Penerapan Data Mining dengan Algoritma C4.5 untuk Penentuan Jurusan Sekolah Menengah Atas

Implementation of Data Mining with C4.5 Algorithm for Determining Senior High School

Rina Kuniasari [1)], Azizah Fatmawati[2)]

[1,2]Program studi Informatika, Fakultas Komunikasi dan Informatika, Universitas Muhammadiyah Surakarta
[1,2]Jl. Ahmad Yani, Pabelan, Kartasura, Surakarta 57102, Telp: +62 (271) 717417

E-mail: rinakurnia768@gmail.com[1)], af157@ums.ac.id[2)]

Abstrak – Penjurusan merupakan suatu proses penempatan atau penyaluran kemampuan, minat dan bakat dari siswa SMA. Penjurusan di SMA juga berpengaruh di masa depan seperti di jenjang karir maupun dalam pemilihan program studi di Perguruan Tinggi. Dalam kurikulum 2013 penjurusan sudah diterapkan pada siswa sejak kelas X. Namun masih banyak siswa yang masih bingung dalam memilih jurusan sesuai kemampuannya dan berakibat mengganggu proses pembelajaran nanti. Dengan demikian pihak sekolah harus tepat dalam mengklasifikasikan jurusan dan diperlukan sistem sebagai pendukung pemilihan jurusan bagi siswa. Penelitian ini mempunyai tujuan untuk membantu dalam penentuan jurusan dan meningkatkan keakuratan serta efisiensi dalam penentuan jurusan bagi siswa sehingga nanti dapat meminimalisir ketidakcocokan dalam penentuan jurusan. Metode yang digunakan dalam penelitian ini dengan penerapan data mining menggunakan Algoritma C4.5. Algoritma C4.5 ini akan membentuk pohon keputusan yang mengahasilkan aturan kemudian diterapkan ke data untuk mengklasifikasikan jurusan. Pengklasifikasian jurusan siswa dibagi menjadi 3 jurusan yaitu MIPA, IPS dan BB. Sedangkan kriteria yang dibutuhkan untuk penentuan jurusan meliputi rata-rata nilai UN SMP, minat siswa dan nilai tes akademik seperti nilai MIPA, nilai IPS, nilai BB. Hasil penelitian ini menghasilkan nilai accuracy sebesar 95% untuk melakukan proses penjurusan bagi siswa.

Kata Kunci: algoritma c4.5, data mining, penjurusan SMA

Abstract *– Majoring is a process of replacement or distribution the abilities, interests, and talents of high school students. Majors in high school become an influence in the future such as in the career paths and selection of study programs at universities. In the curriculum 2013, the majors have been applied to students since tenth grade. However, there are still many students who are still confused in choosing majors based on their abilities and the result of disruption on the learning process later. Thus, school have to concern correctly in classify the department and the system. They are needed as a support to the selection of majors for students. This study aims to assist the determination of majors and improve accuracy and efficiency in determining majors for students, so that later can minimize the incompatibility in majors. The method used the application of data mining using C4.5 Algorithm. This C4.5 algorithm will form a decision tree that results in a rule. Then, it is applied to the data to classify the department. Classification of student majors is divided into 3 majors namely MIPA, IPS and BB. While, the criteria required for determining majors includes the average SMP National Examination score, student interest, and academic test scores such as MIPA scores, IPS scores, BB scores. The results of this study resulted in an accuracy value of 95% for the process of majors for students.*

Keywords: *C4.5 Algorithm, Data Mining, High School Majors*

PENDAHULUAN

Pada jenjang Sekolah Menengah Atas (SMA) dilakukan penjurusan untuk membantu siswa dalam mengetahui dan menyalurkan minat, bakat dan kemampuannya. Proses penjurusan merupakan tahap penting yang digunakan untuk mengelompokkan sisw berdasarkan kemampuan (nilai), bakat dan minat supaya selanjutnya dalam proses pembelajaran di kelas yang diberikan kepada siswa dapat lebih fokus dan terarah (Novianti, Rismawan, & Bahri, 2016). Kurikulum yang diterapkan dalam SMA menggunakan kurikulum 2013, dimana penjurusan siswa dilakukan ketika kelas X dan dari siswa sendiri masih bingung menentukan pilihan sehingga tidak sedikit yang masih asal-asalan dan ikut-ikutan temannya dalam memilih jurusan. Penjurusan SMA memiliki 3 jurusan untuk siswa yaitu Matematika dan Ilmu Pengetahuan Alam (MIPA), Ilmu Pengetahuan Sosial (IPS), dan Bahasa dan Budaya (BB). Faktor utama yang dijadikan kriteria untuk penentuan jurusan

antara lain nilai rata-rata UN SMP, minat siswa, nilai tes akademik seperti nilai MIPA, nilai IPS dan nilai BB.

