KALBIScientia
Jurnal Sains dan Teknologi

ISSN 2356 - 4393

Aplikasi Data Mining Berbasis Android Menggunakan Algoritma K-Means Clustering dan Algoritma C4.5 Untuk Memprediksi Pengambilan Jurusan Siswa SMA Kelas X Pada Sekolah Bunda Mulia

Adi Syah Petera Dewata[1)], Putra Zoel Ibrahim[2)], Halim Agung[3)]

Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia
Jalan Lodan Raya No.2, Pademangan, Jakarta Utara, 14430

[1)]*adiwiratta27@gmail.com,*
[2)]*putraibrahim6@gmail.com,*
[3)]*hagung@bundamulia.ac.id*

Abstract: *Bunda Mulia High School is one of the middle schools in the city of Central Jakarta which currently has 2 majors of science and social studies. This student majors can lead learners to focus more on developing their own abilities and interests. Selection of improper majors can be very detrimental to students of their interests and careers in the future. With the majors are expected to maximize the potential, talent or individual talents, so as to maximize academic value. Based on the background, then by applying data mining techniques are expected to help students to determine the appropriate majors in accordance with the applicable criteria. The data mining techniques used in the determination of this department using 2 pieces of method that is K-Means Algorithm and C4.5 Algorithm. While the attributes used consisted of the Student's Number, the value of the exact lesson of science studies, the exact score of social studies, and the value of Psikotest. The results of the study using two methods show that the values of the exact subjects and the psychotest will influence to determine the majors of the new students*

Keywords: *C4.5, data mining, K-Means, majors*

Abstrak: SMA Bunda Mulia merupakan salah satu sekolah menengah di Kota Jakarta Pusat yang saat ini telah memiliki 2 jurusan yaitu IPA dan IPS. Penjurusan siswa ini dapat mengarahkan peserta didik agar lebih focus dalam mengembangakan kemampuan diri dan minat yang dimiliki. Pemilihan jurusan yang tidak tepat bisa sangat merugikan siswa terhadap minat dan karir mereka di masa mendatang. Dengan penjurusan tersebut diharapkan dapat memaksimalkan potensi, bakat atau talenta individu, sehingga dapat memaksimalkan nilai akademisnya. Berdasarkan latar belakang tersebut, maka dengan menerapkan teknik data mining diharapkan dapat membantu siswa untuk menentukan jurusan yang tepat sesuai dengan kriteria yang di terapkan. Adapun teknik data mining yang digunakan dalam penentuan jurusan ini menggunakan 2 buah metode yaitu Algoritma K-Means dan Algoritma C4.5. Sedangkan atribut yang digunakan terdiri dari Nomor Induk, Nilai pelajaran eksak IPA, Nilai pelarajan eksak IPS, dan nilai Psikotest. Hasil penelitian menggunakan 2 metode menunjukkan bahwa nilai-nilai dari mata pelajaran eksak dan psikotest akan mempengaruhi untuk menentukan penjurusan siswa baru

Kata kunci: *C4.5, data mining, K-Means, penjurusan*

I. PENDAHULUAN

Pemilihan jurusan untuk siswa SMA merupakan salah satu hal yang paling penting dikarenakan penjurusa di SMA menentukan langkah awal dalam menentukan masa depan. Dilihat dari data nilai siswa Sekolah Bunda Mulia pada 5-6 tahun yang lalu, sebagian siswa masuk ke penjurusan yang kurang tepat sehingga membuat siswa sulit untuk mempelajari pelajaran yang ada dalam penjurusan tersebut. Oleh karena itu, dengan memanfaatkan data siswa dan data akademik siswa, dapat diketahui pola pengelompokkan dan pengambilan jurusan berdasarkan nilai mata pelajaran yang terdapat dalam kelas X sebagai syarat memasuki penjurusan tersebut. Untuk dapat mengetahui pola tersebut, dibutuhkan algoritma metode *data mining* yang sesuai dengan kasus tersebut.

Pada penelitian sebelumnya yang berjudul Implementasi Algoritma K-Means *Clustering* Untuk Menentukan Strategi Marketing President University oleh Johan Oscar Ong dapat disimpulkan hasil penelitian berupa, dengan melakukan promosi dengan mengirim tim marketing yang sesuai dengan jurusan yang paling banyak diminati dan melakukan promosi di kota-kota di Indonesia yang didasarkan pada tingkat kemampuan akademik dari calon mahasiswa[1].

Penelitian sebelumnya yang ke-2 yang berjudul Klasifikasi Dan Klastering Siswa Sma Negeri 3 Boyolali, dapat disimpulkan hasil penelitian

Klasifikasi penjurusan siswa dengan menggunakan decision tree menunjukkan bahwa variable yang paling tinggi pengaruhnya terhadap penjurusan siswa adalah nilai rata-rata IPA karena variable nilai tersebut menempati sebagai simpul akar pada diagram pohon keputusan. [2].

Penelitian Sebelumnya yang ke-3 yang berjudul Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta oleh Yusuf Sulistyo Nugroho dengan hasil penelitian berupa mengindikasikan bahwa variabel yang perlu digunakan sebagai pertimbangan bagi Fakultas Komunikasi dan Informatika UMS untuk memperoleh tingkat predikat kelulusan yang maksimal adalah peran serta mahasiswa untuk menjadi asisten. Secara umum probabilitas predikat "Cumlaude" pada kelompok mahasiswa yang pernah menjadi asisten lebih tinggi dibandingkan dengan yang tidak pernah menjadi asisten. Seorang mahasiswa dari kelompok yang pernah menjadi asisten jika berasal dari jurusan IPA semasa sekolah menengah atas memiliki probabilitas predikat kelulusan "Cumlaude" yang lebih tinggi dibandingkan dengan mahasiswa dari jurusan lainnya[3].

Penelitian Sebelumnya yang ke-4 yang berjudul Implementasi *Data mining* Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan Mahasiswa oleh David Hartanto Kamagi, Seng Hasnun dengan hasil penelitian *Data mining* dengan algoritma C4.5 dapat diimplementasikan untuk memprediksi tingkat kelulusan mahasiswa dengan empat kategori yaitu lulus cepat, lulus tepat, lulus terlambat dan drop out. Attribute yang paling berpengaruh dalam hasil prediksi adalah IPS semester enam. Hasil prediksi kelulusan dari aplikasi penelitian ini dapat membantu bagian program studi untuk mengetahui status kelulusan mahasiswa. Hal ini dapat menjadi rekomendasi pengambilan mata kuliah bagi mahasiswa untuk semester berikutnya seperti skripsi dan magang. Dengan hal tersebut mahasiswa bisa lulus minimal tepat waktu [4].

