# Implementasi *Data Mining* Menggunakan Algoritma C4.5 Untuk Klasifikasi Penjurusan Siswa Pada SMA Negeri 2 Pemalang

## IMPLEMENTATION OF DATA MINING USING C4.5 ALGORITHM FOR CLASSIFICATION OF PLACEMENT OF STUDENTS AT SMA STATE 2 PEMALANG

**Prio Pramujio[1], Zaenal Arifin[2]**
Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang
Jl. Nakula I No. 5-11, Kota Semarang, Jawa Tengah 50131,
Telp (024) 3517261/ Fax: 0243520165
e-mail: [1]**112201204589@mhs.dinus.ac.id**, [2]xzaenal@dsn.dinus.ac.id

***Abstrak***

*Penjurusan siswa merupakan suatu prosedur atau proses pengambilan keputusan yang berdasarkan pada minat, pemahaman potensi diri, serta peluang yang tersedia. Penentuan jurusan siswa pada kurikulum 2013 dilakukan pada awal masuk sekolah, yaitu pada kelas X Sekolah Menengah Atas. Implementasi kurikulum 2013 ini ditujukan guna menunjang penyesuaian program pendidikan dengan cirri khas potensi yang terdapat di daerah siswa. Akibat dari penerapan kurikulum 2013 salah satunya pihak sekolah terutama guru BK belum mengetahui bakat, minat, dan karakter siswa dalam mata pelajaran tertentu. Berdasarkan permasalahan tersebut maka dilakukan penerapan data mining menggunakan algoritma C4.5 untuk klasifikasi penjurusan siswa pada SMA Negeri 2 Pemalang. Algoritma C4.5 merupakan salah satu algoritma klasifikasi pohon keputusan yang sering digunakan karena memiliki beberapa kelebihan disbanding algoritma lainnya. Salah satu kelebihan algoritma C4.5 yaitu dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, mempunyai tingkat akurasi yang relative tinggi sehingga dapat diterima, dapat menangani variable bertipe diskret dan numerik, serta efisien dalam menangani variable bertipe diskret. Evaluasi hasil perhitungan algoritma C4.5 dengan 382 data menghasilkan akurasi sebesar 86,1257%.*

***Kata kunci***— Penjurusan, Klasifikasi, Data mining, Pohon keputusan, Algoritma C4.5.

***Abstract***

*Placement of students is a procedure or decision-making process based on interests, understanding their potential, as well as the opportunities available. Determination of student majors in curriculum 2013 performed at the beginning of school, which is in the tenth grade of high school. Implementation of curriculum 2013 is intended to support the adjustment of educational programs with hallmark of the potential contained in the student area. As a result of the implementation of curriculum 2013 for the school, especially BK teachers do not know the talents, interests, and character of students in certain subjects. Based on these problems then the application of data mining using C4.5 algorithm for classification placement of students at SMA Negeri 2 Pemalang. C4.5 algorithm is a decision tree classification algorithm that is often be used because it has several advantages compared to other algorithms. It's advantages are can produce decision tree that is easily interpreted, have a relatively high degree of accuracy so that*

*it can be accepted, can handle variable of type discrete and numerical, and efficient in handling discrete variables of type. Evaluation of the calculation of C4.5 algorithm with 382 data yielded an accuracy of 86.1257%.*

***Keywords***— Students placement, Classification, Data mining, Decision tree, C4.5 Algorithm.

## 1. PENDAHULUAN

Seiring dengan pesatnya perkembangan teknologi dan informasi dewasa ini, mengakibatkan tingkat ketepatan dan akurasi suatu informasi sangat dibutuhkan baik bagi lingkungan bisnis, pendidikan, maupun kehidupan sehari-hari. Tiap-tiap informasi memiliki nilai penting karenaakan berpengaruh pada setiap pengambilan keputusan yang didasarkan pada setiap informasi tersebut. Oleh karena itu diperlukan analisis yang mendalam terhadap suatu data yang akan menghasilkan informasi yang dibutuhkan dalam proses pengambilan keputusan, tak terkecuali dalam proses penjurusan urusan siswa Sekolah Menengah Atas (SMA)[1].

