JURNAL ILMU KOMPUTER DAN TEKNOLOGI
Vol. 2, No. 2, Nov 2021, pp. 1-5

ISSN: 2746-4237
IKOMTI - 1

# Implementasi Teknik Data Mining untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma C4.5

**Novitaria Manullang [1✉], Rahmat Widia Sembiring [2], Indra Gunawan [3], Iin Parlina [4], Irawan [5]**

[1) 3) 4) 5)] *Teknik Informatika, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia*
[1] novitariamanullang8@gmail.com
[3] indra@amiktunasbangsa.ac.id
[4] iin@amiktunasbangsa.ac.id
[5] irawaniwan56@amiktunasbangsa.ac.id

[2)] *Teknik Informatika, AMIK Tunas Bangsa, Pematangsiantar, Indonesia*
rahmatwsphd@gmail.com

***Abstract*— In an effort to provide quality education in a school institution, student interest is needed in selecting the best majors in school. The goal of this student's need for specialization in choosing a major is very important in supporting science to higher education and the enthusiasm of these students in developing knowledge in the field of the department they choose, such as problems in some Vocational High Schools where these students choose majors according to the encouragement of others without any interest in majors, resulting in a decrease in enthusiasm for learning. Data mining is a way to transform data into useful information and can produce new knowledge. The method used is the C4.5 Algorithm in determining the majors to be taken by students according to their own backgrounds, interests, and abilities. The variables used are student majors, student interest and talent test results. By processing data using the RapidMiner application, it was found that precision value of 100% and a recall value of 100%.**

***Keywords*— *Selection of majors, Data Mining, Algorithm C4.5***

***Intisari*— Dalam usaha untuk menyelenggarakan pendidikan yang berkualitas pada suatu instansi sekolah maka diperlukan prediksi peminatan siswa dalam pemilihan jurusan yang terbaik di bangku sekolah. Tujuan peminatan siswa ini dalam pemilihan jurusan sangat penting dalam menunjang ilmu ke perguruan tinggi dan lebih semangatnya siswa tersebut dalam mengembangkan ilmu di bidang jurusan yang dipilihnya, seperti masalah di beberapa Sekolah Menengah Kejuruan dimana siswa tersebut memilih jurusan sesuai dengan dorongan orang lain tanpa ada minat pada jurusan tersebut sehingga mengakibatkan turunnya semangat dalam belajar. Data mining merupakan salah satu cara untuk mengubah data menjadi informasi yang berguna dan dapat menghasilkan ilmu baru. Metode yang digunakan adalah Algoritma C4.5 dalam menentukan jurusan yang akan diambil oleh siswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Variabel yang digunakan adalah jurusan siswa, hasil tes minat dan bakat siswa. Dengan pengolahan data menggunakan aplikasi *RapidMiner* didapat bahwa nilai *precision* sebesar 100% dan nilai *recall* sebesar 100%.**

***Kata kunci* — *Pemilihan jurusan, Data Mining, Algoritma C4.5***

## I. PENDAHULUAN

Penerimaan siswa baru pada suatu institusi pendidikan merupakan sebuah kegiatan yang selalu dilaksanakan setiap tahun ajaran baru, dimana data calon siswa baru tersebut selalu meningkat dari tahun ke tahun (Muwardah dan Pramunendar, 2015) [1]. Melalui jenjang Sekolah Menengah Kejuruan (SMK), setiap siswa diberikan pilihan jurusan untuk konsentrasi pembelajaran kedepannya selama bersekolah.

SMK Swasta Persiapan adalah salah satu SMK di Kota Pematangsiantar yang beralamat di Jalan Pane No. 66 Kelurahan Tomuan Kecamatan Siantar Timur Kota Pematangsiantar. SMK Swasta Persiapan Pematangsiantar menerima siswa baru mulai tahun 1968 dengan menawarkan satu jurusan yaitu Teknik Mesin dan beberapa tahun kemudian karena peminat bertambah terus maka membuka kompetensi keahlian hingga sekarang terdiri dari 7 (tujuh) kompetensi keahlian.

