SIGMA
JURNAL TEKNOLOGI PELITA BANGSA
Vol. 7 No. 1 Maret 2016 ISSN: 2407-3903
Diterima, 16 Februari 2016 | Direvisi, 18 Februari 2016 | Dipublikasikan, 28 Maret 2016

PREDIKSI KELULUSAN TEPAT WAKTU PADA KAMPUS STMIK MIC CIKARANG DENGAN ALGORITMA K-MEANS CLUSTERING

Bei Harira Irawan

Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer
MIC CIKARANG
beiharira@gmail.com

Abstrak

Perguruan tinggi adalah tempat bagi para mahasiswa untuk menimba ilmu sebelum terjun kedalam dunia persaingan kerja. Jumlah mahasiswa yang lulus tepat waktu menjadi indikator keberhasilan dari sebuah perguruan tinggi baik negeri dan swasta. Penelitian dalam hal memprediksi kelulusan mahasiswa dan siswa telah banyak dilakukan. Dalam penelitian ini dilakukan pengukuran metode data mining yaitu menggunakan algoritma K-Means Clustering dengan aplikasi WEKA 3.6.11 dan SPSS 19.0 menggunakan pada data Kampus STMIK MIC CIKARANG dari tahun 2006-2010.

Penelitian dilakukan untuk memprediksi kelulusan menggunakan teknik data mining metode algoritma K-Means Clustering yaitu dengan mengukur jarak terdekat dengan titik pusat cluster. WEKA menghasilkan 153 di cluster pertama, 311 cluster kedua dan 328 di cluster ketiga sedangkan SPSS menghasilkan 107 di cluster pertama, 412 cluster kedua dan 273 cluster ketiga. WEKA dan SPSS mengcluster kelulusan tepat waktu di cluster 3 dengan lama studi kurang lebih 4 tahun dengan komposisi terbanyak adalah mahasiswa dari Bekasi dengan jurusan paling banyak TI, berstatus lulus (LS), jenis kelamin laki-laki dengan jurusan asal SMA/SMK adalah IPS.

Kata Kunci : *K-Means Clustering*, prediksi kelulusan, WEKA, SPSS.

Abstract

College is a place for students to gain knowledge before plunging into the world of job competition. The number of students who graduate on time an indicator of the success of a universities both public and private. Research in terms of predicting graduation and students have been carried out. In this study measured data mining methods that use the K-Means Clustering algorithms with Weka 6.3.11 application and SPSS 19.0 using data Campus STMIK MIC CIKARANG from 2006-2010.

The study was conducted to predict graduation using data mining techniques method K-Means Clustering algorithm that is by measuring the closest distance to the cluster center point. WEKA generate 153 in the first cluster, the second cluster 311 and 328 in third cluster, while SPSS produces 107 in the first cluster, cluster 412 second and 273 third cluster. WEKA and SPSS mengcluster timely graduation in cluster 3 with a long study of approximately 4 years with the highest composition is a student from Bekasi with most IT departments, the status of pass to the male gender with the Department origin SMA/SMK is IPS.

Key words *— Prediction graduation, K-Means Clustering, graduate on time, WEKA, SPSS.*

1. Pendahuluan

Kampus STMIK MIC CIKARANG adalah sebuah lembaga pendidikan yang memiliki ijin menyelenggarakan Program Studi Strata Satu (S1) Teknik Informatika dan Sistem Informasi semenjak

tahun 2004. Peningkatan jumlah mahasiswa terlihat dalam kurun waktu 10 tahun terakhir. Permasalahan yang sering terjadi adalah masih banyaknya jumlah mahasiswa yang lulus dengan lama studi melampaui waktu yang telah ditetapkan. Kriteria kelulusan tepat waktu yang baik adalah mahasiswa sudah mengampu 150 SKS yang ditempuh dalam 8 semester dengan toleransi cuti maksimal 2 semester. Sehingga apabila diukur kurang lebih 4 tahun normalnya. Hal ini seringkali menjadi kendala bagi pihak kampus terutama pada saat pertanyaan Akreditasi muncul berkaitan dengan jumlah output atau lulusan yang sudah diluluskan.