Penentuan jurusan siswa pada kurikulum 2013 dilakukan pada awal masuk sekolah, yaitu pada kelas X Sekolah Menengah Atas. Implementasi kurikulum 2013 ini ditujukan guna menunjang penyesuaian program pendidikan dengan ciri khas potensi yang terdapat di daerah siswa[2]. Akibat dari penerapan kurikulum 2013 salah satunya adalah pihak sekolah terutama guru BK belum mengetahui bakat, minat, dan karakter siswa dalam mata pelajaran tertentu. Hal tersebut dikhawatirkan akan mengakibatkan siswa mengalami kesulitan dalam mengikuti kegiatan pembelajaran yang menyebabkan rendahnya prestasi belajar siswa.

SMA Negeri 2 Pemalang adalah Sekolah Menengah Atas yang telah menerapkan kurikulum 2013 sebagai dasar untuk melakukan proses penjurusan. Penempatan jurusan yang tepat diharapkan meningkatkan minat serta kenyamanan siswa dalam mengikuti kegiatan pembelajaran. Dengan dasar kemampuan yang kurang lebih sama, siswa diharapkan dapat mengikuti proses belajar dengan lancar tanpa adanya gangguan dan tanpa ada yang mengalami kesulitan, serta diharapkan dapat meningkatkan minat dan juga prestasi belajar para siswa.

Penempatan jurusan yang tepat diharapkan meningkatkan minat serta kenyamanan siswa dalam mengikuti kegiatan pembelajaran. Dengan dasar kemampuan yang kurang lebih sama, siswa diharapkan dapat mengikuti proses belajar dengan lancar tanpa adanya gangguan dan tanpa ada yang mengalami kesulitan, serta diharapkan dapat meningkatkan minat dan juga prestasi belajar para siswa.

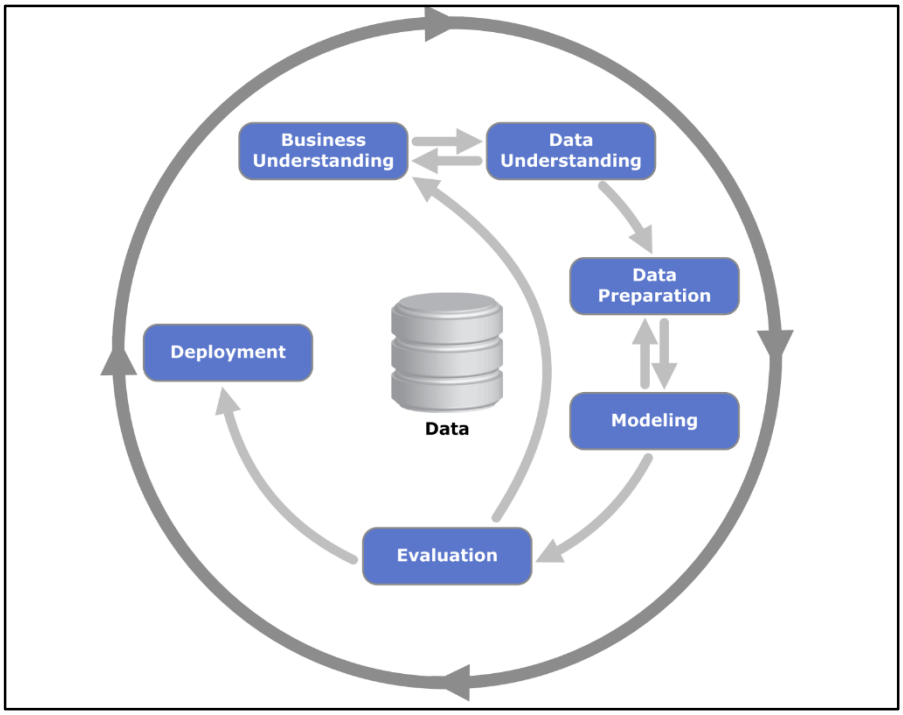
Dari permasalahan yang telah diuraikan diatas, maka penulis akan mengimplementasikan metode *data mining* dengan menggunakan algoritma klasifikasi C4.5 kedalam sebuah sistem yang nantinya akan dijadikan sebagai alat pendukung keputusan dalam klasifikasi penjurusan siswa SMA Negeri 2 Pemalang. Implementasi algoritma C4.5 dalam sistem ini diharapkan dapat membantu proses penjurusan pada SMA negeri 2 Pemalang lebih efektif dan efisien.