Penjurusan yang dilakukan sekolah saat ini masih menggunakan sistem manual dengan tulis tangan, penyimpanan dan pengklasifikasian data yang kurang rapi dan belum menggunakan sistem terkomputerisasi sehingga menjadi kurang efektif dan efisien. (Nugroho, 2014), menyatakan bahwa kegiatan pengklafikasian yang dilakukan secara manual yang dilakukan oleh manusia masih mempunyai keterbatasan, terutama pada kemampuan manusia dalam menampung jumlah data yang ingin diklasifikasikan dan bisa juga terjadi kesalahan dalam pengklasifikasian yang dilakukan.

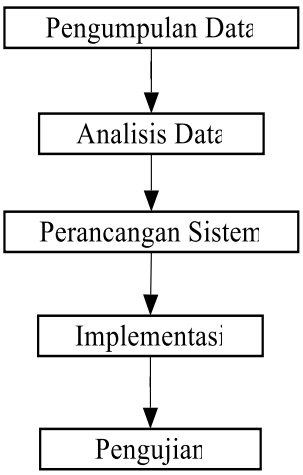
Salah satu cara pengklasifikasian data yaitu menggunakan teknik data mining. Data Mining didefinisikan sebagai sebuah proses untuk menemukan hubungan, pola dan tren baru yang bermakna melalui penyaringan data yang sangat besar, yang tersimpan dalam penyimpanan, menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika (Kamagi & Hansun, 2014). Menurut (Kristanto, 2013), data mining dapat digunakan untuk mengelompokkan data, memprediksi, mengestimasi, dan menentukan kaidah asosiasi dalam suatu data yang ada.

Penerapan data mining yang digunakan dalam penelitian ini menggunakan Algoritma C4.5. Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan yang mempunyai kelebihan dapat menghasilkan pohon keputusan yang mudah diinterpresentasikan, mempunyai tingkat akurasi yang dapat diterima, dan dapat menangani atribut bertipe diskret maupun numerik (Sirait & Hansun, 2017). Berdasarkan penelitian yang dilakukan (Wahyuni, 2018), Algoritma C4.5 dapat memberikan informasi atau pengetahuan tentang aturan yang mudah dipahami karena dideskripsikan dengan pohon keputusan. (Swastina, 2013) dalam penelitiannya tentang penentuan jurusan mahasiswa menghasilkan eksperimen dan evaluasi yang menunjukkan bahwa Algoritma Decision Tree C4.5 akurat diterapkan untuk penentuan kesesuaian jurusan mahasiswa dengan tingkat akurasi 93,31% dan akurasi rekomendasi jurusan sebesar 82,64%. Sehingga Algoritma C4.5 dapat diterapkan dalam penelitian yang dilakukan peneliti karena menghasilkan akurasi yang cukup tinggi.

Dengan dilatarbelakangi masalah tersebut, pada penelitian ini akan dikembangkan sebuah sistem untuk membantu dalam penentuan jurusan. Dengan sistem tersebut diharapkan dapat menampilkan hasil data mining dengan Algoritma C4.5. yang berupa keputusan untuk penentuan penjurusan secara tepat bagi siswa. Keputusan yang dihasilkan tersebut dapat berupa jurusan MIPA, IPS dan BB.

METODOLOGI PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini memiliki beberapa tahapan yang ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Metodologi Penelitian

1. Pengumpulan Data

Pengumpulan data diperoleh berdasarkan hasil wawancara dengan pihak SMA Negeri 6 Surakarta mengenai data dan informasi yang dibutuhkan dalam pengembangan sistem. Data yang dijadikan bahan penelitian ini adalah data siswa kelas X yang berjumlah 280 data. Data tersebut terdiri dari 3 kelompok kelas yaitu kelas Matematika dan Ilmu Pengetahuan Alam (MIPA), Ilmu Pengetahuan Sosial (IPS) dan Bahasa dan Budaya (BB).

