5 dan terdapat 12 atribut data meliputi Kerja Tim, Keterampilan Teknis, Kejujuran, Pengambilan Keputusan, Kepemimpinan, Kedisiplinan, Kemampuan Adaptasi, Komunikasi, Sikap, Verbal, Tanggung Jawab dan Kekuatan. Hasil eksperimen dan evaluasi menunjukkan bahwa keakuratan algoritma C 4.5 sebesar 86%, algoritma Naïve Bayes 84%, K-Star 82% dan Simple Cart 80% dengan sampel sebanyak 110 siswa.

Penelitian berikutnya yaitu Rodriguez, dalam penelitiannya yang berjudul "*Modeling Student Engagement by Means Of Nonverbal Behavior and Decision Tree*" bertujuan untuk meneliti apakah seorang siswa tertarik pada suatu kelas atau tidak yang diperoleh dari informasi perilaku ekspresi wajah siswa yaitu atribut wajah, mata, bahu, dan muulut dan akhirnya diklasifikasikan dengan atribut "tertarik", "tidak tertarik" atau "netral"[6]. Metode klasifikasi yang digunakan yaitu ID3, Random Tree, C4.5, BFTREE, dan REPTree. Serta menerapkan matriks F-Measure untuk mengevaluasi keputusan yang dihasilkan. Hasil penelitian menunjukkan bahwa algoritma Random Tree menghasilkan akurasi tertinggi yaitu 87.03% diantara algoritma lainnya. Sedangkan algoritma C4.5 menghasilkan akurasi yang lebih kecil yaitu 83.33% namun dapat menampung keseluruhan atribut dengan baik. Sementara itu hasil BFTree paling kecil yaitu 74.07% tetapi mempunyai kelebihan dengan cepat membuat inisial klasifikasi.

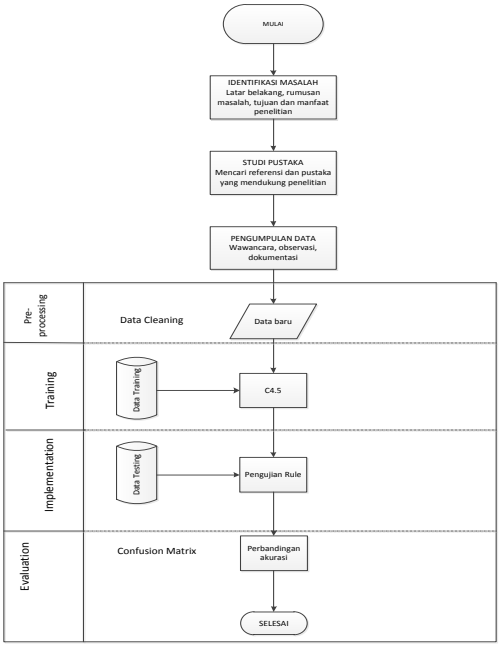
3. METODE PENELITIAN

Jenis penelitian yang dilakukan adalah penelitian eksperimen. Sifat dan pendekatan penelitian menggunakan metode deskriptif kuantitatif. Penelitian deskriptif kuantitatif adalah penelitian yang datanya diperoleh dari suatu sampel populasi yang dianalisis sesuai metode statistik yang digunakan kemudian hasilnya diinterpretasikan ke dalam angka yang memuat pengetahuan.

Ruang lingkup atau objek penelitian yaitu siswa di SMK Negeri 1 Cilacap yang beralamat di Jl. Budi Utomo No.10, Kabupaten Cilacap 53211, Jawa Tengah. Data penelitian diperoleh dengan cara melakukan observasi, pengumpulan dokumen dari pihak terkait dan studi pustaka yang relevan dengan penelitian.

Analisis data dimulai dari penentuan atribut data yang digunakan pada penelitian, didasarkan pada hasil observasi dan wawancara pada guru BK terkait hal yang berpengaruh saat siswa akan melamar pekerjaan ataupun mendaftar kuliah. Dari hasil tersebut diperoleh bahwa akademik, prestasi, nilai PKL serta penghasilan orang tua,

pengalaman organisasi serta minat siswa berpengaruh saat siswa akan melamar pekerjaan ataupun mendaftar kuliah. Sehingga dihasilkan 6 atribut untuk dilakukan olah data pada algoritma C4.5. untuk lebih lengkapnya disajikan pada Gambar 1 berikut :