Pada jurnal sebelumnya yang sudah dilakukan oleh David Hartanto Kamagi dan Seng Hansun dari Program Studi Teknik Informatika, Universitas Multimedia Nusantara, pada tahun 2014 dengan judul "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa" Dari penelitian tersebut membuktikan bahwa algoritma C4.5 dapat diterapkan untuk memprediksi tingkat kelulusan mahasiswa dengan 4 kategori/kelas yaitu *drop out*, lulus terlambat, lulus tepat, dan lulus cepat. Prediksi kelulusan mahasiswa dengan algoritma C4.5 menghasilkan presentase sebanyak 87.5% dari 60 *data training* dan 40 *data testing*[1]. .Pada jurnal selanjutnya dilakukan oleh Liliana Swastina dari Program Studi Sistem Informasi Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Indonesia pada tahun 2013 dengan judul

"Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa" menyimpulkan bahwa algoritma C4.5 memprediksi lebih akurat dibandingkan dengan algoritma Naive Bayes dalam menentukan kesesuaian jurusan dan rekomendasi jurusan mahasiswa. Karena hal itu, maka dapat disimpulkan bahwa algoritma *decision tree* C4.5 akurat diterapkan untuk menentukan kesesuaian jurusan mahasiswa dari pada algoritma *Naïve Bayes.*Tingkat keakuratan yang dihasilkan algoritma C4.5 sebesar 93,31% serta akurasi rekomendasi jurusan sebesar 82,64%, sedangkan rekomendasi jurusan dengan algoritma Nave Bayes hanya sebesar 66,36%[3].

## 2. METODE PENELITIAN

Dalam penelitian ini, metode *data mining* yang diusulkan berupa model standarisasi *data mining* CRISP-DM. CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah konsorsium perusahaan yang berdiri sejak tahun 1996 oleh Komisi Eropa yang ditetapkan sebagai acuan standar data mining untuk seluruh sector industry. Gambar 2.2 menggambarkan pola siklus hidup dalam CRISP-DM dengan penjelasan *data mining*. Ada enam tahapan siklus hidup dalam data mining (Chapman, 2000) [4]:




**Gambar 2.1 Gambar Siklus Hidup CRISP-DM**

**a. Pemahaman Data (*Data Understanding*)**

Dalam penelitian ini diggunakan data primer yaitu data siswa kelas X SMA Negeri 2 Pemalang tahun ajaran 2015/2016 yang diperoleh melalui *softcopy*secara langsung (observasi). Data yang digunakan adalah data siswa baru atau siswa kelas X tahun ajaran 2015/2016 SMA Negeri 2 Pemalang.

**b. Pengolahan Data (*Data Preparation*)**

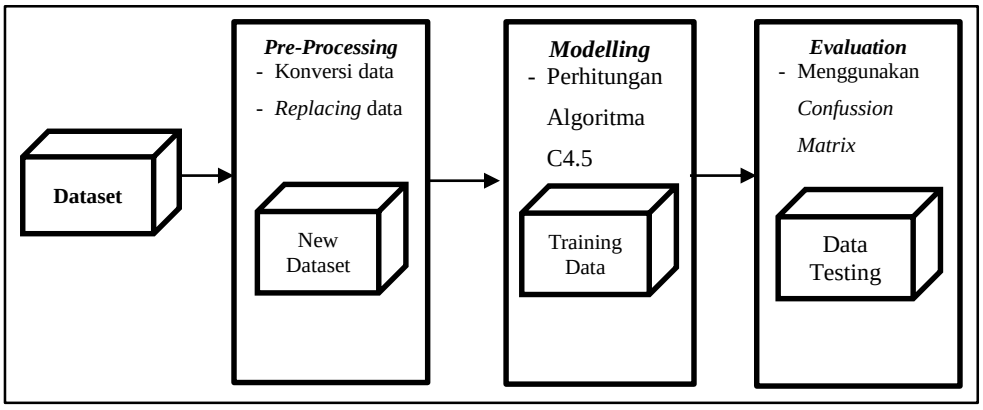
1. Proses pertama yang dilakukan adalah menentuan data yang akan diolah. Dari data yang telah diperoleh, tidak semua data akan diolah karena tidak semua masuk dalam kriteria

penelitian. Penelitian yang akan dilakukan memiliki batasan-batasan data yang akan digunakan. Data awal yang diperoleh terdiri dari 389 *record* data.