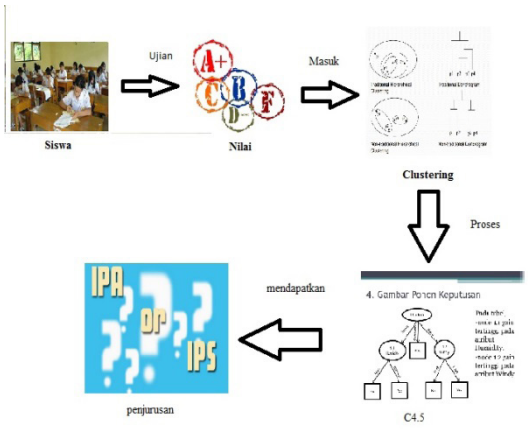
Penelitian sebelumnya yang ke -5 yang berjudul Sistem Pendukung Keputusan Pemilihan Juusan Siswa Dengan Menggunakan Metode Weighted Product oleh Ingot Seen Sianturi dengan hasil penelitian Dengan Weighted Product dapat membantu dalam pengambilan keputusan untuk menentukan penjurusan siswa lebih efisien sehingga siswa lebih cepat mendapatkan informasi tentang penjurusan [5]. Berdasarkan penelitian terdahulu dan jurnal yang dipresentasikan pada International Conference on *Data mining* (ICDM), dengan memperhatikan kelebihan suatu algoritma dalam mempelajari data, terpilihlah Algoritma K-Means *Clustering* sebagai algoritma pengelompokkan data dan Algoritma C4.5 sebagai algoritma pengambilan keputusan dalam *data mining.*

Berdasarkan permasalahan yang dipaparkan, penulis mencoba untuk menerapkan pengelompokkan dan pengambilan keputusan menggunakan algoritma K-Mean *Clustering* dan C4.5 sebagai metode penyelesaian masalah yang ada.

II. METODE PENELITIAN

Metodologi dalam penelitian ini menggunakan metode pengembangan sistem informasi yaitu SDLC (*Software Development Live Cycle*) dimana terbagi menjadi beberapa tahapan, metode tersebut menggunakan metode waterfall model [6].

Penelitian ini dikerjakan dengan metodologi yang digunakan dalam bidang rekayasa software, yang terdiri dari 3 fase, yaitu: (1) Fase requirement atau penelusuran kebutuhan. Pada fase ini penulis mencari tahu beberapa hal, seperti: apa yang dibutuhkan?, apa tujuan dari aplikasi ini?, apa yang ingin dicapai?, apakah ada referensi atau contoh?, siapa sasaran pengguna aplikasi ini?; (2) Fase analisis. Berdasarkan hasil penelitian kebutuhan, maka akan diputuskan seperti apa aplikasi apa yang ingin dibuat, fitur apa saja yang dibutuhkan, masalah yang mungkin dihadapi, dan apa saja yang diperlukan dalam proses pengembangan; (3) Fase Perancangan. Pada tahap ini akan dibuat rancangan aplikasi berdasarkan hasil analisa sebelumnya. Misalnya membuat *Use Case*, *Activity Diagram* [7] dan Normalisasikan data yang ada; dan (4) Fase Pengembangan. Fase ini merupakan tahapan implementasi dari hasil analisa dan perencanaan. Pada tahap ini akan dibuat rancangan algoritma K-Means dan algoritma C4.5 yang



Gambar 1. Situasi sekolah Bunda Mulia

digunakan sebagai basis dalam membuat aplikasi memprediksi penjurusan dalam penelitian lanjutan.

Dari Fase diatas maka situasi permasalahan yang ada pada sekolah SMA Bunda Mulia digambarkan pada Gambar 1. Dari gambar situasi masalah dengan gambar 1 terlihat bahwa diperlukannya: (1) Pengolahan kembali data-data siswa yang diperoleh dari hasil tes IQ dan nilai akademik siswa; (2) Pengidentifikasian data siswa yang diperlukan untuk mengetahui siswa tersebut jurusan IPA atau IPS atau masih belum memiliki penjurusan.

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan dalam basis data. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi dan pengetahuan yang terkait dari basis data besar[8].

Data mining sering juga disebut *Knowledge Discovery in Database* (KDD) yaitu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar.

Data mining, menurut *Gartner Group* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika.

Data mining dikelompokkan ke dalam Teknik Pohon Keputusan, Bayesian (*Naive Bayesian* dan *Bayesian Belief Networks*), Jaringan Saraf Tiruan (*Backpropagation*), Teknik yang berbasis konsep dari penambangan aturan-aturan asosiasi, dan teknik lain (*K-Nearest Neighboor*, algoritma genetik, teknik dengan pendekatan himpunan *rough* dan *fuzzy*)[9]. Secara umum, Proses Klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari *data training* dan klasifikasi kasus.

Dari beberapa definisi diatas dapat disimpulkan bahwa *data mining* adalah sekumpulan informasi yang didapatkan dari berbagai sistem operasi dalam informasi yang didapatkan, transformasi agar data terintegrasi, dan dapat digunakan untuk melakukan analisis dalam pengambilan suatu keputusan.

Decision Tree merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas bercabang, dimana simpul *internal* maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi label nilai atribut yang mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut: (1) Pilih atribut sebagai *Node* akar; (2)Buat cabang untuk tiap-tiap nilai; (3)Bagi kasus dalam cabang; dan (4)Ulangi proses untuk setiap setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama[10].

Untuk memilih atribut sebagai *Node* akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Proses pembuatan rumus akan di jelaskan seperti pada Gambar 2:

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

Gambar 2. Perhitungan Gain

Keterangan:

S: himpunan kasus

A: Atribut

N: jumlah partisi atribut

|Si|: jumlah kasus pada partisi ke-i

|S|: jumlah kasus dalam S

Dan cara mencari *entropy* yang akan digunakan dalam algoritma ini dapat di cari dengan menggunakan rumus yang dapat dilihat pada Gambar 3.

