**Analisis Kinerja *Data Mining* Algoritma C4.5 Dalam Menentukan Tingkat Minat Siswa yang Mendaftar di Kampus ABC**

**Yudhi Andrian**1**, M. Rhifky Wayahdi**2

1Dosen Teknik Informatika, STMIK Potensi Utama

2Mahasiswa Sistem Informasi, STMIK Potensi Utama

1,2Jl. K.L. Yos Sudarso Km 6,5 No. 3A Tanjung Mulia-Medan

1[yudhi.andrian@gmail.com](mailto:yudhi.andrian@gmail.com), 2[rhifky.wayahdi@yahoo.com](mailto:rhifky.wayahdi@yahoo.com)

*Abstract*

*In determining the level of interest of the students who will enroll in the ABC campus can be predicted by the application of data mining. Data mining is a method of information retrieval (knowledge) which is contained in a very large data. Decision Tree is one of the classification and prediction methods are very powerful in the application of data mining, one of the algorithms in the formation of a decision tree is the C4.5 algorithm. The advantage in this method is effective in analyzing a large number of attributes of the data available and easily understood by the end user. In this study, the author will analyze the performance of the C4.5 data mining algorithm in determining the level of interest of the students who enroll in college ABC, and the obtained results of data mining algorithms can be applied in building C4.5 decision tree (decision tree) are both in the case of interest prediction students enrolled in the ABC campus. Students with C4.5 classification algorithm can classify students' interest to enroll or not enroll in college ABC. Of the 50 tested data there are 40 students enrolled and 10 students do not apply, where the highest level of student interest in the attribute 'year of graduation' highest gain value based on the algorithm classification process C.45.*

*Keyword: Data Mining, Decision Tree, C4.5 Algorithm*

**1. Pendahuluan**

Siswa merupakan suatu substansi yang perlu diperhatikan karena sangat erat kaitannya dengan dunia pendidikan, karena siswa merupakan penerjemah terhadap dinamika ilmu pengetahuan, dan melaksanakan tugas yaitu mendalami ilmu pengetahuan tersebut. Semakin meningkatnya ilmu pengetahuan, semakin meningkat pula pola pikir dan minat seseorang dalam mencapai suatu tujuan.

Salah satu minat seorang siswa adalah melanjutkan ke bangku perkuliahan. Dalam menentukan tingkat minat siswa yang akan mendaftar pada kampus ABC dapat diprediksi dengan penerapan *data mining*. *Data mining* merupakan metode pencarian informasi (pengetahuan) baru yang terkandung dalam data yang sangat besar. *Decision Tree* merupakan salah satu metode klasifikasi dan prediksi yang sangat kuat dan terkenal dalam penerapan *data mining*.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, salah satunya adalah algoritma C4.5. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, di mana atribut paling atas merupakan akar, dan yang paling bawah dinamakan daun. Keuntungan dalam metode ini adalah efektif dalam menganalisis sejumlah besar atribut dari data yang ada dan mudah dipahami oleh pengguna akhir.

Lesmana, I Putu Dody (2012) dalam penelitiannya mengungkapkan tujuan dari *data mining* adalah untuk mendapatkan pola informasi yang tersimpan dalam suatu basis data yang dapat digunakan untuk pengolahan selanjutnya dan sebagai bahan pendukung keputusan [3].

Andriani, Anik (2012) dalam penelitiannya mengatakan penerapan *rule* dari algoritma C4.5 yang digunakan dalam klasifikasi mahasiswa potensi *dropout* terhadap data baru diperoleh hasil evaluasi dan validasi dengan tingkat akurasi sebesar 90% [1].