2. Analisis Data

Pada tahapan analisis data setelah didapatkan data kemudian melakukan proses analisis untuk menentukan kriteria yang dibutuhkan untuk penentuan jurusan siswa. Kriteria tersebut dijadikan sebagai atribut yang antara lain rata-rata nilai UN, minat siswa, nilai tes akademik seperti nilai mipa, nilai ips, dan nilai bb. Masing-masing atribut memiliki tipe dan memiliki beberapa *value*. Value tersebut ada yang memiliki *range* rendah dengan nilai 0-50, r*ange* sedang dengan nilai 51-75 dan *range* tinggi dengan

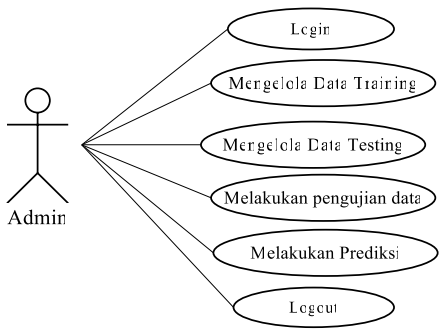
nilai 76-100. Penjelasan atribut data siswa ditunjukkan pada Tabel 1.

Tabel 1. Penjelasan atribut yang dibutuhkan

Atribut	Tipe	*Value* (Nilai)
Rata-rata Nilai	Polinomial	a. Rendah (0-50) b. Sedang (51-75) c. Tinggi (76-100)
Minat	Polinomial	a. MIPA b. IPS c. BB
Nilai MIPA	Polinomial	a. Rendah (0-50) b. Sedang (51-75) c. Tinggi (76-100)
Nilai IPS	Polinomial	a. Rendah (0-50) b. Sedang (51-75) c. Tinggi (76-100)
Nilai BB	Polinomial	a. Rendah (0-50) b. Sedang (51-75) c. Tinggi (76-100)
Jurusan	Label	a. MIPA b. IPS c. BB

3. Perancangan Sistem

Pada tahap perancangan sistem diperlukan *use case* untuk menggambarkan penggunaan dan pengelolaan data dalam sistem. *Use case* yang digunakan dalam sistem ini ditunjukkan pada Gambar 2.



Gambar 2. Use Case Diagram

Pada *use case* tersebut pihak yang berperan sebagai *admin* adalah pihak sekolah. *Admin* dapat melakukan aktivitas yang berkaitan dengan pengelolaan sistem seperti melihat, menambah, dan menghapus data dari sistem.

4. Implementasi

Pada tahap ini dilakukan pengembangan sistem yang dibangun menggunakan bahasa pemrograman PHP *(Hypertext Prepocessor)* dengan *framework Code Igniter* dan menggunakan tools seperti *xampp*, *sublime* sebagai *text editor* dan *Google chrome* atau *mozila firefox* sebagai *web browser*. Sedangkan untuk penyimpanan *database* menggunakan *MySQL*. Pengembangan sistem ini akan diterapkan menggunakan algoritma C4.5. Secara umum Algoritma C4.5 membangun keputusan pohon keputusan sebagai berikut (Nasari, 2014):

a. Perhitungan *Entropy* dan *Gain*
b. Pemilihan *Gain* tertinggi sebagai akar (*Node*)
c. Ulangi proses perhitungan *Entropy* dan *Gain.* untuk mencari cabang sampai semua kasus pada cabang memiliki kelas yang sama yaitu pada saat semua variable telah menjadi bagian dari pohon keputusan atau masing-masing variable telah memiliki keputusan.
d. Membuat *Rule* berdasarkan pohon keputusan.

Sebelum mendapatkan nilai *Gain,* terlebih dulu mencari *Entropy*. Perhitungan *Entropy* dirumuskan pada persamaan berikut:

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

Keterangan:
S = Himpunan kasus
n = Jumlah partisi S
P_i = proporsi S_i terhadap S

Untuk memilih atribut sebagai akar didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada dan dirumuskan sebagai berikut:

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

Keterangan:
S = Himpunan kasus
A = Fitur
n = jumlah partisi atribut A
$|Si|$ = proporsi Si terhadap S
$|S|$ = jumlah kasus dalam S

Data training yang digunakan dalam penelitian ini berjumlah 198 data. Jumlah keseluruhan data dengan kriteria jurusan MIPA sebanyak 80 data, jurusan IPS sebanyak 90 data dan jurusan BB sebanyak 26 data. Berikut contoh hasil perhitungan menggunakan algoritma c4.5 untuk mencari akar atau node pertama ditunjukkan pada Gambar 3.