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

Gambar 3. Perhitungan entropy

Keterangan:

S: himpunan kasus

A: fitur

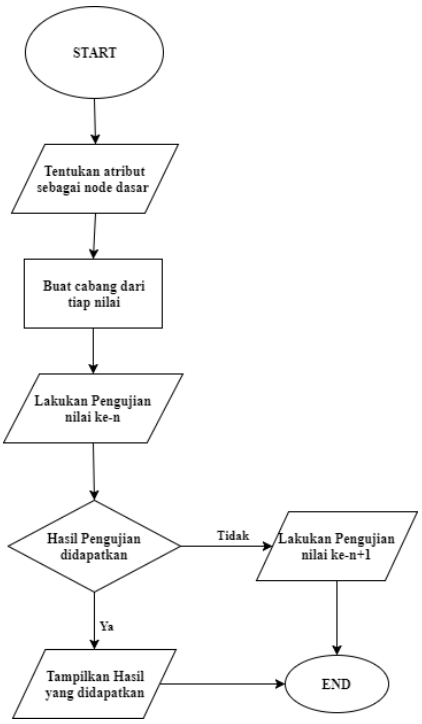
N: jumlah partisi S

Pi: proporsi dari Si terhadap S

Kedua rumus diatas akan di pakai untuk mencari nilai-nilai yang akan di jadikan sebagai atribut dalam mengambil keputusan. Berikut adalah flowchart diagram dari algoritma C4.5 yang dapat dilihat pada Gambar 4.

K-Means merupakan algoritma yang umum digunakan untuk *clustering* dokumen. Prinsip utama *K-Means* adalah menyusun *k prototype* atau pusat massa (*centroid*) dari sekumpulan data berdimensi *n*". Sebelum diterapkan proses algoritma *K-means*, dokumen akan di*preprocessing* terlebih dahulu. Kemudian dokumen direpresentasikan sebagai vektor yang memiliki *term* dengan nilai tertentu.

K-Means merupakan salah satu metode data *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok. Metode ini mempartisi data ke dalam *cluster* atau kelompok sehingga data yang memiliki karakteristik sama dikelompokkan ke dalam satu *cluster* yang sama.



Gambar 4. Flowchart C4.5

Langkah melakukan *clustering* dengan ketode K-Means adalah sebagai berikut: [1]

1). Pilih jumlah *cluster* K

2). Inisialisasi K pusat *cluster* ini bias dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat *cluster* diberi nilai awal dengan angka-angka random.

3). Alokasi semua data atau objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan pada Gambar 5:

$$D(ij) = \sqrt{(X1i - X1j)2 + (X2i - X2j)2 + \cdots + (Xki - Xkj)2}$$

Gambar 5. Jarak data ke cluster

Keterangan:

D(ij): Jarak data-(i) ke pusat *cluster* (j)

Xki: data-(i) pada atribut-(k)

Xkj: titik pusat-(j) pada atribut-(k)

4.) Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data atau objek dalam *cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata bukan satu-satunya ukuran yang bisa dipakai. Rumus rata-rata Cluster dapat dilihat pada Gambar 6

$$Rk = \frac{1}{Nk}(X1k + X2k + \cdots + Xnk)$$

Gambar 6. Rata-rata cluster

Keterangan:

Rk: rata-rata *cluster* baru

Nk: jumlah training pattern pada *cluster*-(k)

Xnl: pola-(n) yang menjadi bagian dari *cluster*-(k)

5.) Tugaskan lagi tiap objek memakai ousat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai. Atau, kembali ke langkah 3 sampai pusat *cluster* tidak berubah lagi.

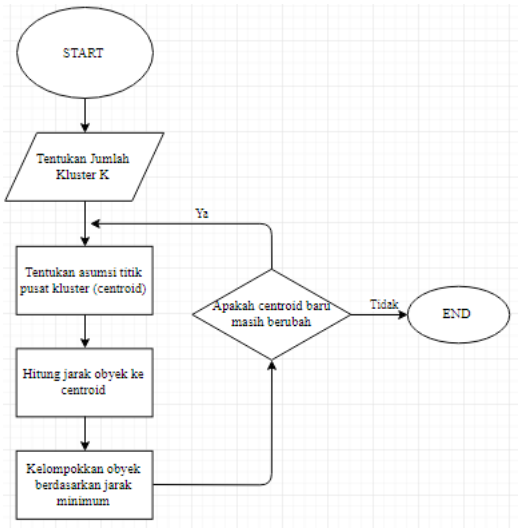
Data yang diperlukan dalam penelitian ini diperoleh melalui bagian akademik SMA Bunda Mulia yang berupa data siswa yang telah lulus ujian masuk SMA dimulai pada tahun ajaran 2014-2015. Data tersebut berisi data diri siswa yang telah lulus