2. Proses kedua yang dilakukan adalah mementukan variable atau atribut yang akan digunakan dari proses pertama. Terdapat 16 atribut/variabel pada data awal, atribut/variabel yang akan digunakan sebanyak 8 atribut/variabel adalah nama, n_mtk, un_ipa, rtotal_bhs, rtotal_ips, rtotal_ipa, mnt_1, jurusan.
3. Proses ketiga yang dilakukan adalah penanganan data *missing value*. *Missing value* adalah data yang tidak lengkap dikarenakan attribut tidak tercatat maupun attribut memang tidak dimiliki dsb. Penanganan *missing value* dilakukan dengan penghapusan *record* yang kosong. Jumlah data awal adalah 389 data, terdapat data yang *missing value* sehingga menjadi 382 data *record* yang dapat digunakan.
4. Proses keempat yang dilakukan adalah melakukan konversi data. Data yang telah dipilih kemudian dikonversi gunamempermudah proses penambangan data pada sebagian atribut.Data akan diproses menggunakan alat bantu *data mining*. Konversi dilakukan pada atribut un_mtk, un_ipa, rtotal_bhs, rtotal_ips, rtotal_ipa, dan mnt_1.

**c. Pemodelan (*Modelling*)**

*Decision tree* C4.5 dalam penelitian ini merupakan metode klasifikasi yang digunakan, sedangkan untuk evaluasi serta pengukuran tingkat akurasi menggunakan kerangka kerja *confussion matrix*. Berikut adalah gambar pemodelan *data mining* :




Gambar 2.2 Pemodelan

**d. Evaluasi (*Evaluation*)**

Tahap dimana dilakukan validasi dan pengukuran tingkat akurasi hasil yang dicapai oleh model yang telah ditetapkan. Tahap ini dilakukan menggunakan *framework Confusion Matrix* dengan *tools* Matlab.

**e. Penyebaran (*Deployment*)**

*Output* dari penelitian ini akan menunjukan hasil klasifikasi penjurusan siswa SMA Negeri 2 Pemalang yang sudah diproses menggunakan *software* matlab. Hasil tersebut akan disajikan dalam bentuk sebuah sistem pendukung keputusan yang dapat digunakan oleh pihak sekolah sebagai dasar pertimbangan untuk melakukan penjurusan siswa dengan variabel atau atribut yang telah ditentukan.

## 3. HASIL DAN PEMBAHASAN

a. Persiapan Data

Pada penelitian ini, data yang digunakan adalah data siswa baru SMA 1 Kajen tahun ajaran 2015/2016 dengan jumlah 382 record.

| nama | un_mtk | un_ipa | rtotal_bhs | rtotal_ips | rtotal_ipa | mnt_1 | jurusan |
|---|---|---|---|---|---|---|---|
| AGUNG TRILAKSONO | 1 | 2 | 2 | 1 | 1 | 1 | IPA |
| JULIETIKA PUTRI MELINIA | 3 | 3 | 1 | 2 | 2 | 1 | IPS |
| Ayu Septiana Amaliyah | 2 | 3 | 2 | 2 | 2 | 1 | IPA |
| syahfara ashari putri | 2 | 2 | 1 | 1 | 1 | 1 | IPA |
| Diva Madini | 3 | 2 | 1 | 2 | 2 | 3 | BHS |
| Jantra Wisesa Gati | 3 | 3 | 2 | 2 | 2 | 1 | IPS |
| VIKA OKTAVIA | 3 | 2 | 1 | 2 | 2 | 1 | IPA |
| Rofiatul Adawiyah | 3 | 3 | 2 | 2 | 2 | 2 | IPS |
| PEPARING GUSTI MASHAR ATMAJA | 2 | 2 | 2 | 1 | 1 | 1 | IPS |
| DANANG WAHYU UTOMO | 3 | 2 | 2 | 1 | 2 | 1 | IPS |
| SOLEH TEGUH MARGONO | 2 | 2 | 2 | 2 | 2 | 1 | IPA |
| TAUFIK HIDAYAT | 2 | 1 | 2 | 2 | 2 | 1 | IPA |
| Elsa Rahma Prameswari | 3 | 2 | 2 | 2 | 2 | 1 | IPA |
| ANTY SETIANING TYAS | 2 | 2 | 2 | 1 | 1 | 1 | IPA |
| Maya Kartikaningtias | 2 | 3 | 1 | 2 | 1 | 1 | IPA |
| DIVA RISKY ANANDA | 3 | 2 | 2 | 2 | 2 | 2 | IPS |
| Dias Indah Melisawati | 2 | 2 | 2 | 1 | 1 | 1 | IPA |