Atribut	Nilai	Jumlah Kasus	MIPA	IPS	BB	*Entropy*	*Gain*
Total		198	80	90	26	1,42985036	
Minat	MIPA	71	55	12	4	0,952643282	
	IPS	83	15	66	3	0,8691615	**0,386628284**
	BB	42	11	12	19	1,540319314	
Rata-rata Nilai	Rendah	9	0	9	0	0	
	Sedang	83	30	46	7	1,30344844	**0,090036005940326**
	Tinggi	104	50	35	19	1,4844782624	
Nilai MIPA	Rendah	25	0	10	15	0,970950597	
	Sedang	85	0	76	10	0,518569732	**0,920081259**
	Tinggi	86	80	4	1	0,365223317	
Nilai IPS	Rendah	14	13	0	1	0,37123322327	
	Sedang	89	60	4	25	1,099218369	**0,721385036**
	Tinggi	93	7	86	0	0,385285119	
Nilai BB	Rendah	22	5	7	0	0,773226674287 63	
	Sedang	141	75	66	0	0,997059057	**0,500266673**
	Tinggi	33	0	7	26	0,745517843	

Gambar 3. Hasil perhitungan *node* 1

Berdasarkan hasil perhitungan c4.5 diperoleh nilai gain paling tinggi dimiliki atribut nilai MIPA. Kemudian nilai MIPA dijadikan sebagai akar atau node dari pohon keputusan. Proses perhitungan dilanjutkan hingga semua atribut sudah memiliki pohon keputusan. Setelah semua atribut dihitung kemudian akan diperoleh aturan dan diterapkan dalam pengujian data.

5. Pengujian

1) Pengujian Algoritma

Pengujian Algoritma C4.5 yaitu pengujian berdasarkan kebutuhan data mining yang digunakan untuk mengukur performa klasifikasi yaitu mengukur *accuracy*, *precission*, dan *recall dataset* hasil penjurusan menggunakan algoritma C4.5. Perhitungan *accuracy, recall* dan *precission* dapat dirumuskan dalam persamaan-persamaan berikut: (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015)

Precission adalah perhitungan terhadap perkiraan proporsi kasus yang benar dan dirumuskan dalam persamaan 3.

$$Precission = \frac{TP}{TP+TN} \quad (3)$$

Recall adalah perhitungan terhadap perkiraan proporsi kasus positif yang diidentifikasikan benar dan dirumuskan dalam persamaan 4.

$$Recall = \frac{TP}{TP+TN} \quad (4)$$

Akurasi *(Accuracy)* adalah perhitungan terhadap proporsi dari jumlah total prediksi yang benar dan dirumuskan dalam persamaan 5.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

Keterangan:

TP (True Positif) = Jumlah objek positif yang benar diklasifikasikan

TN (True Negatif) = Jumlah objek positif yang salah diklasifikasikan

FP (False Positif) = Jumlah objek negatif yang benar diklasifikasikan

FN (False Negatif) =Jumlah objek negatif yang salah diklasifikasikan

2) Pengujian *black box*

Pengujian sistem yang dilakukan pada penelitian ini yaitu menggunakan *black box testing*. *Black box testing* dilakukan untuk menguji apakah fungsi-fungsi dalam sistem dapat berjalan bagaimana mestinya atau tidak. Pada pengujian juga dapat melihat kesalahan-kesalahan yang terdapat dalam sistem dan selanjutnya dapat diperbaiki.