No	Nis	Mat	Fis	Kim	Bio	Sos	Eko	Geo	Vrb	Lgk	Art	Total	Jurusan
1	3955	8,0	8,6	7,8	7,5	8,5	8,0	8,2	8,0	8,0	9,0	81,6	BMIPA
2	3958	7,0	7,8	7,7	6,8	7,5	7,8	8,8	8,0	9,0	8,0	78,4	BMIPA
3	3961	7,8	8,2	8,5	8,0	8,2	8,2	8,2	9,0	7,0	8,0	81,1	BMIPA
4	3950	7,5	7,5	7,5	7,5	7,8	7,5	7,6	6,0	5,0	5,0	68,9	BMIPS
5	3951	7,5	7,5	7,6	7,5	7,5	6,9	7,6	7,0	8,0	6,0	73,1	BMIPS
6	3952	7,5	7,5	7,5	7,5	8,8	8,0	8,3	7,0	7,0	4,0	73,1	BMIPS
7	3899	6,7	6,6	7,0	7,1	7,5	7,8	8,2	8,2	8,6	8,6	76,3	BMIPS
8	3900	7,3	6,6	7,0	7,4	7,6	7,9	7,8	7,5	7,5	7,7	74,3	BMIPS
9	3901	7,3	6,9	7,1	7,6	8,1	8,0	8,2	7,0	7,5	7,0	74,7	BMIPS
10	3479	9,1	9,1	8,3	9,4	7,5	7,0	7,8	6,2	7,6	7,3	79,3	BMIPA
11	3548	9,2	8,4	8,6	9,0	8,1	7,4	8,1	8,2	7,6	8,3	82,8	BMIPA
12	3516	8,5	8,2	8,1	7,8	7,6	7,2	7,9	8,7	8,0	8,1	80,1	BMIPA
13	3549	9,2	9,9	9,2	9,5	7,7	8,1	7,4	7,3	7,9	7,2	83,3	BMIPA
14	3481	8,8	9,1	7,6	8,5	8,6	7,6	7,2	7,9	7,8	8,4	81,4	BMIPA
15	3690	7,5	9,0	7,7	8,3	8,2	8,1	7,5	6,0	7,1	7,1	76,5	BMIPA
16	3908	9,4	8,2	8,6	8,0	8,2	8,2	8,8	7,5	7,7	7,6	82,2	BMIPA
17	3909	7,4	7,9	7,1	7,9	8,2	7,6	8,7	7,5	7,5	7,2	77,0	BMIPA
18	3910	7,5	7,6	7,9	7,6	8,0	8,1	8,6	7,5	7,5	7,5	77,8	BMIPA
19	3911	7,3	6,0	7,4	7,3	7,8	7,5	7,9	7,5	7,5	7,5	73,7	BMIPS
20	3912	7,5	7,6	7,5	8,5	8,8	8,0	8,7	7,6	7,5	7,8	79,5	BMIPA
21	3913	7,3	7,1	7,3	8,2	8,6	7,9	8,6	8,0	8,1	8,1	79,2	BMIPA
22	3914	7,5	8,2	7,0	8,0	8,3	8,0	8,0	7,5	8,1	7,8	78,4	BMIPA
23	3915	7,6	8,3	7,0	8,1	8,1	8,1	7,8	8,0	8,5	8,4	79,9	BMIPA
24	3916	7,3	7,2	7,0	7,7	8,1	7,8	8,0	9,2	8,1	7,5	77,9	BMIPA
25	3917	8,0	9,0	8,0	8,6	8,5	8,4	8,7	5,4	8,5	8,2	81,3	BMIPA
26	3918	7,4	7,5	7,6	8,2	8,1	7,5	8,5	8,4	7,9	7,5	78,6	BMIPA
27	3919	8,0	8,3	8,4	8,5	8,5	8,2	8,3	6,6	7,8	7,6	80,2	BMIPA
28	3920	7,3	7,1	7,0	7,5	8,1	8,2	7,9	7,8	7,9	7,7	76,5	BMIPA
29	3921	7,0	6,0	7,0	6,5	7,5	7,0	7,7	9,4	9,4	9,2	76,7	BMIPS
30	3922	6,5	6,8	7,0	6,0	7,0	7,0	7,9	8,5	8,8	7,6	73,1	BMIPS
31	3923	7,3	6,0	7,0	7,1	7,5	7,0	7,9	9,0	8,8	8,3	75,9	BMIPS
32	3924	7,5	6,6	7,0	7,3	7,5	7,7	7,9	7,5	7,7	7,6	74,3	BMIPS
33	3925	7,3	6,0	7,0	6,5	7,5	7,3	7,9	8,0	8,1	8,1	73,7	BMIPS
34	3926	7,0	6,0	7,0	7,1	7,5	7,1	8,1	7,5	8,1	7,8	73,2	BMIPS
35	3927	7,6	7,1	7,2	7,4	7,9	7,6	7,9	8,0	8,1	8,1	76,9	BMIPA
36	3928	6,1	6,2	6,5	6,5	7,5	7,2	7,9	8,0	8,1	8,1	72,1	BMIPS
37	3929	7,8	7,8	7,3	7,8	7,8	7,0	8,3	7,5	8,1	7,8	77,2	BMIPA
38	3930	7,3	6,6	7,0	7,1	7,7	7,7	7,9	8,0	8,5	8,4	76,2	BMIPS
39	3931	9,9	9,2	9,6	9,9	9,6	9,0	9,3	9,2	8,1	7,5	91,3	BMIPA
40	3932	7,5	7,0	7,3	7,3	7,8	7,9	7,9	5,4	8,5	8,2	74,8	BMIPA
41	3933	7,7	8,2	7,1	7,4	8,2	7,7	7,9	8,4	7,9	7,5	78,0	BMIPA
42	3934	7,3	7,1	7,0	7,4	7,5	7,2	7,9	6,6	7,8	7,6	73,4	BMIPA
43	3935	7,9	6,8	7,4	7,5	7,9	7,4	8,3	7,8	7,9	7,7	76,6	BMIPS
44	3936	7,3	7,1	7,3	7,4	7,6	7,8	8,1	6,2	7,6	7,3	73,7	BMIPA
45	3937	7,3	6,0	7,0	7,2	7,6	7,6	7,7	8,2	7,6	8,3	74,5	BMIPS
46	3938	7,3	6,0	6,5	7,0	7,7	7,2	7,9	8,7	8,0	8,1	74,4	BMIPS
47	3939	7,4	6,4	6,5	7,1	7,7	7,8	8,2	7,3	7,9	7,2	73,5	BMIPS
48	3940	8,0	8,1	8,4	8,0	8,5	8,1	9,3	7,9	7,8	8,4	82,5	BMIPA
49	3941	8,1	7,4	7,1	7,4	7,6	7,6	8,6	6,0	7,1	7,1	74,0	BMIPA
50	3942	7,3	6,2	6,5	7,4	7,5	7,9	7,9	7,1	7,3	7,0	72,1	BMIPS
51	3944	9,0	8,4	8,5	9,1	9,0	8,2	8,9	6,1	7,0	6,5	80,7	BMIPA
52	3945	7,7	6,6	7,3	7,3	7,6	8,2	7,9	7,3	7,1	7,2	74,2	BMIPS

Gambar 7. Data siswa SMA Bunda Mulia

dari ujian, namun pada penelitian ini hanya beberapa atribut saja yang digunakan seperti nomor induk siswa, nilai mata pelajaran eksak IPA, nilai mata pelajaran eksak IPS, dan psikotest. Data siswa yang telah menjadi siswa SMA Bunda Mulia dapat dilihat pada Gambar 7.