Yunus, Mahmud, et al. (2014) dalam penelitiannya menerapkan algoritma *data mining* C4.5 untuk melakukan explorasi data guna menemukan hubungan tersembunyi antara sejumlah variabel input dengan variabel target dalam bentuk pohon keputusan *(decision tree)* [6]*.*

Nasari, Fina (2014) dalam penelitiannya menggunakan algoritma C4.5 menjelaskan bahwa hasil uji coba terhadap *rule* dari pohon keputusan pengolahan data *training*, terdapat kecocokan 100 % dengan data *testing* [5]*.*

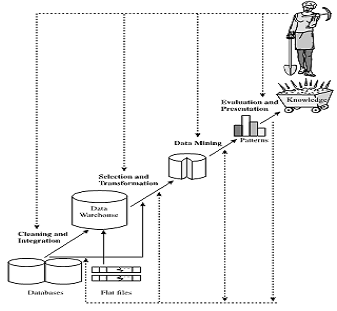
Dari penelitian yang dilakukan Andriani, Anik (2012), Yunus, Mahmud, et al (2014), dan Nasari, Fina (2014) menunjukkan bahwa algoritma C4.5 dapat diterapkan dalam pembentukan pohon keputusan dengan tingkat akurasi yang tinggi. Hal ini yang mendasari penulis untuk menganalisa lebih lanjut kinerja *data mining* algoritma C4.5.

Tujuan penelitian ini dilakukan adalah untuk menganalisis kinerja *data mining* algoritma C4.5 dalam memprediksi tingkat minat siswa yang mendaftar di kampus ABC, penulis ingin mengetahui apakah algortima C4.5 dapat membentuk *decision tree* dengan baik agar dapat diketahui tingkat minat siswa yang mendaftar pada kampus ABC.

**2. *Data Mining***

*Data mining* adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting. Secara teknis, *data mining* dapat disebut sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuat *field* dari sebuah relasional *database* yang besar [4].

*Data mining* didefinisikan sebagai proses menemukan pola dalam data. Proses ini harus otomatis atau biasanya secara semi-otomatis. Pola yang dihasilkan harus berarti bahwa pola tersebut memberikan beberapa keuntungan. Pola tersebut diidentifikasi, divalidasi, dan digunakan untuk membuat sebuah prediksi [1]. Gambar 1 menunjukkan bentuk tahapan dalam *data mining* [6].



Gambar 1. Tahapan dalam *Data Mining*

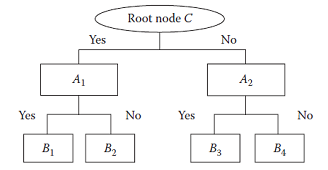
**3. *Decision Tree***

*Decision Tree* (Pohon Keputusan) merupakan metode klasifikasi (taksonomi) yang menggunakan representasi struktur pohon *(tree)* di mana setiap *node* merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari *decision tree* disebut sebagai *root.* Pada *decision tree* terdapat 3 jenis *node,* yaitu [1]:

1. *Root Node,* merupakan *node* paling atas, pada *node* ini tidak ada input dan bisa tidak mempunyai output atau mempunyai output lebih dari satu.
2. *Internal Node,* merupakan *node* percabangan, pada *node* ini hanya terdapat satu input dan mempunyai output minimal dua.
3. *Leaf Node* atau *Terminal Node,* merupakan *node* akhir, pada *node* ini hanya terdapat satu input dan tidak mempunyai output.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, salah satunya ada algoritma C4.5. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record.* Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon.

*Decision tree* tergantung pada aturan *if-then,* tetapi tidak membutuhkan parameter dan metrik. Strukturnya yang sederhana dan dapat ditafsirkan memungkinkan *decision tree* untuk memecahkan masalah atribut *multi-type. Decision tree* juga dapat mengelola nilai-nilai yang hilang atau data *noise.* Gambar 2 menunjukkan contoh struktur *decision tree* [2].



Gambar 2. Contoh Struktur *Decision Tree*

**4. Algoritma C4.5**

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan dibutuhkan algoritma C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3. Algoritma C4.5 dan ID3 diciptakan oleh seorang peneliti di bidang kecerdasan buatan bernama J. Rose Quinlan pada akhir tahun 1970-an. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, di mana atribut paling atas merupakan akar *(root)*, dan yang paling bawah dinamakan daun *(leaf)*.

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam *data mining* adalah [5]:

1. Pilih atribut sebagai simpul akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Pemilihan atribut sebagai simpul, baik akar *(root)* atau simpul internal didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Perhitungan nilai *Gain* digunakan rumus seperti dalam Persamaan 1.