Penulis menganalisa juga seberapa besar pengaruh dari latar belakang SMA para mahasiswa tersebut manjadi faktor penyebab tingginya angka retensi seperti latar belakang lulusan IPA atau IPS dalam memilih Program Studi. Selain itu hasil prediksi dapat menjadi rekomendasi untuk para dosen pembimbing akademik (Dosen PA), Kaprodi dan dosen pengajar agar kedepannya menjadi lebih intens membimbing, memonitoring dan memotivasi mahasiswa agar tingkat retensi berkurang. Tingkat *drop out* yang tinggi (mahasiswa yang tidak jelas statusnya) menyebabkan kampus kehilangan pendapatan, serta kelulusan yang tidak tepat waktu menjadi penyebab turunnya reputasi kampus di mata stakeholder (Delen, 2010). Beberapa hal yang sudah selayaknya dilakukan untuk memacu mahasiswa yang beresiko dengan mekanisme pendukung seperti orientasi, menasehati, monitoring, motivasi dan lain-lain dapat digunakan untuk meningkatkan kegigihan mahasiswa agar meningkatkan tingkat kelulusan. Tugas prediksi dapat dianggap sebagai menjadi dua kelas yaitu berarti "sukses" yakni mahasiswa yang lulus tepat waktu dan berarti "gagal" bagi mahasiswa yang lulus terlambat.

Penelitian yang sudah pernah dilakukan yang berhubungan dengan penelitian ini adalah:

1) Penelitian Qudril, M. N., & Kalyankar, N. V. (2010) yang memprediksi tingkat *drop out* siswa menggunakan metode *Decission Tree* sebagai pilihan yang terbaik dengan hasil bahwa faktor pendapatan orang tua menjadi faktor utama dalam kasus *drop out* siswa.
2) Karamouzis, T. S., & Vrettos, A. (2008) melakukan penelitian tentang prediksi kelulusan menggunakan *Artificial Neural Networks* (ANN) dengan 1.407 data sampel menggunakan 12 parameter. Hasilnya dari 1.100 data training diperoleh persentase kesuksesan 86.04% dan 307 data testing diperoleh presentase kesuksesan 70.27%.
3) Suhartinah, S. M., & Ernastuti. (2010) memprediksi mahasiswa yang lulus atau tidak lulus sesuai dengan waktu studi dengan membandingkan menggunakan algoritma *Naïve Bayes* dan *C4.5* dengan hasil akurasi ketepatan 80,85% prediksi menggunakan *Naïve Bayes* dan akurasi ketepatan 85,70% prediksi menggunakan *C4.5*.
4) Penelitian yang dilakukan oleh Kikie Riesky dan M. Akbar pada penerapan data mining untuk mengolah informasi konsentrasi keahlian dengan metode Clustering pada Universitas Bina Darma menggunakan metode Algoritma K-Means. Cluster yang ditentukan sebanyak 5 cluster. Dari data mahasiswa mempunyai data 697 record, Tabel KHS mempunyai 20,646 record sedangkan untuk tabel Matkul mempunyai 365 record selama 2 tahun. Dari semua atribut yang ada pada tabel Matkul dan tabel KHS terdapat 4 atribut yang digunakan dalam proses knowledge discovery in databases (KDD) yaitu kd_matkul (tabel KHS), nilai (tabel KHS), kd_matkul (tabel matkul), matkul (tabel matkul). Dari hasil penelitian disimpulkan bahwa mahasiswa yang mengambil konsentrasi jurusan didapat berdasarkan nilai dan minat mahasiswa. Berdasarkan nilai yaitu nilai diantara 60-70 dapat mengambil konsentrasi jurusan *database management system*, nilai diantara 70-80 dapat mengambil konsentrasi jurusan *software engineering* sedangkan nilai diantara 80-90 dapat mengambil konsentrasi jurusan *IT infrastructure*. Dari grafik juga disimpulkan bahwa informasi prediksi konsentrasi keahlian yang paling banyak diambil mahasiswa adalah konsentrasi jurusan *database management system* sebanyak 80%.
5) Penelitian yang dilakukan oleh Lillyan Hadjaratie pada prediksi dan pemetaan data mahasiswa Fakultas Teknik Universitas Negeri Gorontalo menggunakan pendekatan data mining menggunakan 3 metode yaitu *Decission Tree, Artificial Neural Network* dan *K-Nearest Neighbour*. Pada ketiga metode klasifikasi didasarkan pada IPK mahasiswa. Pada metode *Decission Tree*, klasifikasi data mahasiswa aktif berdasarkan IPK dengan metode *Decision Tree* menghasilkan 7 klasifikasi dengan karakteristik yang berbeda. Hasilnya didapat bahwa mahasiswa dengan IPK rendah didapat pada klasifikasi ke-6 yaitu sebanyak 275 mahasiswa dengan parameter mahasiswa dengan Jurusan Elektro, Arsitektur dan Industri. IPK sedang pada klasifikasi ke-4 sebanyak 171 mahasiswa dengan parameter mahasiswa dengan Jurusan Informatika, Jenis Kelamin Laki-Laki dan IPK tinggi pada klasifikasi ke-7 sebanyak 10 mahasiswa dengan parameter mahasiswa dengan Jurusan Kriya.