Melihat masalah yang dihadapi Sekolah dalam menentukan jurusan disetiap siswa, maka perlu diterapkan suatu metode untuk menyelesaikan masalah tersebut. Salah satunya adalah dengan melakukan prediksi terhadap peminatan jurusan siswa, metode yang cocok dalam melakukan prediksi peminatan jurusan yang diminati adalah penerapan data mining dengan menggunakan algoritma. Dalam penelitian ini algoritma yang digunakan adalah Algoritma C4.5.

Algoritma ini merupakan algoritma yang popular digunakan dan memiliki tingkatan akurasi yang lebih tinggi. Algoritma C4.5 merupakan pengembangan dari Algoritma ID3, ID3 sendiri dikembangkan oleh J.Ross Quinlan. Dalam prosedur algoritma ID3, inputannya berupa sampel training, label training dan atribut (Marwana, 2017) [2].

## II. METODOLOGI PENELITIAN

### *A. Data Mining*

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Swastina, 2013) [3]. Data mining disisi lain adalah kegiatan meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan [4].

### *B. Algoritma C4.5*

Algoritma C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk mem- prediksi nilai atribut bertipe diskret dari record yang baru. Algortima C4.5 sendiri merupakan pengembangan dari algortima ID3, dimana pengembangan dilakukan dalam hal, bisa mengatasi missing data, bisa mengatasi data continue dan pruning [5].

Secara umum Algorima C4.5 untuk membangun pohon keputusan adalah sebagai berikut [6]:

1. Pilih atribut sebagai node 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.

Untuk memilih atribut sebagai node akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Untuk menghitung Gain digunakan rumus seperti tertera dalam persamaan berikut:

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

Keterangan:
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah S dalam S

Setelah mendapatkan nilai Gain, ada satu hal lagi yang perlu kita lakukan perhitungan, yaitu mencari nilai Entropy [7]. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan keluaran atribut. Rumus dasar dari Entropy tersebut adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} -pi\ .\ log2\ pi \quad (2)$$

Ketarangan :
S : Himpunan kasus
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

### *C. RapidMiner*

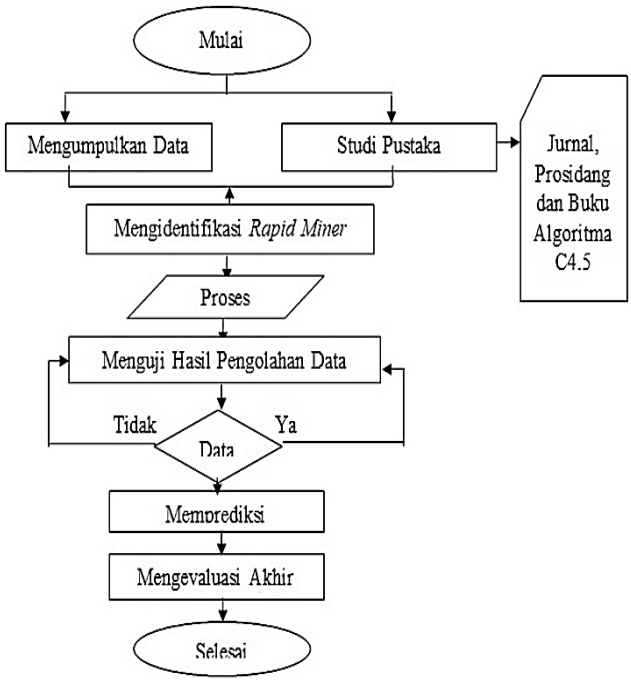
*RapidMiner* merupakan perangakat lunak yang bersifat terbuka (open source). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. *RapidMiner* ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi [8].

### *D.* Lokasi dan Waktu Penelitian

Adapun yang menjadi tempat penelitian adalah SMK Swasta Persiapan Jalan Pane No. 66 Kel. Tomuan Kecamatan Siantar Timur Kota Pematangsiantar. Kegiatan penelitian ini dilakukan sejak tanggal 10 November 2020.