Diagram alir dari algoritma K-Means *Clustering* dapat dilihat pada Gambar 8.



Gambar 8. Flowchart K-Means cluster

III. HASIL DAN PEMBAHASAN

Setelah semua data siswa pada tahun ajaran 2014-2015 ditransformasikan dalam bentuk angka, maka data tersebut telah dapat dikelompokkan dengan menggunakan data tersebut menjadi beberapa dengan menggunakan algoritma K-Means *Clustering*. Untuk dapat melakukan pengelompokkan data tersebut menjadi beberapa *cluster*, perlu dilakukan beberapa langkah, yaitu:

1) Tentukan jumlah *cluster* yang diinginkan. Dalam penelitian ini data yang ada akan di kelompokkan menjadi dua *cluster*.
2) Tentukan titik pusat awal dari setiap *cluster*. Dalam penelitian ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap *cluster* dapat dilihat pada gambar dibawah ini.

Titik Pusat Awal	Nis	Jurusan	Mtk	Fis	Kim	Bio	Sos	Eko	Geo	Vrb	Lgk	Art	Total
Cluster 1	3958	BMIPA	7,0	7,8	7,7	6,8	7,5	7,8	8,8	8,0	9,0	8,0	78,4
Cluster 2	3951	BMIPS	7,5	7,5	7,6	7,5	7,5	6,9	7,6	7,0	8,0	6,0	73,1

3) Tempatkan setiap data pada *cluster*. Dalam penelitian ini digunakan metode hard k-means untuk mengalokasikan setiap data ke dalam suatu *cluster*, sehingga data akan dimasukan dalam suatu *cluster* yang memiliki jarak paling dekat dengan titik pusat dari setiap *cluster* . Untuk mengetahui *cluster* mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap *cluster*. Sebagai contoh, akan dihitung jarak dari data siswa pertama ke pusat *cluster* pertama yang dapat dilihat pada Gambar 10.

$$D(1,1)$$
$$= \sqrt{(8,0-7,0)^2+(8,6-7,8)^2+(7,8-7,7)^2+(7,5-6,8)^2+(8,5-7,5)^2+(8,0-7,8)^2+(8,2-8,8)^2+(8,0-8,0)^2+(8,0-9,0)^2+(9,0-8,0)^2}$$
$$= 2,354$$

Gambar 10. Perhitungan data pertama dengan pusat cluster pertama

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data siswa pertama dengan pusat *cluster* pertama adalah 2,354.

Jarak data siswa pertama ke pusat *cluster* kedua seperti pada gambar 11

$$D(1,2)$$
$$= \sqrt{(8,0-7,5)^2+(8,6-7,5)^2+(7,8-7,6)^2+(7,5-7,5)^2+(8,5-7,5)^2+(8,0-6,9)^2+(8,2-7,6)^2+(8,0-7,0)^2+(8,0-8,0)^2+(9,0-6,0)^2}$$
$$= 3,751$$

Gambar 11. Perhitungan data pertama dengan pusat cluster kedua

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data siswa pertama dengan pusat *cluster* kedua adalah 3,751.

Berdasarkan hasil kedua perhitungan di atas dapat disimpulkan bahwa jarak data siswa pertama yang paling dekat dengan *cluster* 1, sehingga data siswa pertama dimasukkan ke dalam *cluster* 1. Hasil perhitungan selengkapnya untuk 50 data siswa dapat dilihat pada gambar 12.