Gambar 1. Alur Penelitian

Penjelasan alur penelitian gambar 1 di atas :

1. Identifikasi Masalah
   Melakukan identifikasi pada suatu masalah merupakan tahap awal dalam melakukan proses penelitian. Menentukan latar belakang masalah, merumuskan masalah, tujuan dan manfaat penelitian yang diharapkan.
2. Studi Pustaka
   Dilakukan dengan mempelajari dan memahami semua teori dan referensi yang berhubungan dengan penelitian yang akan diteliti, diantaranya meliputi Data Mining, Metode Decision Tree, Algoritma C 4.5, parameter yang berpengaruh, serta data-data yang diperlukan dalam penelitian.
3. Pengumpulan Data
   Tahap ini merupakan cara untuk mengumpulkan data yang dilakukan dengan observasi, wawancara dan dokumentasi dengan pihak-pihak yang terkait dengan penelitian ini yaitu pihak sekolah SMK Negeri 1 Cilacap.
4. Data Baru
   Pada tahap ini yaitu *pre-processing* dengan tahapan *data cleaning* (memilih data yang tidak memiliki *missing value* untuk digunakan menjadi data training) pada data sampel sehingga diperoleh data baru.
5. Data Training
   Kemudian data baru tersebut menjadi data training yang akan dilakukan proses algoritma C4.5.
6. Data Testing
   Setelah rule terbentuk maka akan dilakukan pengujian menggunakan data testing.
7. Perbandingan akurasi
   Pada tahapan Evaluation, data testing akan diproses sehingga terbentuk hasil, yang kemudian akan dilakukan perbandingan akurasi dari beberapa percobaan.

4. HASIL DAN PEMBAHASAN

Implementasi pada penelitian ini menggunakan RapidMiner versi 7.4. Percobaan dalam penelitian ini ada dua pembagian, percobaan pertama hanya menggunakan 5 atribut, dan percobaan kedua menggunakan 6 atribut.

Berikut ini tabel 1 yang berisi mengenai atribut yang dipakai dalam penelitian :

Tabel 1. Pembagian Variabel dan Kelas Data

Variabel	Nama Field	Jenis Kelas Data	Kelas Data yang Digunakan
Y	Keputusan	Binominal	Kerja; Kuliah
X1	Akademik	Polynominal	Sangat baik,

Variabel	**Nama Field**	**Jenis Kelas Data**	**Kelas Data yang Digunakan**
			Baik, Cukup, Kurang
X2	Penghasilan	Polynominal	Sangat tinggi, Tinggi, Sedang, Rendah
X3	Organisasi	Polynominal	Tidak ikut, ikut satu, ikut lebih dari satu
X4	PKL	Polynominal	A, B, C, D
X5	Prestasi	Binominal	Ya, Tidak
X6	Minat	Binominal	Bekerja, Kuliah

Nilai kelas pada field Akademik dikategorikan berdasarkan Peraturan Menteri No.104 tentang Penilaian Hasil Belajar selama semester 1 sampai semester 5, yaitu:

1) Sangat baik 86-100
2) Baik 71-85
3) Cukup 56-70
4) Kurang <= 55

Nilai kelas pada field Penghasilan dikategorikan berdasarkan Penggolongan Gaji menurut BPS tahun 2014 yaitu :

1. Golongan Sangat Tinggi jika gaji > 3.500.000 per bulan
2. Golongan Tinggi jika gaji 2.500.000 – 3.500.000 per bulan
3. Golongan Sedang jika gaji 1.500.000-2.500.000 per bulan
4. Golongan Rendah jika gaji < 1.500.000 per bulan

Nilai kelas pada field PKL dikategorikan berdasarkan nilai PKL yang diperoleh dari Permendikbud No.103 tahun 2014 tentang Proses Pembelajaran pada Pendidikan Dasar dan Pendidikan Menengah, yaitu :

1) Amat baik (A) 90-100
2) Baik (B) 75-89
3) Cukup (C) 60-74
4) Kurang (D) <= 59

Berikut ini adalah hasil percobaan pada penelitian di atas :

Tabel 2. Hasil Percobaan

Banyaknya Data	Dengan 5 atribut		Dengan 5 atribut + minat	
	Tingkat akurasi	Root	Tingkat akurasi	Root
442 Data	79,86%	Prestasi	81,22%	Prestasi
500 Data	82,20%	Akademik	82,40%	Akademik
1025 Data	82,54%	PKL	83,51%	PKL
2050 Data	82,54%	PKL	83,71%	PKL
3050 Data	82,33%	Akademik	83,02%	Akademik