Gambar 3.1 Data Baru

b. Perhitungan Algoritma C4.5

1. Perhitungan *Entropy* dan *Gain*

Langkah pertama menghitung nilai *Entropy*, kemudian nilai *Entropy* seluruh nilai fitur per atribut dengan rumus :

$$Entropy(X) = \sum_{j=1}^{k} -pj \times \log 2\, pj$$

Keterangan :
X : Himpunan Kasus
k : jumlah partisi X
pj : Proporsi Xj terhadap X

Kemudian mengitung nilai *Gain* setiap atribut dengan rumus :

$$Gain(X, A) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i)$$

Keterangan :
X : Himpunan Kasus
A : atribut
Xi : Proporsi atribut ke X terhadap jumlah kasus

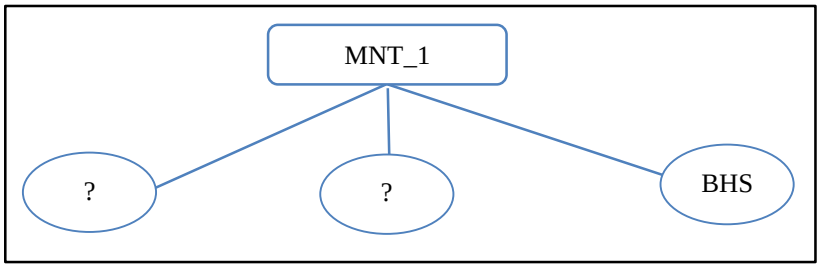
Berikut ini adalah tabel hasil perhitungan *Entropy* dan *Gain* pada node 1 dari 382 data :

**Tabel 3.1** Hasil Perhitungan nilai *Entropy* dan *Gain*

| | **Atribut** | **JML_KASUS** | **IPA** | **IPS** | **BHS** | ***Entropy*** | ***Gain*** |
|---|---|---|---|---|---|---|---|
| Total | | 382 | 220 | 120 | 42 | 1.3331 | |
| UN_MTK | 1 | 5 | 5 | 0 | 0 | 0 | 0.0755 |
| | 2 | 146 | 105 | 28 | 13 | 1.1092 | |
| | 3 | 223 | 110 | 87 | 26 | 1.3941 | |
| | 4 | 8 | 0 | 5 | 3 | 0.9544 | |
| UN_IPA | 1 | 5 | 4 | 1 | 0 | 0.7219 | 0.1364 |
| | 2 | 263 | 180 | 56 | 27 | 1.1365 | |
| | 3 | 113 | 36 | 63 | 14 | 1.3689 | |
| | 4 | 1 | 0 | 0 | 1 | 0 | |
| RTOTAL_BHS | 1 | 180 | 104 | 57 | 19 | 1.3249 | -0.004 |
| | 2 | 202 | 116 | 63 | 23 | 1.348 | |
| | 3 | 0 | 0 | 0 | 0 | 0 | |
| | 4 | 0 | 0 | 0 | 0 | 0 | |
| RTOTAL_IPS | 1 | 131 | 78 | 46 | 7 | 1.2011 | 0.0141 |
| | 2 | 251 | 142 | 74 | 35 | 1.3806 | |
| | 3 | 0 | 0 | 0 | 0 | 0 | |
| | 4 | 0 | 0 | 0 | 0 | 0 | |
| RTOTAL_IPA | 1 | 123 | 82 | 33 | 8 | 1.1555 | 0.0137 |
| | 2 | 259 | 138 | 87 | 34 | 1.3973 | |
| | 3 | 0 | 0 | 0 | 0 | 0 | |
| | 4 | 0 | 0 | 0 | 0 | 0 | |
| MNT_1 | 1 | 286 | 211 | 75 | 0 | 0.83 | 0.6011 |
| | 2 | 55 | 9 | 45 | 1 | 0.7693 | |
| | 3 | 41 | 0 | 0 | 41 | 0 | |