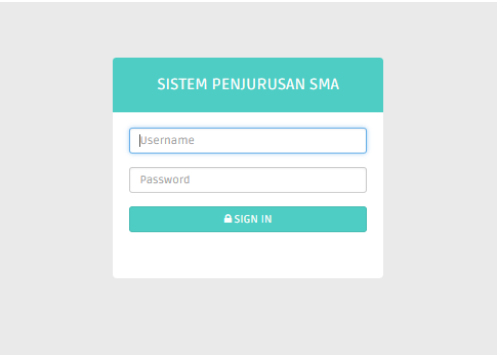
HASIL DAN PEMBAHASAN

A. Implementasi Sistem

Implementasi sistem merupakan tahap yang digunakan untuk mengetahui hasil dari rancangan sistem yang dikembangkan.

1. Halaman *Login*

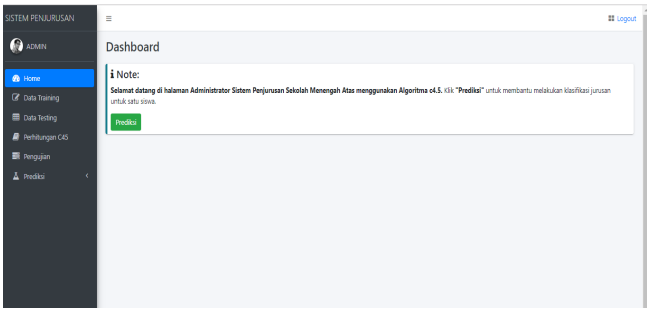
Untuk melakukan proses *login* yaitu pihak *administrator* memasukkan *username* dan *password* agar dapat mengakses menu yang disediakan dalam sistem. Tampilan halaman *login* tersebut ditunjukkan pada Gambar 4.



Gambar 4. Halaman *Login*

2. Halaman Utama

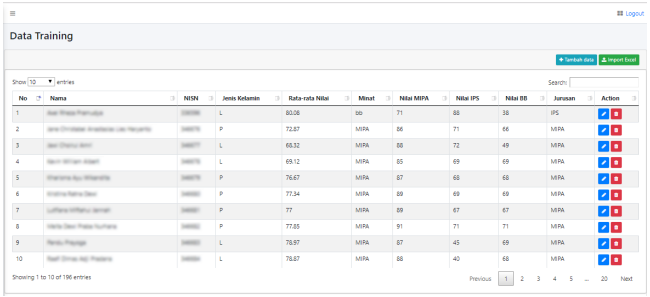
Setelah admin sukses melakukan *login*, maka akan muncul halaman utama sistem sebagai menu utama. Tampilan halaman utama tersebut ditunjukkan pada Gambar 5.



Gambar 5. Halaman Utama

3. Halaman Data *Training*

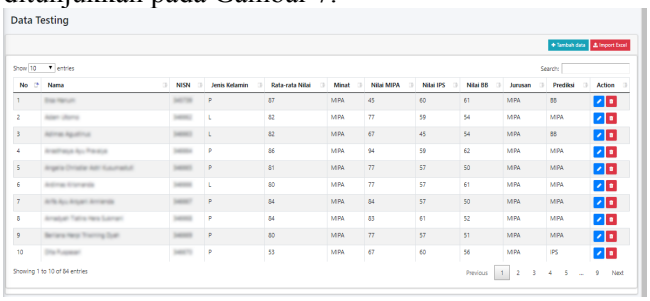
Halaman data *training* berfungsi untuk menampilkan daftar data *training* dari data siswa yang telah diinputkan. Pada halaman ini admin dapat menambahkan data training melalui form yang disediakan atau dapat menambahkan melalui import excel. Data yang ditambahkan sebelumnya dapat diubah maupun dihapus jika admin salah dalam memasukkan data. Tampilan halaman data training ditunjukkan pada Gambar 6.



Gambar 6. Halaman Data *Training*

4. Halaman Data *Testing*

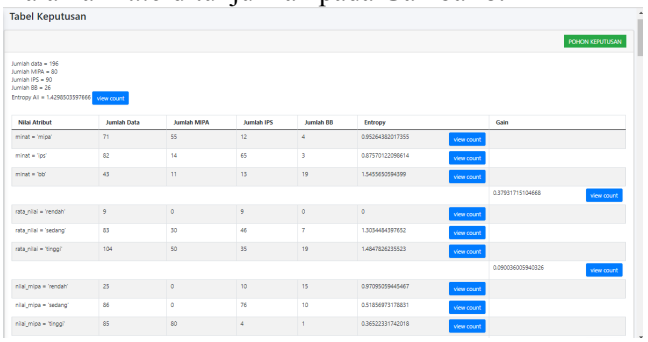
Halaman data *testing* berfungsi untuk menampilkan daftar data *testing* dari data siswa yang telah diinputkan. Pada halaman ini admin dapat menambahkan data *training* melalui *form* yang disediakan atau dapat menambahkan melalui *import excel*. Data yang dimasukkan sebelumnya dapat diubah maupun dihapus jika admin salah dalam memasukkan data. Tampilan halaman data *testing* ditunjukkan pada Gambar 7.