Dari hasil tabel 2 di atas dapat diketahui bahwa semakin banyak data yang digunakan dalam penelitian maka akan meningkatkan nilai prediksi/akurasi dari penelitian tersebut, sehingga dapat dikatakan percobaan tersebut dapat membuktikan bahwa algoritma C4.5 dapat menyelesaikan permasalahan prediksi siswa SMK yang akan kuliah maupun bekerja. Namun pada percobaan dengan data sebanyak 3050 mengalami penurunan akurasi dari yang sebelumnya, hal tersebut dikarenakan terdapat data yang nilainya sama (duplikat). Selain itu dengan menambahkan atribut baru yaitu atribut minat yang berisi keinginan siswa, hal tersebut dapat mempengaruhi tingkat akurasinya sehingga semakin membesar

daripada tingkat akurasi yang hanya menggunakan lima atribut saja.

Pada tabel 2 di atas juga dapat dilihat bahwa root (node teratas dari *decision tree*) berbeda-beda jika datanya bertambah banyak. Namun untuk pemberian atribut baru yaitu atribut minat, tidak berpengaruh terhadap root. Karena root/ node tertinggi dihitung dari infomasi gain yang tertinggi.

Atribut yang berpengaruh dalam penelitian menggunakan metode Forward Selection dapat disajikan pada tabel 3 berikut ini :

Tabel 3. Hasil Forward Selection

Banyaknya Data	Forward Selection (Pembobotan)	
	Dengan 5 atribut	Dengan 5 atribut + minat
442 Data	Atribut Prestasi	Atribut PKL
500 Data	Atribut Prestasi	Atribut Penghasilan, Organisasi, PKL
1025 Data	Atribut Organisasi	Atribut Penghasilan
2050 Data	Atribut Organisasi	Atribut Penghasilan
3050 Data	Atribut Penghasilan	Atribut Akademik

Pada pembobotan dengan fitur Forward Selection dapat dinyatakan bahwa setiap percobaan dapat memiliki pembobotan yang berbeda sesuai dengan banyaknya data maupun atribut yang digunakan.

5. KESIMPULAN

Berdasarkan hasil implementasi algoritma C4.5 di bagian sebelumnya dapat disimpulkan sebagai berikut :

a) Pada percobaan penelitian di atas dapat disimpulkan bahwa tingkat akurasi tertinggi yang paling mendekati dengan prediksi data yang sebenarnya yaitu 82,54% yaitu pada saat menggunakan 2050 data serta menggunakan lima atribut seperti akademik, PKL, organisasi, penghasilan dan prestasi. Atribut yang berpengaruh dalam percobaan tersebut yaitu Organisasi.
b) Pada percobaan berikutnya ditambahkan atribut minat, sehingga dalam penelitian terdapat enam atribut. Tingkat akurasi tertinggi dihasilkan saat jumlah datanya mencapai 2050 data. Tingkat akurasinya sebesar 83,71% dan atribut yang berpengaruh di dalam percobaan tersebut yaitu PKL.

DAFTAR PUSTAKA

C. A. V. Rodriguez, M. M. Lavalle, and R. P. Elias, "Modeling Student Engagement by Means of Nonverbal Behavior and Decision Trees," *Proc. - 2015 Int. Conf. Mechatronics, Electron. Automot. Eng. ICMEAE 2015*, pp. 81–85, 2016.

L. S. Katore, B. S. Ratnaparkhi, and J. S. Umale, "Novel professional career prediction and recommendation method for individual through analytics on personal traits using C4.5 algorithm," *Glob. Conf. Commun. Technol. GCCT 2015*, pp. 503–506, 2015.

L. Swastina, "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa," *Gema Aktual.*, vol. 2, no. 1, p. 6, 2013.

M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," *Commun. Netw. Technol. (ICCNT), 2014 Int. Conf.*

Comput. Intell. Comput. Res., pp. 113–118, 2014.

N. Zuwida, "Tinjauan Pemanfaatan Pemberian Beasiswa Bantuan Khusus Murid (BKM) Pada Siswa SMK Negeri 1 Pariaman," *Cived ISSN 2302-3341*, vol. 2, no. 3, pp. 389–394, 2014.

T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, pp. 255–262, 2014.