Dari beberapa jurnal dan penelitian yang penulis gunakan sebagai *related work* sebagai *study*

literature dan penelusuran ilmiah tersebut diatas maka yang dapat dijadikan sebagai identifikasi masalah dalam penelitian ini adalah :

a. Bagaimana mengolah data kampus agar bisa mengetahui kelompok kategori mana saja potensi mahasiswa yang kemungkinan akan lulus dengan waktu lama sehingga dapat menjadi acuan bagi calon mahasiswa lulusan SMA/SMK dalam memilih Program Studi agar kelak di kemudian hari dapat lulus tepat waktu.
b. Sebagai alat peringatan dini (*early warning*) bagi mahasiswa tertentu yang berdasarkan hasil prediksi dinyatakan berpotensi lulus dengan melampaui ketentuan lama studi.
c. Dapat dijadikan rekomendasi bagi Dosen Pembimbing Akademik maupun Kaprodi untuk memantau dan melihat perkembangan mahasiswa yang memiliki potensi retensi sehingga bisa diarahkan agar potensi tersebut bisa diminimalisir.

Berdasarkan latar belakang dan indentifikasi masalah diatas, maka dilakukanlah perumusan masalah sebagai berikut:

Bagaimana penerapan metode *K-Means Clustering* untuk memprediksi kelulusan tepat waktu pada Kampus STMIK MIC CIKARANG?

2. Teori

2.1. Landasan Teori

Han & Kamber (2001) mendefinisikan data mining sebagai *The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner*. Istilah data mining kadang disebut juga *Knowledge Discovery*. Inti dari data mining adalah kegiatan penggalian pengetahuan data. Pengertian dari istilah lain yang hampir mirip dengan data mining adalah *knowledge discovery* dan *pattern recognition*. *Knowledge discovery* adalah menemukan pengetahuan dari bongkahan data yang masih tersembunyi sedang *pattern recognition* adalah pengenalan pola. Pengetahuan yang digali masih berbentuk pola-pola yang mungkin masih perlu digali dalam bongkahan data.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1) Deskripsi
   Menggambarkan sekumpulan data secara ringkas. Data yang digambarkan berupa:
   a. Deskripsi grafis : diagram titik, histogram
   b. Deskripsi lokasi : *mean* (rata-rata), *median* (nilai tengah), *modus*, *kuartil*, *persentil*
   c. Deskripsi keberagaman : *range* (rentang), *varians* dan standar deviasi
2) Estimasi
   Memperkirakan suatu hal dari sejumlah sampel yang kita miliki (yang tidak kita ketahui), Estimasi hampir sama dengan klasifikasi, kecuali variable target. Estimasi lebih kearah numerik dari pada kearah kategori.
3) Prediksi
   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa datang (memperkirakan hal yang belum terjadi). Kita bisa menunggu hingga hal itu terjadi untuk membuktikan seberapa tepat prediksi kita.
4) Klasifikasi
   Kegiatan menggolongkan dengan menggunakan data historis (sebagai data yang digunakan untuk latihan dan sebagai pengalaman). Dalam klasifikasi terdapat variabel prediktor dan target variabel.
5) Pengklasteran (*Clustering*)
   Pengklasteran merupakan pengelompokan *record*, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Klaster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainya dan memiliki ketidakmiripan dengan *record-record* dalam klaster.
6) Asosiasi
   Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang pasar.