(1)

di mana:

*S* : Himpunan kasus

*A* : Atribut

*n* : Jumlah partisi atribut *A*

|*Si*| : Jumlah kasus pada partisi ke-*i*

|*S*| : Jumlah kasus dalam *S*

Untuk menghitung nilai *Entropy* dapat dilihat pada Persamaan 2.

(2)

di mana:

*S* : Himpunan kasus

*n* : Jumlah partisi *S*

*pi* : Proporsi dari *Si* terhadap *S*

**5. Metode Penelitian**

Tujuan penelitian ini yaitu untuk menganalisis kinerja *data mining* algoritma C4.5 dalam memprediksi tingkat minat siswa yang mendaftar di kampus ABC. Untuk mencapai tujuan tersebut, penulis akan melakukan tahapan-tahapan sebagai berikut:

1. Mengumpulkan data.

Data penelitian yang digunakan adalah data identitas siswa yang mendaftar pada kampus ABC di mana data yang digunakan adalah sebanyak 50 data sebagai data uji. Data yang didapat adalah data yang masih utuh di mana terdapat banyak *field* atau atribut seperti ‘nama, alamat rumah, asal sekolah, jenis kelamin, agama, no. telephone, tahun masuk, program studi, waktu kuliah, ujian, dan lain sebagainya’.

1. Melakukan proses *cleaning* data.

Proses *cleaning* data merupakan proses pembersihan pada data seperti memeriksa kondisi data, menghilangkan redudansi data, memperbaiki kesalahan pada data, dan menghilangkan beberapa data yang tidak lengkap ataupun data yang tidak diperlukan. Adapun data yang telah dilakukan proses *cleaning* yaitu menyisakan beberapa atribut penting seperti ‘alamat rumah, asal sekolah, jenis kelamin, dan tahun masuk’.

1. Melakukan proses *transformation* data.

Proses *transformation* data merupakan pengelompokkan atau penyederhanaan data. Variabel penelitian meliputi variabel input dan variabel output atau target. Variabel input dinyatakan oleh beberapa atribut yang ditransformasi seperti alamat rumah ditransformasi menjadi ‘jauh, sedang, dan dekat’. Sedangkan asal sekolah ditransformasi menjadi ‘SMA Negeri, SMA Swasta, SMK Negeri, SMK Swasta, MA Negeri, dan MA Swasta’. Variabel ‘alamat rumah’ dan variabel ‘asal sekolah’ dilakukan transformasi data karena jangkauan nilainya yang lebar, sehingga dapat mengakibatkan proses pengenalan pola dan dan pembentukan pohon keputusan menjadi lama. Variabel output atau targetnya adalah atribut ‘status’ yang berisikan nilai ‘daftar’ atau ‘tidak daftar’.

1. Melakukan proses perhitungan algoritma C4.5.

Setelah data dikumpulkan, dilakukan proses *cleaning*, dan dilakukan *transformation* data, selanjutnya data akan diproses menggunakan perhitungan algoritma C4.5 untuk menghasilkan pohon keputusan *(decision tree)*.

1. Pembentukan pohon keputusan *(decision tree).*

Pohon keputusan *(decision tree)* terbentuk setelah dilakukan proses perhitungan algoritma C4.5 dengan memilih atribut yang dijadikan sebagai akar *(root)* didasarkan pada nilai *Gain* tertinggi dari beberapa atribut yang digunakan.

**6. Hasil dan Analisa**

Pada tahap pengujian, data yang sudah ditransformasikan ke dalam bentuk kelompok data atribut yang lebih sederhana. Data yang digunakan adalah data siswa yang mendaftar dan tidak mendaftar pada kampus ABC yang diambil secara acak/random sebanyak 50 data sampel. Data akan diolah dengan menerapkan *data mining* algoritma C4.5 pada proses memprediksi minat siswa untuk mendaftar ke kampus ABC. Beberapa atribut atau komponen variabel yang digunakan yaitu terlihat pada Tabel 1 **(Lampiran)**.