2. Pembentukan Pohon Keputusan

   Pembentukan pohon dimulai dari *root* dengan memilih atribut yang memiliki nilai *Gain* tertinggi. Berikut ini adalah pohon yang terbentuk pada node 1 :

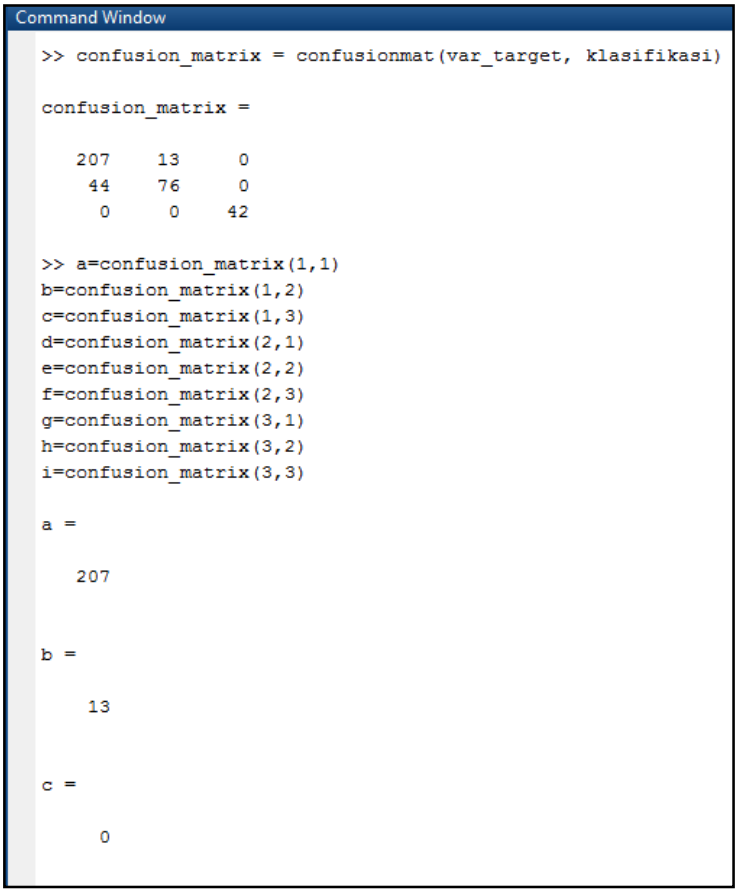



Gambar 3.2 Pohon Keputusan Hasil Perhitungann Node 1

Setelah itu ulangi langkah diatas kembali disetiap cabang sehingga semua kasus memiliki kelas yang sama.

c. Pengujian dengan Matlab

Dari data *training*yang berjumlah 382 data dan 7 variabel antara lain nilai UN Matematika, UN IPA, rata-rata nilai raport Bahasa (Bahasa Indonesia dan Bahasa Inggris), rata-rata nilai raport IPS, rata-rata nilai raport IPA, dan minat, yang dimodelkan dengan algoritma C4.5 diperoleh hasil sebagai berikut :