Gambar 7. Halaman Data *Testing*

5. Halaman Perhitungan C4.5

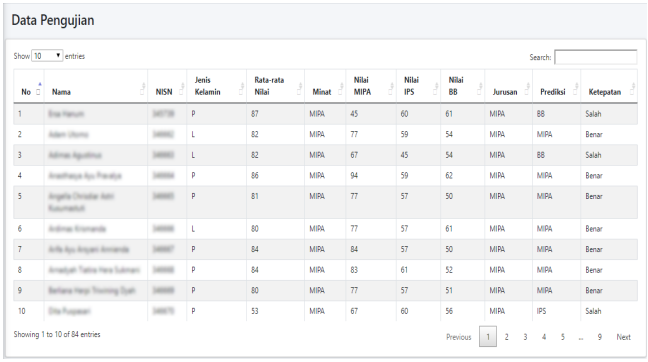
Halaman perhitungan C4.5 berfungsi untuk menampilkan perhitungan algoritma c4.5 dari data training. Pada halaman ini menampilkan jumlah data, nilai *entrophy*, dan nilai *gain* dari masing-masing atribut yang ditunjukkan pada Gambar 5. Selain itu ada tombol pohon keputusan yang mana akan menampilkan halaman baru yang berisi hasil *rule*/aturan yang didapat dari proses perhitungan C4.5. *Rule* ini kemudian akan diterapkan dalam proses pengujian untuk menentukan hasil penjurusan. Halaman *rule* ditunjukkan pada Gambar 8.



Gambar 8. Halaman Perhitungan C4.5

6. Halaman Pengujian

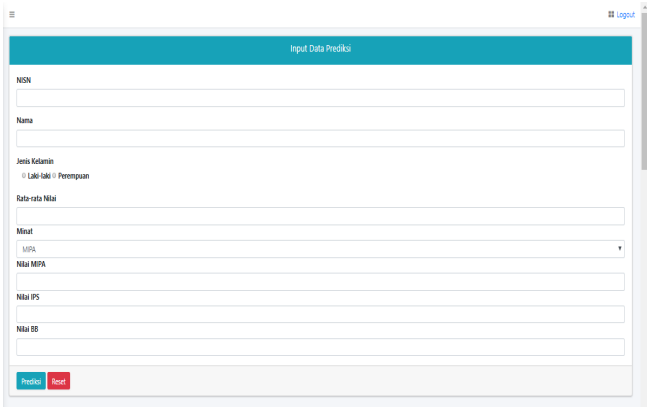
Halaman pengujian berfungsi untuk melihat proses pengujian data dari data testing disertai dengan hasil prediksi dari sistem beserta ketepatannya. Pada kolom ketepatan terdapat dua kondisi yaitu "benar" atau "salah" untuk membandingkan antara kolom jurusan dan hasil prediksi. Halaman pengujian ditunjukkan pada Gambar 9.



Gambar 9. Halaman Pengujian

7. Halaman Prediksi

Halaman prediksi digunakan untuk memprediksi dari satu siwa dan melihat *rule* yang diperoleh dari hasil perhitungan C4.5 berdasarkan data yang sudah dimasukkan oleh admin. Tampilan halaman *input* prediksi untuk penjurusan satu siswa ditunjukkan pada Gambar 10. Sedangkan hasil prediksi jurusan ditunjukkan pada Gambar 11.



Gambar 10. Halaman *input* Prediksi

Gambar 11. Halaman Prediksi

B. Pengujian

Pengujian yang dilakukan pada sistem yang dibangun dalam penelitian ini menggunakan 2 cara yaitu pengujian pertama dengan *precision, recall, accuracy* dan pengujian kedua menggunakan *black box testing*.