4.) Setelah semua data ditempat ke dalam *cluster*

No	Nis	Jurusan	Mtk	Fis	Kim	Bio	Sos	Eko	Geo	Vrb	Lgk	Art	Total	Jarak Ke		Jarak terdekat ke *Cluster*
														C1	C2	
1	3955	BMIPA	8,0	8,6	7,8	7,5	8,5	8,0	8,2	8,0	8,0	9,0	81,6	2,354	3,751	1
2	3961	BMIPA	7,8	8,2	8,5	8,0	8,2	8,2	8,2	9,0	7,0	8,0	81,1	2,982	3,630	1
3	3950	BMIPS	7,5	7,5	7,5	7,5	7,8	7,5	7,6	6,0	5,0	5,0	68,9	5,612	3,385	2
4	3952	BMIPS	7,5	7,5	7,5	7,5	8,8	8,0	8,3	7,0	7,0	4,0	73,1	4,884	2,898	2
5	3899	BMIPS	6,7	6,6	7,0	7,1	7,5	7,8	8,2	8,2	8,6	8,6	76,3	1,741	3,421	1
6	3900	BMIPS	7,3	6,6	7,0	7,4	7,6	7,9	7,8	7,5	7,5	7,7	74,3	2,447	2,379	2
7	3901	BMIPS	7,3	6,9	7,1	7,6	8,1	8,0	8,2	7,0	7,5	7,0	74,7	2,629	1,960	2
8	3479	BMIPA	9,1	9,1	8,3	9,4	7,5	7,0	7,8	6,2	7,6	7,3	79,3	4,517	3,404	2
9	3548	BMIPA	9,2	8,4	8,6	9,0	8,1	7,4	8,1	8,2	7,6	8,3	82,8	3,730	3,818	1
10	3516	BMIPA	8,5	8,2	8,1	7,8	7,6	7,2	7,9	8,7	8,0	8,1	80,1	2,501	3,064	1
11	3549	BMIPA	9,2	9,9	9,2	9,5	7,7	8,1	7,4	7,3	7,9	7,2	83,3	4,793	4,235	2
12	3481	BMIPA	8,8	9,1	7,6	8,5	8,6	7,6	7,2	7,9	7,8	8,4	81,4	3,598	3,677	1
13	3690	BMIPA	7,5	9,0	7,7	8,3	8,2	8,1	7,5	6,0	7,1	7,1	76,5	3,826	2,815	2
14	3908	BMIPA	9,4	8,2	8,6	8,0	8,2	8,2	8,8	7,5	7,7	7,6	82,2	3,305	3,445	1
15	3909	BMIPA	7,4	7,9	7,1	7,9	8,2	7,6	8,7	7,5	7,5	7,2	77,0	2,328	2,170	2
16	3910	BMIPA	7,5	7,6	7,9	7,6	8,0	8,1	8,6	7,5	7,5	7,5	77,8	2,025	2,356	1
17	3911	BMIPS	7,3	6,0	7,4	7,3	7,8	7,5	7,9	7,5	7,5	7,5	73,7	2,722	2,379	2
18	3912	BMIPA	7,5	7,6	7,5	8,5	8,8	8,0	8,7	7,6	7,5	7,8	79,5	2,722	2,997	1
19	3913	BMIPA	7,3	7,1	7,3	8,2	8,6	7,9	8,6	8,0	8,1	8,1	79,2	2,186	3,068	1
20	3914	BMIPA	7,5	8,2	7,0	8,0	8,3	8,0	8,0	7,5	8,1	7,8	78,4	2,182	2,571	1
21	3915	BMIPA	7,6	8,3	7,0	8,1	8,1	8,1	7,8	8,0	8,5	8,4	79,9	2,156	3,197	1
22	3916	BMIPA	7,3	7,2	7,0	7,7	8,1	7,8	8,0	9,2	8,1	7,5	77,9	2,291	2,993	1
23	3917	BMIPA	8,0	9,0	8,0	8,6	8,5	8,4	8,7	5,4	8,5	8,2	81,3	3,767	3,997	1
24	3918	BMIPA	7,4	7,5	7,6	8,2	8,1	7,5	8,5	8,4	7,9	7,5	78,6	2,107	2,517	1
25	3919	BMIPA	8,0	8,3	8,4	8,5	8,5	8,2	8,3	6,6	7,8	7,6	80,2	3,103	2,911	2
26	3920	BMIPA	7,3	7,1	7,0	7,5	8,1	8,2	7,9	7,8	7,9	7,7	76,5	2,067	2,482	1
27	3921	BMIPS	7,0	6,0	7,0	6,5	7,5	7,0	7,7	9,4	9,4	9,2	76,7	3,038	4,673	1
28	3922	BMIPS	6,5	6,8	7,0	6,0	7,0	7,0	7,9	8,5	8,8	7,6	73,1	2,128	3,146	1
29	3923	BMIPS	7,3	6,0	7,0	7,1	7,5	7,0	7,9	9,0	8,8	8,3	75,9	2,548	3,583	1
30	3924	BMIPS	7,5	6,6	7,0	7,3	7,5	7,7	7,9	7,5	7,7	7,6	74,3	2,313	2,200	2
31	3925	BMIPS	7,3	6,0	7,0	6,5	7,5	7,3	7,9	8,0	8,1	8,1	73,7	2,406	3,053	1
32	3926	BMIPS	7,0	6,0	7,0	7,1	7,5	7,1	8,1	7,5	8,1	7,8	73,2	2,429	2,610	1
33	3927	BMIPA	7,6	7,1	7,2	7,4	7,9	7,6	7,9	8,0	8,1	8,1	76,9	1,814	2,550	1
34	3928	BMIPS	6,1	6,2	6,5	6,5	7,5	7,2	7,9	8,0	8,1	8,1	72,1	2,625	3,385	1
35	3929	BMIPA	7,8	7,8	7,3	7,8	7,8	7,0	8,3	7,5	8,1	7,8	77,2	1,970	2,110	1
36	3930	BMIPS	7,3	6,6	7,0	7,1	7,7	7,7	7,9	8,0	8,5	8,4	76,2	1,838	3,025	1
37	3931	BMIPA	9,9	9,2	9,6	9,9	9,6	9,0	9,3	9,2	8,1	7,5	91,3	5,674	6,101	1
38	3932	BMIPA	7,5	7,0	7,3	7,3	7,8	7,9	7,9	5,4	8,5	8,2	74,8	3,043	3,035	2
39	3933	BMIPA	7,7	8,2	7,1	7,4	8,2	7,7	7,9	8,4	7,9	7,5	78,0	2,088	2,513	1
40	3934	BMIPA	7,3	7,1	7,0	7,4	7,5	7,2	7,9	6,6	7,8	7,6	73,4	2,488	1,875	2
41	3935	BMIPS	7,9	6,8	7,4	7,5	7,9	7,4	8,3	7,8	7,9	7,7	76,6	2,084	2,247	1
42	3936	BMIPA	7,3	7,1	7,3	7,4	7,6	7,8	8,1	6,2	7,6	7,3	73,7	2,700	1,965	2
43	3937	BMIPS	7,3	6,0	7,0	7,2	7,6	7,6	7,7	8,2	7,6	8,3	74,5	2,707	3,184	1
44	3938	BMIPS	7,3	6,0	6,5	7,0	7,7	7,2	7,9	8,7	8,0	8,1	74,4	2,742	3,357	1
45	3939	BMIPS	7,4	6,4	6,5	7,1	7,7	7,8	8,2	7,3	7,9	7,2	73,5	2,528	2,311	2
46	3940	BMIPA	8,0	8,1	8,4	8,0	8,5	8,1	9,3	7,9	7,8	8,4	82,5	2,443	3,666	1
47	3941	BMIPA	8,1	7,4	7,1	7,4	7,6	7,6	8,6	6,0	7,1	7,1	74,0	3,256	2,269	2
48	3942	BMIPS	7,3	6,2	6,5	7,4	7,5	7,9	7,9	7,1	7,3	7,0	72,1	3,158	2,354	2
49	3944	BMIPA	9,0	8,4	8,5	9,1	9,0	8,2	8,9	6,1	7,0	6,5	80,7	4,751	3,758	2
50	3945	BMIPS	7,7	6,6	7,3	7,3	7,6	8,2	7,9	7,3	7,1	7,2	74,2	2,839	2,261	2

Gambar 12. Hasil perhitungan setiap data ke setiap cluster

yang terdekat, kemudian hitung kembali pusat *cluster* yang baru berdasarkan rata-rata angggota yang ada pada *cluster* tersebut.

5.) Setelah didapatkan titik pusat yang baru dari setiap *cluster*, lakukan kembali dari langkah ketiga hingga titik pusat dari setiap *cluster* tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu *cluster* ke *cluster* yang lain.