Kemudian hitung *Entropy* dengan rumus sebagai berikut:

Jadi,

Tabel 2. Hasil Perhitungan pada *Dataset*

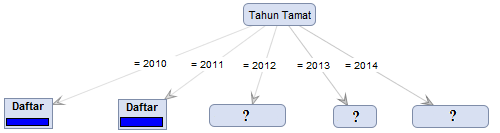
|  |  |  |  |
| --- | --- | --- | --- |
| **Total Kasus** | **Sum(Daftar)** | **Sum(Tidak Daftar)** | ***Entropy*** |
| 50 | 40 | 10 | 0.721928095 |

Setelah mendapatkan *Entropy* dari keseluruhan kasus seperti terlihat pada Tabel 2, lakukan analisis pada setiap atribut dan nilai-nilainya. Kemudian hitung *Entropy*-nya dan hitung *Gain* pada setiap atribut.

Untuk menghitung *Gain* setiap atribut menggunakan rumus:

Dan akan didapatkan hasil seperti terlihat pada Tabel 3 **(Lampiran)**.

Pada Tabel 3 terlihat nilai *Gain* terbesar adalah *Gain* ‘tahun tamat’. Maka atribut ‘tahun tamat’ menjadi *root node* atau *node* akar. Kemudian pada atribut ‘tahun tamat’ 2010 dan 2011 masing-masing memiliki 4 dan 5 kasus di mana semuanya memiliki jawaban ‘daftar’. Dengan demikian ‘tahun tamat’ 2010 dan 2011 menjadi daun atau *leaf.* Maka akan terbentuk pohon keputusan *node* 1 seperti terlihat pada Gambar 3.



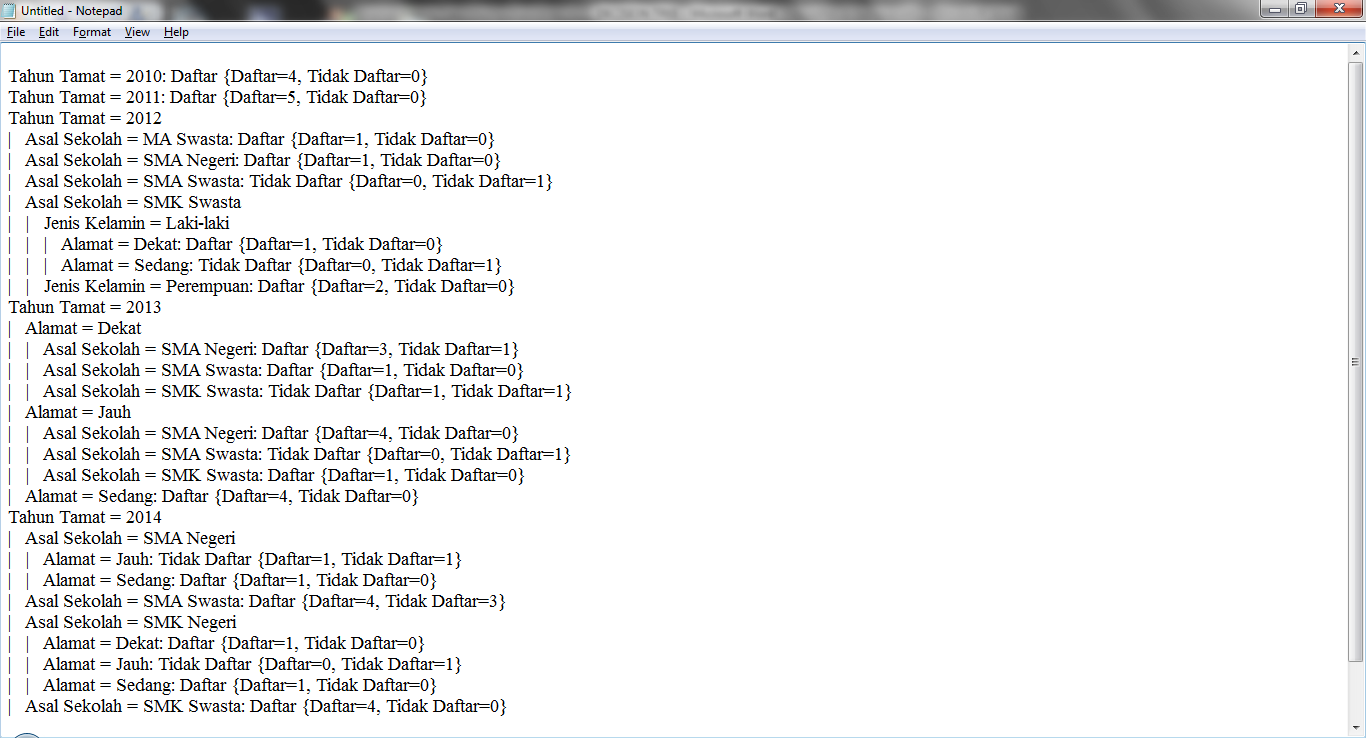
Gambar 3. Pohon Keputusan *node* 1 *(root node)*