K-Means Clustering merupakan salah satu metode data *clustering* non-hirarki yang mengelompokan data dalam bentuk satu atau lebih cluster/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokan dalam satu cluster/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan cluster/kelompok yang lain sehingga data yang berada dalam satu cluster/kelompok memiliki tingkat variasi yang kecil.

K-Means pertama kali dipublikasikan oleh Stuart Lloyd pada tahun 1984 dan merupakan algoritma clustering yang banyak digunakan. K-Means bekerja dengan mensegmentasi objek yang ada kedalam kelompok atau yang disebut dengan segmen sehingga objek yang berada dalam masing-masing kelompok lebih serupa satu sama lain dibandingkan dengan objek dalam kelompok yang berbeda. Algoritma Clustering adalah meletakkan nilai yang serupa dalam satu segmen, dan meletakkan nilai yang berbeda dalam cluster yang berbeda (Wu & Kumar, 2009). K-Means memisahkan data dengan optimal dengan perulangan yang memaksimalkan hasil dari partisi hingga tidak ada perubahan data dalam setiap

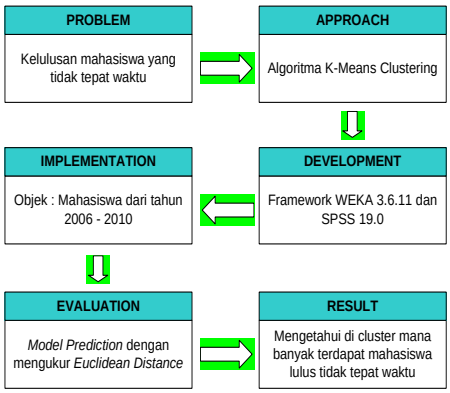
segmentasi. K-Means bekerja dengan pendekatan Top-Down karena memulai dengan segmentasi yang sudah ditentukan terlebih dahulu (Myatt, 2007). Sehingga hasil data sebuah segmen tidak mungkin tercampur antara satu segmen dengan segmen lainnya (Xu & Wunsch II, 2009). Pendekatan ini juga mempercepat proses komputasi untuk data dalam jumlah besar.

Cluster Analysis merupakan salah satu metode objek mining yang bersifat tanpa latihan (*unsupervised analysis*), sedangkan *K-Means Cluster Analysis* merupakan salah satu metode cluster analisis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih cluster atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam cluster yang lain. Tujuan pengelompokan adalah untuk meminimalkan objective function yang di set dalam proses clustering, yang pada dasarnya berusaha untuk meminimalkan variasi dalam satu cluster dan memaksimalkan variasi antar cluster.

2.2. Model Kerangka Pemikiran

Adapun tujuan yang ingin dicapai pada penulisan ini adalah sebagai berikut:

a. Mengetahui permasalahan yang terjadi mengapa masih banyak jumlah lulusan mahasiswa yang tidak tepat waktu (lebih dari 4 tahun).
b. Sebagai alat peringatan dini (*early warning*) bagi mahasiswa tertentu yang berdasarkan hasil prediksi dinyatakan berpotensi lulus dengan melampaui ketentuan lama studi.
c. Menghasilkan informasi parameter-parameter apa saja model mahasiswa yang sering tidak lulus tepat waktu.