Dalam penelitian ini, iterasi *clustering* data siswa terjadi sebanyak 8 kali iterasi. Pada iterasi ke-8 ini, titik pusat dari setiap *cluster* sudah tidak berubah dan tidak ada lagi data yang berpindah dari satu *cluster* ke *cluster* lain. Berdasarkan hasil pengelompokkan data menggunakan metode k-means *clustering* didapatkan hasil *cluster* hingga iterasi ke-7 dimana titik pusat tidak lagi berubah dan tidak ada data yang berpindah antar *cluster*. Pada penelitian ini, jika proses *cluster* sudah selesai, maka akan dilanjutkan pada proses c4.5.

Dari hasil penelitian kami berdasarkan data *cluster* yang sudah didapatkan, proses c4.5 dapat dilihat pada Gambar 13.

Node	Atribut	Nilai Atribut	Jumlah Kasus Total	BMIPA	BMIPS	Entropy	Gain
1	Total	Total	52	31	21	0,973156	
	Mtk						0,079271
		Lulus	49	31	18	0,948613	
		Tidak Lulus	3	0	3	0	
	Fis						0,691641
		Lulus	34	31	3	0,430552	
		Tidak Lulus	18	0	18	0	
	Kim						0,107555
		Lulus	48	31	17	0,937734	
		Tidak Lulus	4	0	4	0	
	Bio						0
		Lulus	52	31	21	0,973156	
		Tidak Lulus	0	0	0	0	
	Sos						0
		Lulus	52	31	21	0,973156	
		Tidak Lulus	0	0	0	0	
	Eko						0,025559
		Lulus	51	31	20	0,966177	
		Tidak Lulus	1	0	1	0	
	Geo						0
		Lulus	52	31	21	0,973156	
		Tidak Lulus	0	0	0	0	
	Vrb						0,031924
		Lulus	42	23	19	0,993447	
		Tidak Lulus	10	8	2	0,721928	
	Lgk						0,025559
		Lulus	51	31	20	0,966177	
		Tidak Lulus	1	0	1	0	
	Art						0,107555
		Lulus	48	31	17	0,937734	
		Tidak Lulus	4	0	4	0	

Gambar 13. Proses C4.5

Pada Total kolom *entropy* pada tabel diatas, dihitung dengan persamaan yang dapat dilihat pada Gambar 14.

Sementara itu, nilai *Gain* pada barus Matematika (Mtk) dihitung dengan menggunakan persamaan sebagai berikut:

$$Entropy\ (Total) = \left(-\frac{31}{52} * log2\left(\frac{31}{52}\right)\right) + \left(-\frac{21}{52} * log2\left(\frac{21}{52}\right)\right)$$

$$Entropy\ (Total) = 0{,}9731560354$$

Gambar 14. Perhitungan Entropy Total

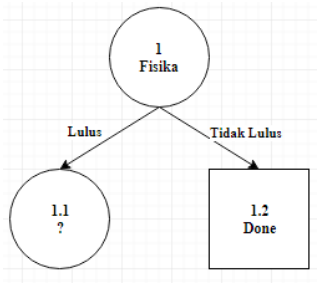
$$Gain(Total, Mtk) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Mtk|}{|Total|} * Entropy(Mtk_i)$$

$$Gain(Total, Mtk) = 0{,}9731560354 - \left(\left(\frac{49}{52} * 0{,}948613\right) + \left(\frac{3}{52} * 0\right)\right)$$

$$Gain(Total, Mtk) = 0{,}079271$$

Gambar 15. Perhitungan Gain Total-Matematika

Dari hasil pada Gambar 13 dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah Fisika, yaitu sebesar 0,69. Dengan demikian, Fisika akan kami ambil sebagai *Node* akar. Ada dua nilai atribut dari Fisika, yaitu Lulus dan Tidak Lulus. Dari kedua nilai atribut tersebut, nilai atribut yang memiliki nilai *Entropy* terkecil akan mengklasifikasikan kasus menjadi satu keputusan, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut dengan *Entropy* yang besar masih perlu dilakukan perhitungan lagi. Setelah dilakukan perhitungan, maka terbentuklah pohon keputusan sementara seperti gambar dibawah.



Gambar 16. Pohon Keputusan Node 1

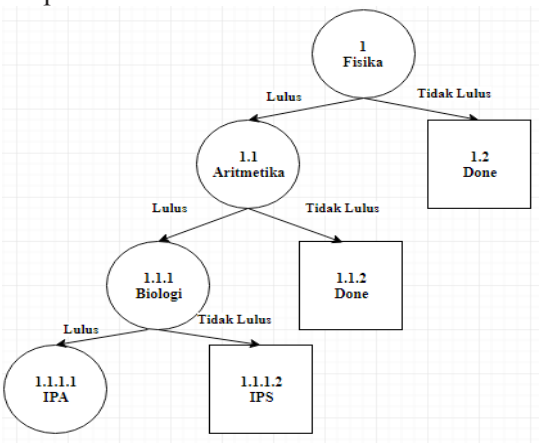
Pada Gambar 16, belum diketahui *node* 1.1 akan dilakukan pada mata pelajaran apa, sehingga harus dilakukan dengan penghitungan kembali. Hasil pengujian akan dipaparkan dalam gambar 17.

Pada Gambar 17 terlihat jelas bahwa mata pelajaran Fisika dengan atribut Lulus mempunyai atribut lain dengan *gain* yang tinggi yaitu Aritmetika. Dikarenakan pada mata pelajaran Matematika,Kimia, Sosiologi, Geografi, sudah memiliki nilai atribut yang pasti yakni 1 dan 0, maka mata pelajaran tersebut tidak di ikut sertakan dalam pengujian berikutnya. Sehingga dapat disimpulkan bahwa pada aplikasi ini algoritma C4.5 memprediksi penjurusan dengan mempertimbangkan nilai dalam mata pelajaran Fisika, Aritmetika, sedangkan Matematika dan Kimia akan di ikutsertakan dalam memprediksi penjurusan IPA karena mereka memiliki nilai pasti dimana siswa