Berdasarkan Gambar 3 dapat dilihat siswa yang mendaftar ada pada tahun tamat 2010 dan 2011. Berdasarkan pembentukan pohon keputusan *node* 1 *(root node),* *node-node* atribut lainnya akan dianalisis lebih lanjut. Dengan memulai menganalisis pada atribut ‘tahun tamat’ 2012, 2013, dan 2014 dengan melakukan proses yang sama seperti sebelumnya dengan mencari nilai *Entropy* dan *Gain-*nya.

Setelah dilakukan perhitungan dan pengujian data pada masing-masing atribut dengan algoritma C4.5, maka didapatkan pola pohon keputusan akhir seperti ditunjukkan pada Gambar 4 **(Lampiran)**.

Pada Gambar 4 dapat dilihat bahwa tingkat dominan minat siswa untuk mendaftar adalah berdasarkan ‘tahun tamat’, yaitu ‘tahun tamat’ 2010 dan 2011. Sedangkan tingkat minat siswa untuk mendaftar pada *node* terakhir adalah berdasarkan ‘alamat rumah’ sebagai pertimbangan akhir. Apabila jarak ‘alamat rumah’ dekat dan sedang dari kampus sebagian besar mendaftar, yaitu masing-masing sebanyak 14 orang. Sedangkan apabila jarak alamat jauh dari kampus yang mendaftar sebanyak 12 orang siswa.

Gambar 5 menunjukkan deskripsi lengkap dari pohon keputusan *(decision tree)* yang terbentuk dari 50 data siswa untuk menentukan minat siswa yang mendaftar pada kampus ABC.



Gambar 5. Deskripsi *Decision Tree*

Gambar 5 menunjukkan hasil deskripsi secara lengkap dari pohon keputusan *(decision tree)* yang telah terbentuk dengan menggunakan algoritma C4.5. Dari hasil dekripsi pada Gambar 5 juga menunjukkan bahwa penggunaan *data mining* algoritma C4.5 baik digunakan dalam proses menggali data *(data mining process)* untuk menarik beberapa kesimpulan yang divisualisasikan dengan pohon keputusan *(decision tree).*

**7. Kesimpulan**

Dari hasil penelitian dapat diambil beberapa kesimpulan antara lain:

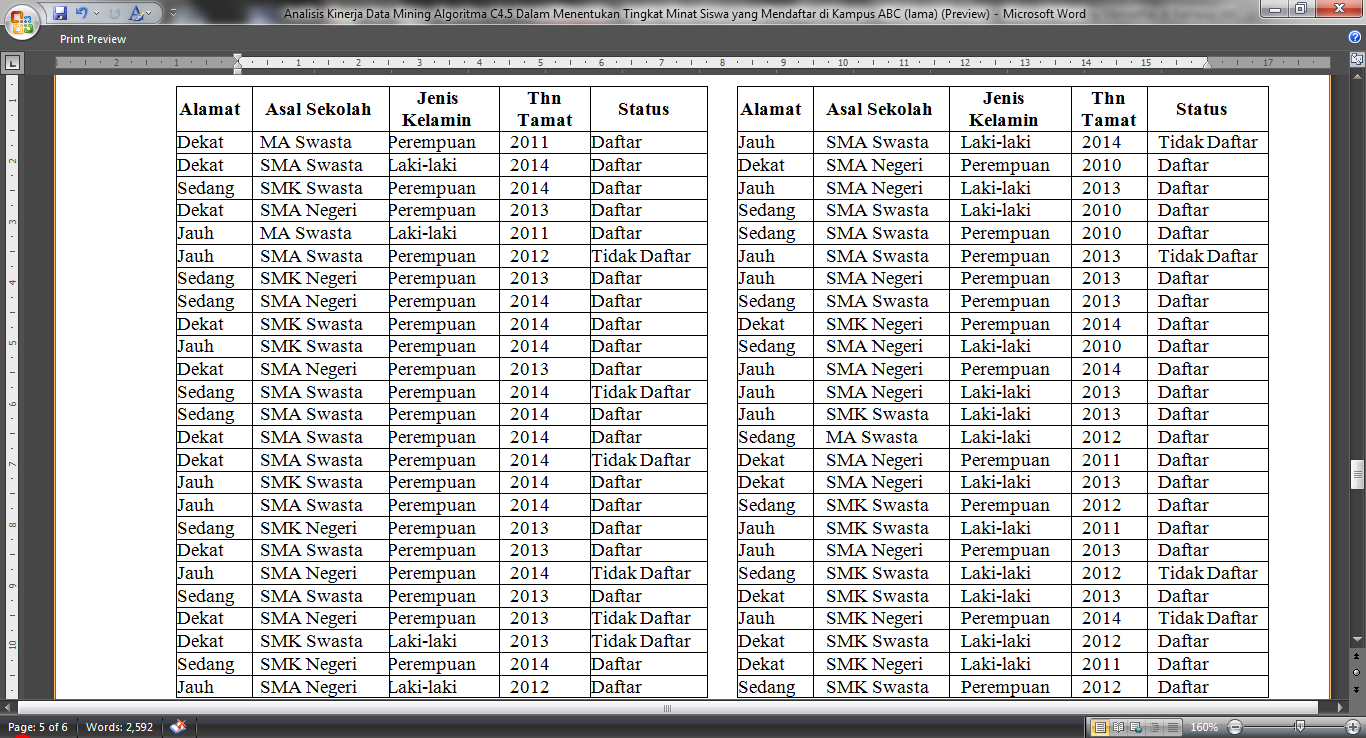
1. *Data mining* algoritma C4.5 dapat diterapkan dalam membangun pohon keputusan *(decision tree)* yang baik pada kasus prediksi minat siswa yang mendaftar pada kampus ABC.
2. Klasifikasi siswa dengan algoritma C4.5 dapat mengklasifikasikan minat siswa untuk mendaftar atau tidak mendaftar pada kampus ABC.
3. Dengan adanya penerapan *data mining* algoritma C4.5 diharapkan mampu memberikan solusi dalam memprediksi tingkat minat siswa yang mendaftar pada kampus ABC.
4. Dari 50 data yang diuji terdapat 40 siswa yang mendaftar dan 10 siswa tidak mendaftar, di mana tingkat minat siswa mendaftar tertinggi pada atribut ‘tahun tamat’ berdasarkan nilai *Gain* tertinggi pada proses klasifikasi dengan algoritma C.45.

**Daftar Pustaka:**

|  |  |
| --- | --- |
| [1] | Andriani, Anik (2012). “Penerapan Algoritma C4.5 pada Program Klasifikasi Mahasiswa *Droupout*”. Prosiding Seminar Nasional Matematika. |
| [2] | Dua, S. and Xian Du (2011). “*Data Mining and Machine Learning in Cybersecurity”*. USA: Taylor & Francis Group. ISBN-13: 987-1-4398-3943-0. USA. |