### *E.* Rancangan Penelitan

Perancangan penelitian ini digunakan untuk menguraikan dan menyelesaikan masalah dalam penelitian yang dapat dilihat pada Gambar 1 berikut:




*Gambar 1. Rancangan Penelitian*

Beberapa perancangan penelitian pada gambar 1 diatas, maka masing-masing langkah dapat diuraikan sebagai berikut:

1) Mengumpulkan Data
Pada tahap ini data diperoleh dari SMK Swasta Persiapan Pematangsiantar.

2) Studi Pustaka
Tahap ini merupakan langkah untuk melengkapi pengetahuan dasar dan teori-teori yang digunakan dalam penelitian.

*3)* Mengidentifikasi Masalah
Tahap ini merupakan langkah untuk memproses tahap konservasi data yang diperoleh sesuai dengan bobot yang sudah ditentukan.

*4)* Proses
Tahap ini bertujuan untuk mempermudah permahaman terhadap isi record.

5) Menguji Hasil Pengolahan Data
Pada tahapan ini melakukan uji coba terhadap hasil pengolahan data dengan menggunakan software Rapid Miner.

*6)* Memprediksi
Prediksi dilakukan untuk melihat data akurat dengan Algoritma C4.5.

*7)* Mengevaluasi Akhir
Mengevaluasi akhir dilakukan untuk mengetahui apakah testing hasil pengolahan data sesuai dengan yang diharapkan.

## III. HASIL DAN PEMBAHASAN

Dalam penyelesaian masalah prediksi peminatan jurusan pada kompetensi keahlian di SMK Swasta Persiapan Pematangsiantar, ada beberapa tahapan yang harus dilakukan, antara lain sebagai berikut :

- Menyiapkan *data training*;
- Menetukan atribut dari data yang diperoleh;
- Melakukan perhitungan nilai *entropy* dan *gain*;
- Melakukan prediksi dengan metode Algoritma C4.5 menggunakan aplikasi *RapidMiner*.

### *A.* Pengolahan Data

- Kriteria Data

Kriteria data yang digunakan dapat dilihat pada Tabel 1 berikut ini:

*Tabel 1. Kriteria Data*

| No. | Atribut | Ket. |
|---|---|---|
| 1. | Bahasa Indonesia | Digunakan |
| 2. | Matematika | Digunakan |
| 3. | Bahasa Inggris | Digunakan |
| 4. | Teknik Instalasi Tenaga Listrik (TITL) | Digunakan |
| 5. | Teknik Audio Video (TAV) | Digunakan |
| 6. | Teknik Pemesinan (TP) | Digunakan |
| 7. | Multimedia (MM) | Digunakan |
| 8. | Teknik Komputer dan Jaringan (TKJ) | Digunakan |
| 9. | Teknik Kendaraan Ringan Otomotif (TKRO) | Digunakan |
| 10. | Teknik dan Bisnis Sepeda Motor (TBSM) | Digunakan |

- *Split Validation*

*Split Validation* adalah teknik validasi yang membagi data menjadi dua bagian secara acak, sebagian sebagai data training dan sebagian sebagai *data testing*. Dengan menggunakan *split validation* klasifikasi diperoleh dengan menggunakan teknik *systematic random sampling*, yaitu dengan membagi ukuran populasi dengan ukuran sampel yang dikehendaki. Penentuan unsur selanjutnya ditempuh dengan cara menggunakan *interval* sampel [9].

Pada pembagian data dibagi menjadi data training dan data testing, hasil pembagian tersebut didapat jumlah *data training* sebanyak 157 *record* data dan jumlah *data testing* sebanyak 100 *record* data.

### *B.* Hasil Percobaan

Data Uji adalah data yang telah dipersiapkan sebelumnya untuk dilakukan pengujian. Dari hasil data uji yang ada, kemudian dilakukan pengkategorian dengan variable dan atribut yang kemudian dijadikan *data training* sebanyak 157 *record data* dan *data testing* sebanyak 100 *record data*. Dari proses tersebut kemudian di hitung dengan Algoritma C4.5 untuk mengetahui prediksi peminatan jurusan siswa di SMK Swasta Persiapan Pematangsiantar.