1. Pengujian *Precission, Recall*, dan *Accuracy*

Pengujian *precission, recall* dan *accuracy* digunakan untuk mengukur performa klasifikasi. Pengujian yang dilakukan peneliti menggunakan *dataset* yang dibagi menjadi dua yaitu data *training* 70% yang terdiri dari 196 data dan data *testing* 30% yang terdiri dari 84 data dan menghasilkan *precission, recall* dan *accuracy* data seperti pada penjelasan Tabel 2.

Tabel 2. Hasil *Precision*, *Recall* dan *Accuracy*

No	Kriteria	Hasil Pengujian Data *Testing*
1	*Dataset*	84 data
2	*Accuracy*	95%
3	*Recall*	95%
4	*Precission*	100%

2. Pengujian *Black box*

Pengujian *black box* dilakukan untuk melakukan uji coba apakah fungsi-fungsi yang terdapat dalam sistem dapat berjalan bagaimana mestinya atau tidak. Tampilan pengujian *black box* ditunjukkan pada Tabel 3.

Tabel 3. Pengujian *black box*

Menu	*Input*	*Output*	Hasil
Login	*Username dan password benar*	Masuk ke halaman utama	Valid
Home	Masuk dalam menu home	Muncul halaman tampilan home	Valid
Data *Training*	Melakukan *input, edit* dan hapus data	Data *training* ditambah, ditamilkan pada form, diubah dan dihapus	Valid
Data *Testing*	Melakukan *input, edit* dan hapus data	Data *testing* ditambah, ditamilkan pada form, diubah dan dihapus	Valid
Perhitungan C45	Melakukan proses perhitungan C4.5	Menampilkan rule/aturan	Valid
Pengujian	Melakukan pemrosesan data yang telah terisi dari data testing	Menampilkan proses pengujian data yang berisi hasil prediksi dan ketepatan data	Valid
Prediksi	Memasukkan data untuk diprediksi	Data tersimpan di database dan masuk ke halaman prediksi	Valid
Logout	Keluar dari sistem	Muncul ke halaman *login*	Valid

KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa:

1. Algoritma C4.5 dapat digunakan untuk mengklasifikasikan penentuan jurusan siswa SMA.
2. Pengujian *Black Box* yang digunakan untuk memeriksa fungsionalitas dari sistem bernilai valid dan dapat berjalan baik.
3. Pengujian menggunakan Algoritma C4.5. membuktikan bahwa sistem dyang dibuat adalah benar dan sesuai dengan perhitungan manual.
4. Penelitian ini menghasilkan proses penjurusan dengan hasil akurasi sebesar 95%.

Adapun saran yang dapat diajukan yaitu sistem ini dapat dikembangkan menjadi lebih efekti dan efisies serta lebih *detail* termasuk pada kriteria dalam penjurusan dan data yang digunakan lebih banyak agar *rule* yang dihasilkan lebih bervariasi sehingga pada penerapan data uji dapat meminimalisir kesalahan.

DAFTAR PUSTAKA

Kamagi, D. H., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4 . 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa, *VI*(1), 15–20.

Kristanto, O. (2013). Penerapan Algoritma Klasifikasi Data Mining Id3 Untuk Menentukan Penjurusan Siswa Sman 6.

Nasari, F. (2014). Penerapan algoritma c4.5 dalam pemilihan bidang peminatan program studi sistem informasi di stmik potensi utama medan, 30–34.

Novianti, B., Rismawan, T., & Bahri, S. (2016). Implementasi Data Mining Dengan Algoritma C4 . 5 Untuk Penjurusan Siswa (Studi Kasus : Sma Negeri 1 Pontianak), *04*(3).

Nugroho, Y. S. (2014). Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta, (November), 1–6.

Sirait, G., & Hansun, S. (2017). Decision Support System In Choosing Study Program At University Using C4.5 Algorithm (A Case Study At Universitas Multimedia Nusantara), *5*, 357–365.

Swastina, L. (2013). Penerapan Algoritma C4 . 5 Untuk Penentuan Jurusan Mahasiswa, *2*(1).

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction, *1–9*, 55.

Wahyuni, S. (2018). Implementation of Data Mining to Analyze Drug Cases Using C4 . 5 Decision Tree Implementation of Data Mining to Analyze Drug Cases Using C4 . 5 Decision Tree, 0–6.