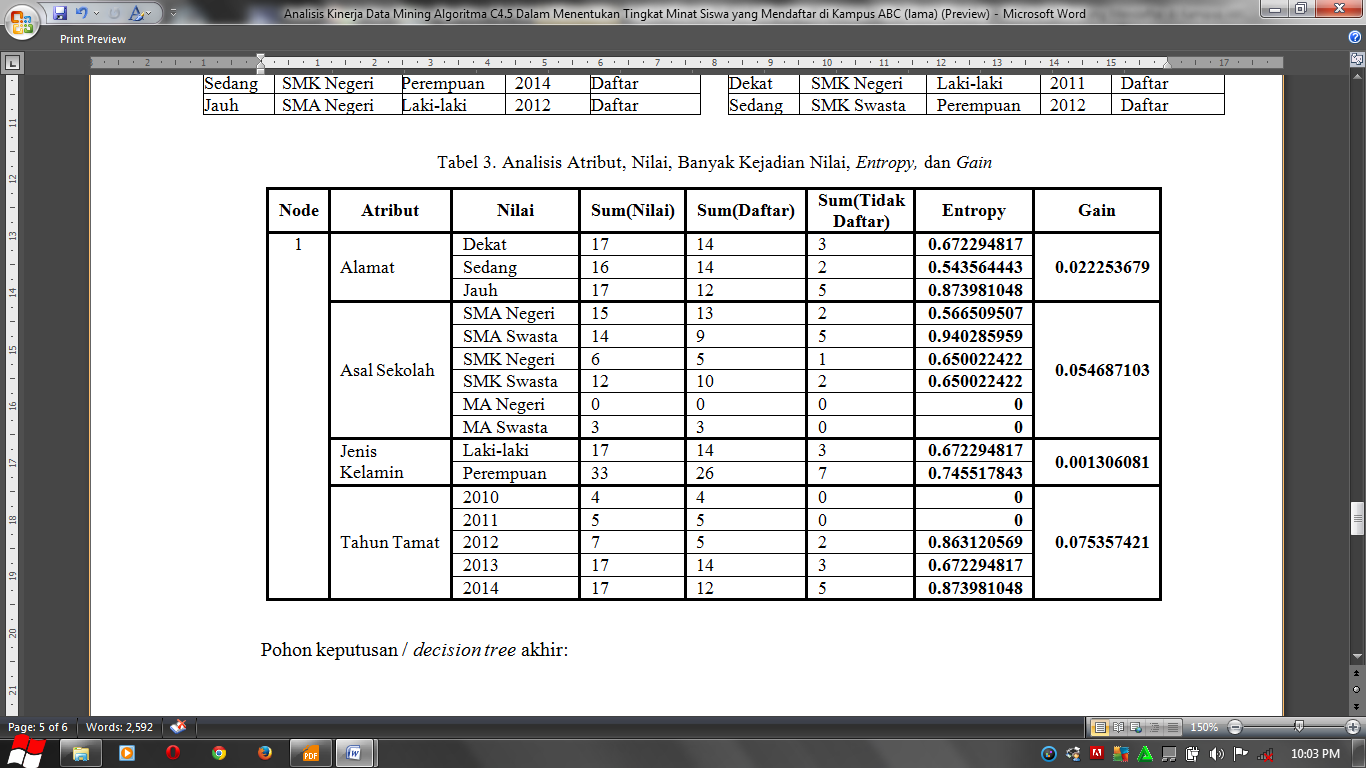
**Lampiran:**

Tabel 1. *Learning Dataset*

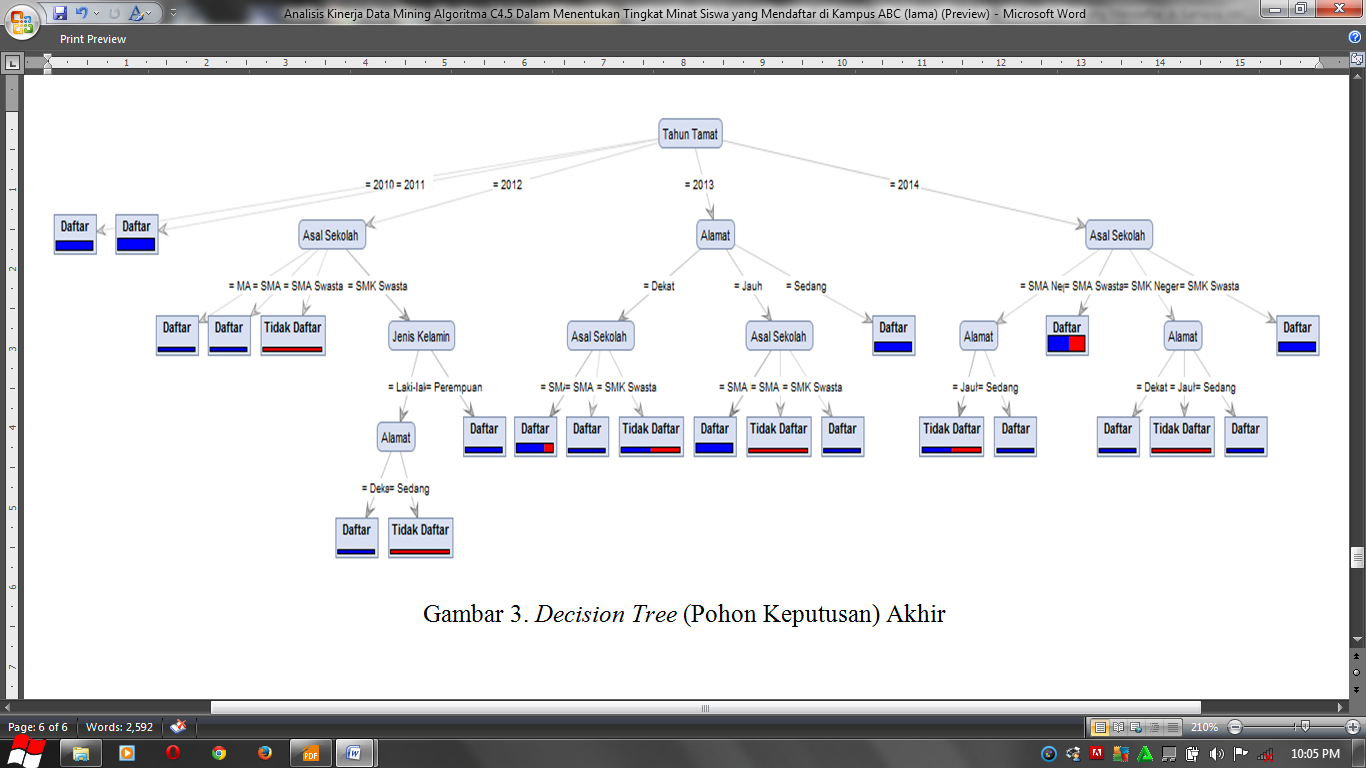


|  |  |
| --- | --- |
| [3] | Lesmana, I Putu Dody (2012). “Perbandingan Kinerja *Decision Tree* J48 dan ID3 dalam Pengklasifikasian Diagnosis Penyakit Diabetes Mellitus”. Jurnal Teknologi dan Informatika. Vol. 2, No. 2. |
| [4] | Mabrur, Angga Ginanjar dan Riani Lubis (2012). “Penerapan *Data Mining* untuk Memprediksi Kriteria Nasabah Kredit”. Jurnal Komputer dan Informatika. Edisi 1, Vol. 1. |
| [5] | Nasari, Fina (2014). “Analisa Faktor Penyebab Tingginya Pemakaian Listrik Rumah Tangga Menggunakan Algoritma C4.5”. UPI YPTK. Padang. |
| [6] | Yunus, Mahmud, et al (2014). “SPK Pemilihan Calon Pendonor Darah Potensial dengan Algoritma C4.5 dan Fuzzy Tahani”. Jurnal EECCIS Vol. 8, No. 1. |

Tabel 3. Analisis Atribut, Nilai, Banyak Kejadian Nilai, *Entropy,* dan *Gain*



Pohon keputusan *(decision tree)* akhir:



Gambar 4. *Decision Tree* (Pohon Keputusan) Akhir