Dalam penelitian prediksi jurusan siswa menggunakan Algoritma C4.5, pohon keputusan dibuat berdasarkan dari hasil perhitungan *Entropy* dan *Gain*, setelah pohon keputusan tersebut terbentuk, langkah selanjutnya adalah mencari rule berdasarkan cabang pohon keputusan. Berikut ini akan dibahas tentang langkah-langkah perhitungan manual dan analisa yang digunakan dengan mengguanakan *tools Rapid Miner*. Berikut ini adalah Tabel 2, merupakan hasil perhitungan menggunakan metode Algoritma C4.5.

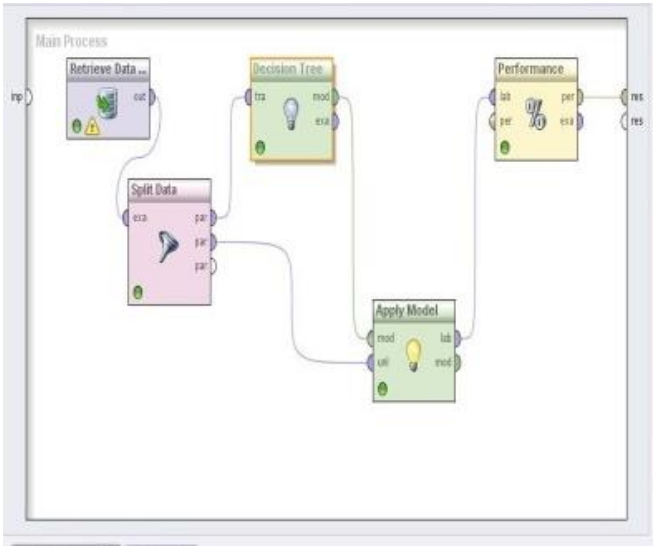
*Tabel 2. Perhitungan C4.5*

| C.45 | | | | | | |
|---|---|---|---|---|---|---|
| **Langkah** | | **Jlh. Kasus** | **Yes** | **No** | **Entropy** | **Information Gain** |
| Total | | 100 | 44 | 56 | 0,98958752 | |
| Jurusan | | | | | | 0,293091345 |
| | TITL | 11 | 11 | 0 | 0 | |
| | TAV | 12 | 3 | 9 | 0,81127812 | |
| | TP | 16 | 1 | 15 | 0,33729006 | |
| | MM | 14 | 6 | 8 | 0,98522813 | |
| | TKJ | 12 | 6 | 6 | 1 | |
| | TKRO | 17 | 4 | 13 | 0,78712658 | |
| | TBSM | 18 | 13 | 5 | 0,85240517 | |
| Bhs Indonesia | | | | | | 0 |
| | A | 44 | 44 | 0 | 0 | |
| | B | 21 | 21 | 0 | 0 | |
| | C | 35 | 35 | 0 | 0 | |
| Matematika | | | | | | 0,015304757 |
| | A | 27 | 9 | 18 | 0,91829583 | |
| | B | 21 | 9 | 12 | 0,98522813 | |
| | C | 52 | 25 | 27 | 0,99893265 | |
| Bhs Inggris | | | | | | 0,015232905 |
| | A | 18 | 6 | 12 | 0,91829583 | |
| | B | 27 | 13 | 14 | 0,99901027 | |
| | C | 55 | 23 | 32 | 0,98059744 | |

Pada Tabel 2 dijelaskan bahwa dari hasil data siswa atau data testing di hitung dengan algoritma C4.5 dan dikelompokan perkelas sesuai atribut yang menghasilkan *information gain* tertinggi, yaitu jatuh pada jurusan siswa. Jadi dapat diambil kesimpulan bahwa nilai information gain tertinggi untuk dijadikan pohon keputusan adalah jurusan siswa jurusan dengan jurusan yang paling diminati adalah jurusan Teknik dan Bisnis Sepeda Motor.