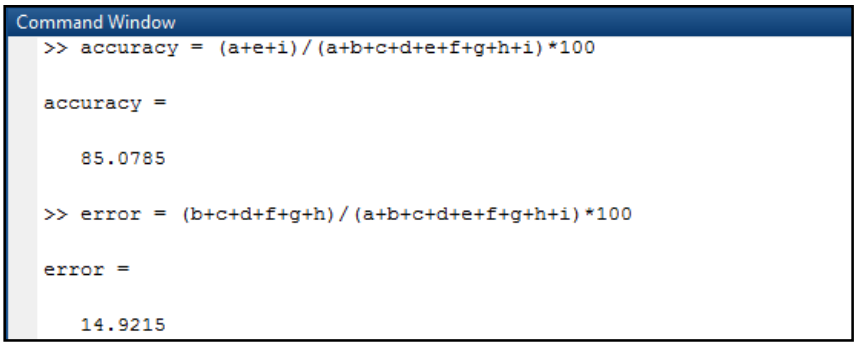



Gambar 3.3 *Confusion Matrix*

Jumlah siswa berdasarkan data asli yang diklasifikasikan IPA yaitu sebanyak 220 *record*, namun hasil dari prediksi untuk kelas IPA diklasifikasikan IPA sebanyak 198 *record*, kelas IPA diklasifikasikan IPS sebanyak 44 *record*, dan kelas IPA diklasifikasikan BHS sebanyak 0 *record*. Berdasarkan data asli yang diklasifikasikan IPS yaitu sebanyak 120 *record*, namun hasil dari prediksi untuk kelas IPS diklasifikasikan IPS seebanyak 76 *record*, kelas IPS diklasifikasikan IPA sebanyak 13 *record*, dan kelas IPS diklasifikasikan BHS sebanyak 0 *record*. Sedangkan untuk kelas yang diklasifikasikan BHS dari data asli berjumlah 42 *record*.namun hasil dari prediksi untuk kelas BHS diklasifikasikan BHS seebanyak 42 *record*, kelas BHS diklasifikasikan IPA sebanyak 0 *record*, dan kelas BHS diklasifikasikan IPS sebanyak 0 *record*.

Berdasarkan *Confusion Matrix*, *accuracy* yang diperoleh dari perhitungan algoritma C4.5 yaitu sebesar 85,0785 %, sedangkan error rate yang dihasilkan sebesar 14,9215 %.

Berikut ini adalah perhitungan *Accuracy* dan *Error* :




Gambar 3.4 Perhitungan *Accuracy* dan *Error*

d. Implementasi Aplikasi

Berikut adalah tampilan aplikasi klasifikasi penjurusan siswa SMA Negeri 2 Pemalang :


Gambar 3.5 Tampilan Aplikasi

## 4. KESIMPULAN

Dari analisis data penjurusan siswa SMA Negeri 2 Pemalang menggunakan algoritma C4.5 berdasarkan literatur yang digunakan maka dapat disimpulkan bahwa teknik *data mining* dengan algoritma C4.5 dapat diimplementasikan untuk mengklasifikasikan jurusan siswa SMA Negeri 2 Pemalang,dengan menggunakan *Confusion Matrix* sebagai evaluasi model.

Hasil menunjukkan bahwa algoritma C4.5 yang diterapkan pada data siswa kelas X SMA Negeri 2 Pemalang tahun ajaran 2015/2016, nilai akurasi *Confusion Matrix* yang didapat sebesar 85,0785 %, sedangkan *error rate*yang dihasilkan sebesar 14,9215 %.

Dengan adanya penerapan *Decision Tree* C4.5 diharapkan berguna bagi pihak sekolah sebagai sistem pendukung keputusankhususnya guru Bimbingan Konseling (BK) dalam membantu menentukan jurusan siswa.

## 5. SARAN

Berikut merupakan hal-hal yang dapat dilakukan untuk melakukan penelitian selanjutnya:

a. Agar menghasilkan klasifikasi yang lebih akurat dapat menggunakan data set yang lebih banyak, misalnya menambahkan data siswa tahun ajaran terbaru.
b. Untuk menemukan tingkat akurasi tertinggi dengan jumlah dataset yang sama dapat menggunakan metode *cross validation* dalam tahap evaluasi dan validasi.
c. Penggunaan metode *data mining* klasifikasi yang lain dapat diterapkan untuk mengembangkan penelitian ini, sehingga nantinya dapat dilakukan perbandingan/komparasi.
d. Penelitian ini setidaknya dapat dijadikan sebagai pedoman oleh pihak SMA Negeri 2 Pemalang dalam pengambilan keputusan penentuan jurusan.

## UCAPAN TERIMA KASIH

Penulis mengucapkan teima kasih kepada Universitas Dian Nuswantoro, Rektor UDINUS, Dekan Fakultas Ilmu Komputer, Kaprodi Sistem Informasi-S1, Dosen pembimbing, Dosen-dosen pengampu kuliah di Fakultas Ilmu Komputer, serta teman-teman dan sahabat yang selama ini telah mendampingi penulis selama kuliah di Universitas Dian Nuswantoro.

## DAFTAR PUSTAKA

[1] D. H. Kamagi and S. Hansun, “Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa,” vol. VI, no. 1, pp. 15–20, 2014.

[2] K. P. dan Kebudayaan, “Permendikbud Nomor 81A Tahun 2013 Tentang Implementasi Kurikulum,” 2013.

[3] L. Swastina, “Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa,” vol. 2, no. 1, 2013.

[4] P. Chapman, *CRISP-DM 1.0: Step-by-step Data Mining Guide, SPSS*. 2000.