Node	Atribut	Nilai Atribut	Jumlah Kasus Total	BMIPA	BMIPS	Entropy	Gain
1.1	Fis-Lulus	Total	34	31	3	0,430552	
	Mtk						0
		Lulus	34	31	3	0,430552	
		Tidak Lulus	0	0	0	0	
	Kim						0
		Lulus	34	31	3	0,430552	
		Tidak Lulus	0	0	0	0	
	Bio						0,110407
		Lulus	33	31	2	0,329846	
		Tidak Lulus	1	0	1	0	
	Sos						0
		Lulus	34	31	3	0,430552	
		Tidak Lulus	0	0	0	0	
	Eko						0,110407
		Lulus	33	31	2	0,329846	
		Tidak Lulus	1	0	1	0	
	Geo						0
		Lulus	34	31	3	0,430552	
		Tidak Lulus	0	0	0	0	
	Vrb						0,041833
		Lulus	24	23	1	0,249882	
		Tidak Lulus	10	8	2	0,721928	
	Lgk						0,110407
		Lulus	33	31	2	0,329846	
		Tidak Lulus	1	0	1	0	
	Art						0,430552
		Lulus	31	31	0	0	
		Tidak Lulus	3	0	3	0	

Gambar 17. Perhitungan Node 1.1

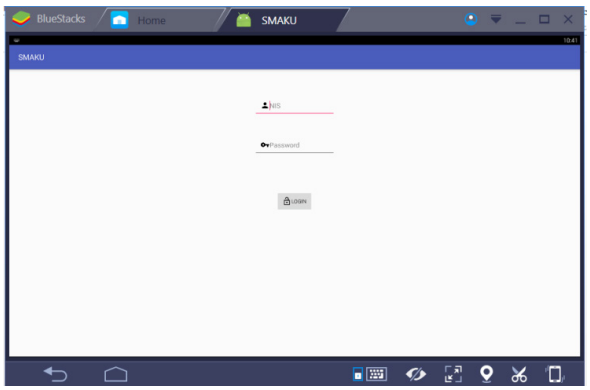
semua lulus. Sehingga didapatkan pohon keputusan seperti pada Gambar 18.



Gambar 18. Pohon keputusan penjurusan

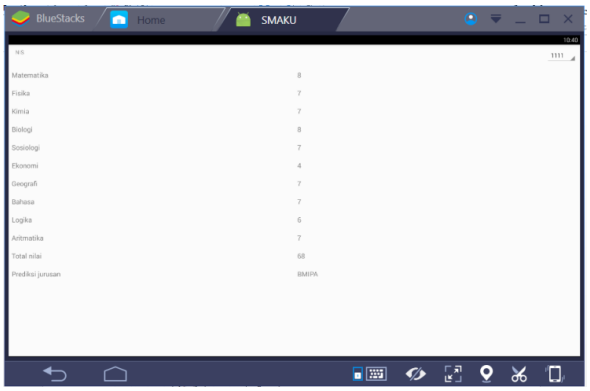
Pada gambar 18 didapatkan hasil dari algoritma *K-Means Clustering* dan C4.5 sehingga dapat diambil kesimpulan apa saja aspek yang akan digunakan dalam penentuan penjurusan IPA maupun IPS.

Pada gambar 19, gambar 20 dan gambar 21 di paparkan hasil pembuatan aplikasi dimana siswa baru akan memulai *login*. Siswa baru akan memasukkan username dan password yang sudah diberikan saat mereka melakukan pendaftaran. Setelah *login* berhasil, siswa dapat memulai sesi 1. Ujian sesi 1 ini berisi 50 butir soal dengan waktu yang diberikan selama 2jam 30menit. Setelah siswa sudah mengerjakan sesi 1, siswa di perbolehkan beristirahat dan lanjut sesi 2. Setelah siswa selesai mengerjakan sesi 1 dan sesi 2, siswa dapat melihat nilai dari apa yang sudah dikerjakan, tetapi siswa tersebut tidak dapat melakukan ujian ulang.



Gambar 19. Tampilan login siswa baru

Gambar 20. Tampilan menu awal siswa baru



Gambar 21. Tampilan hasil ujian

IV. SIMPULAN

Berdasarkan penelitian ini terdapat kesimpulan yang didapat yaitu sebagai berikut: (1) Pada penelitian ini terdapat 559 data akademis siswa yang dapat dipakai untuk melakukan pengujian; (2) Menggunakan 50 siswa untuk menguji dengan algoritma K-Means *Clustering* dan mendapatkan hasil akhir setelah melakukan 8 kali iterasi; dan (3) Mendapatkan kriteria pengambilan keputusan jurusan IPA dari hasil pohon keputusan C4.5.

V. DAFTAR RUJUKAN

[1] J. O. Ong. Implementasi Algoritma K Means Clustering Untuk Menentukan Strategi Marketing President University. Jurnal Ilmiah Teknik Industri, Vol. 12, No. 1, ISSN: 1412-6869. Hlm 65, 2013.

[2] Y. S. Nugroho & S. N. Haryanti. Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali. Jurnal Ilmu Komputer dan Informatika, Vol. 1, No. 1, ISSN: 2477-698X. Hlm 47, 2015.

[3] Y. S. Nugroho. Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta. Prosiding Seminar Nasional Aplikasi Sains & Teknologi, ISSN: 11979-911X. Hlm 59, 2014.

[4] D. H. Kamagi & S. Hansun. Implementasi Data mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. ULTIMATICS, Vol. VI, No. 1, ISSN: 2085-4552. Hlm 79, 2014.

[5] I. S. Sianturi. Sistem Pendukung Keputusan Untuk Menentukan Pemilihan Jurusan Siswa Dengan Menggunakan Metode Weighted Product. Jurnal Informasi dan Teknologi Ilmiah, Vol. 1, No. 1, ISSN: 2339-210X. hlm 67, 2013.

[6] Y. Bassil. A Simulation Model for the Waterfall Software Development Life Cycle, Internaional Journal of Engineering & Technology(iJET), Vol. 2,hal 75, 2015.

[7] G. Booch, R. A. Maksimchuk, M. W. Engle, B.J. Young, J. Conallen & K. A. Houston, Object Oriented Analysis and Design with Applications. United States: Addison-Wesley. ISBN: 978-0201895513. Hlm 431, 2007.

[8] D. T. Larose. Discovering Knowledge in Data: An Introduction to Data mining. Canada: Wiley Publishing. Hal 451, 2005.

[9] J. Han. Data mining: Concepts and Techniques, USA: Morgan Kaufmann Publishers. 2012.

[10] Kusrini, Luthfi & E. Taufiq. Algoritma Data mining, Yogyakarta: Penerbit Andi. Hlm 157, 2009.