*C.* Proses Pengujian Data Menggunakan *RapidMiner*

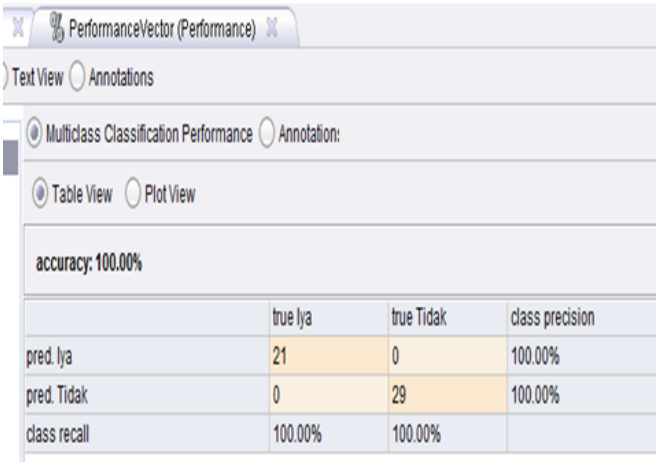
Setelah data dianalisis dan diklasifikasikan menggunakan metode Algoritma C4.5. Maka untuk tahap selanjutnya adalah dengan pembuktian dari analisis perhitungan manual tersebut. Adapun aplikasi yang digunakan dalam pengujian klasifikasi jurusan siswa ini adalah menggunakan aplikasi *RapidMiner* seperti ditunjukkan pada Gambar 2 berikut:




*Gambar 2. Penghubungan port decision tree, apply model dan performance*

Pertama dilakukan *reatrive data* atau *input data*, kemudian dilakukan *split data*, selanjutnya tahapan pada gambar diatas adalah menghubungkan port-port dari operator *decision tree*, *operator apply model* dan *operator performance*, lalu klik *icon run* pada *toolbar* untuk menampilkan hasil.

Pada Gambar 3 dapat dilihat bahwa Jurusan yang paling diminati adalah Teknik dan Bisnis Sepeda Motor. Sedangkan untuk nilai akurasi Algoritma C4.5 dapat dilihat pada Gambar 4 berikut:


PerformanceVector (Performance)

Text View  Annotations

Multiclass Classification Performance  Annotation:

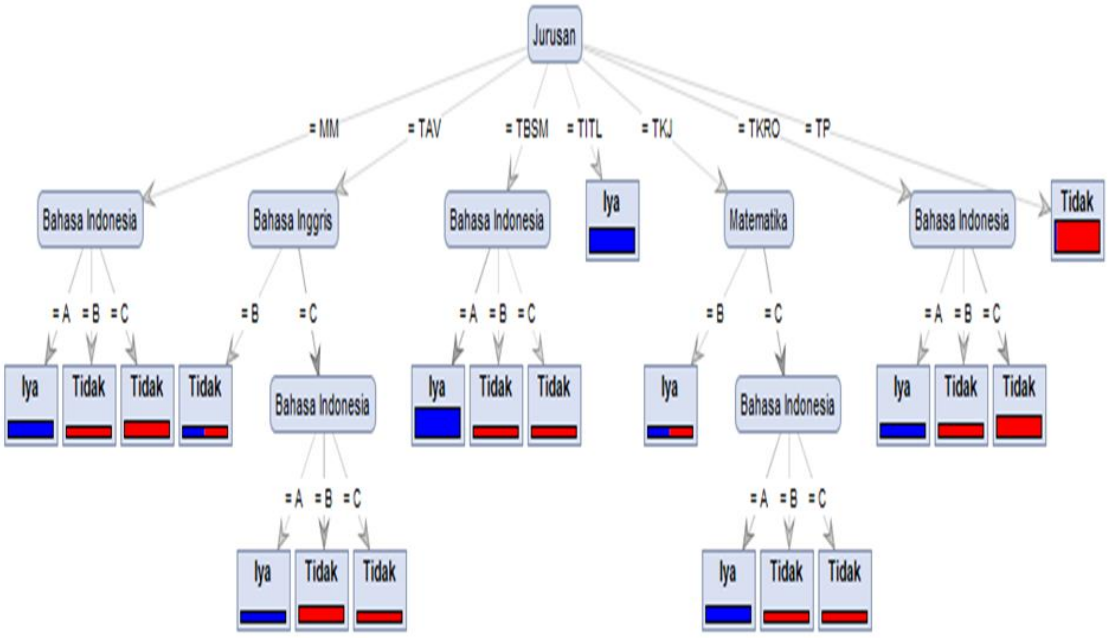
Table View  Plot View

accuracy: 100.00%

| | true Iya | true Tidak | class precision |
|---|---|---|---|
| pred. Iya | 21 | 0 | 100.00% |
| pred. Tidak | 0 | 29 | 100.00% |
| class recall | 100.00% | 100.00% | |

*Gambar 4. Nilai Akurasi Algoritma C4.5*

Dengan pengolahan data menggunakan aplikasi RapidMiner didapat nilai akurasi sistem sebesar 100%. Dari gambar di jelaskan bahwa prediksi tidak adalah 29 dan prediksi iya adalah 21 dengan nilai precision sebesar 100% dan nilai recall sebesar 100%. Maka dari perhitungan manual dan pengujian menggunakan aplikasi RapidMiner yang telah dibandingkan di dapat hasil yang sama yaitu jurusan Teknik dan Bisnis Sepeda Motor.




*Gambar 3. Tampilan hasil decision tree*

## IV. KESIMPULAN

Dari hasil penelitian yang dilakukan dalam proses pengujian sebanyak 100 record data testing yang diuji menyatakan bahwa algoritma C4.5 dapat menghasilkan tingkat akurasi sebesar 100,00%. Penerapan data mining dengan menggunkan metode Algoritma C4.5 ini dapat mempercepat pengambilan keputusan dalam memprediksi peminatan jurusan siswa saat proses masuk.

Sebagai topik penelitian selanjuntnya adalah memaksimalkan atau menambah atribut yang lebih spesifik dan lebih banyak dalam menentukan peminatan jurusan siswa seperti data siswa baru dan penambahan dataset dalam data training dan data testing serta perlu adanya penelitian lebih lanjut dengan melakukan pengujian dengan metode lain seperti naive bayes, ID3 dan lain sebagainya. Agar memperoleh perbandingan dengan tingkat akurasi yang paling tinggi dalam membuat kualifikasi prediksi jurusan siswa ditingkat Sekolah Menengah Kejuruan.

## REFERENSI

[1] Dwi Aditya and Shaufiah, "Pemanfaatan Data Mining pada Penerimaan Mahasiswa Baru Menggunakan Metode AHP dan Algoritma C4.5 Decision Tree," pp. 1–12, 2013

[2] Anik Andriani, " Penerapan Algoritma C4.5 Pada Progam Klasifikasi Mahasiswa Dropout," AMIK BSI Jakarta, 2012.

[3] Dicky Nofriansyah, Gunadi Widi Nurcahyo. 2015. Algoritma *Data Mining* dan Pengujian. Yogyakarta : Deepublish.

[4] David Hartanto Kamagi, Seng Hansun, "Implementasi *Data Mining* dengan Algoritma C4.5 untuk memprediksi Tingkat Kelulusan Mahasiswa," Universitas Multimedia Nusantara, Juni 2014.

[5] Ibnu Fathur Rochman, "Penerapan Algoritma C4.5 Pada Kepuasan Pelanggan Perum DAMRI," Universitas Dian Nuswantoro, 2015

[6] Mochamad Rizki Ilham, P. (2016). Implementasi *Data Mining* Menggunakan Algoritma C4.5 Untuk Prediksi Kepuasan Pelanggan Taksi Kosti. Simplementasi *Data Mining* Menggunakan Algoritma C4.5 Untuk Prediksi Kepuasan Pelanggan Taksi Kosti, Vol. 4, No (5), 11.

[7] Siska Haryati, Aji Sudarsono, and Eko Suryan, "Implementasi *Data Mining* Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5, (Studi Kasus: Universitas Dehasen Bengkulu)," Jurnal Media Infotama", Vol. 11 No. 2, September 2015.

[8] Retno. (2017). *Data Mining* Teori dan Aplikasi *Rapidminer*. Yogyakarta. Gava Media.

[9] Suherman, B. (2018). Implementasi *data mining* untuk memprediksi pemasaran produk helmet dengan algoritma C4.5 pada PT. Indosafety Manufacture.