Gambar 1. Kerangka Pemikiran

Adapun manfaat yang diharapkan dapat diperoleh dari penulisan ini adalah sebagai berikut:

a. Dengan adanya penerapan data mining dengan menentukan Cluster untuk memprediksi kelulusan tepat waktu kampus diharapkan dapat membantu memberikan gambaran dan informasi yang mendukung untuk mengambil keputusan (bagi manajemen dan lembaga penyelenggara pendidikan) yang tepat dalam melihat potensi-potensi kelulusan sehingga tingkat retensi bisa ditekan.
b. Dapat diarahkan kemana penjurusan yang tepat bagi mereka diantaranya dengan menilik latar belakang sekolah asal SMA/SMK bagi calon mahasiswa, misal bagi lulusan dengan jurusan IPA/Kejuruan yang linear dapat diarahkan ke program Teknik Informatika (TI) dan jurusan IPS untuk progran Sistem Informasi (SI).
c. Dapat dijadikan rekomendasi bagi Dosen Pembimbing Akademik maupun Kaprodi untuk memantau dan melihat perkembangan mahasiswa yang memiliki potensi retensi sehingga bisa diarahkan agar potensi tersebut bisa diminimalisir.

3. Hasil Pembahasan dan Analisa

3.1. Proses *Knowledge Discovery in Databases*

Data awal terdiri dari tabel mahasiswa, tabel data_pribadi dan tabel KHS mulai dari tahun 2006 sampai tahun 2010. Semua data di gabungkan menjadi satu yang terdiri dari beberapa field antara lain NIM, nama, tempat lahir, program, jurusan, status, jenis kelamin dan jurusan SMA/SMK. Jumlah data sebelum di lakukan proses *pre-processing/cleaning* adalah sebanyak 1.519 record data.

Semua data tersebut digabung menjadi satu melalui proses menggunakan metode Knowledge Diccovery in Databases (KDD) yaitu proses untuk mengekstrak pengetahuan apa yang dianggap sesuai dengan spesifikasi ukuran dan batas, menggunakan database bersama dengan pre-processing yang diperlukan.

Data ditransformasi sehingga sesuai dengan spesifikasi yang akan digunakan dalam penelitian ini yang terdiri dari NIM mahasiswa dengan digit NIM sesuai ketentuan kampus, nama mahasiswa sesuai dengan nama pada Ijazah SMA/SMK pada saat mendaftar, tempat lahir, program studi pilihan mahasiswa, jurusan, status terakhir mahasiswa sampai penelitian ini disusun, jenis kelamin mahasiswa dan jurusan SMA/SMK asal.

Tabel 1. Dataset awal sebelum dilakukan proses KDD (masih banyak terjadi redundansi data)

NIM	Nama	Tempat Lahir	Program	Jurusan	Status	JK	Jurusan Asal
0606040021	ABDUL	Leuwimunding	S1	TI	LS	L	IPS
0606040003	ADE SARIF MUDAIN		S1	TI	NA	L	IPS
0606040006	AGUS MUSLIMIN	Garut	S1	TI	NA	L	IPA
0604010004	Ahmad Abdulloh Attaimiyah	Brebes	D3	MI	NA	L	Elektro
0605010007	Ahmad Mustahal	Wonosobo	D3	MI	NA	L	IPS
0605010006	Ahyad	Bekasi	D3	MI	NA	L	IPS
0605010008	AHYAR	Sangia	D3	MI	LS	L	IPA
0606040004	ANDHI DJOKO PURNOMO	Semarang	S1	TI	NA	L	Mesin
0606040005	ANDRI PRASETYO	Wonogiri	S1	TI	LS	L	IPA
0604020001	ANISTIDA	Gunungkidul	D3	KA	LS	P	IPS
0605010009	ANISTIDA	Gunungkidul	D3	MI	LS	P	IPS
0605010010	Aproni	Brebes	D3	MI	PJ	L	IPS
0606030024	Aproni	Brebes	S1	SI	LS	L	IPS
0606040001	ARI SANTOSO	Ciamis	S1	TI	PJ	L	IPA
0606164027	ARI SANTOSO	Ciamis	S1	TI	LS	L	IPA

Dalam melakukan pre-processing/cleaning yaitu proses data mining yang bertujuan untuk menggabungkan data, membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak dan lainnya, maka penulis melakukan tahapan antara lain menggabungkan, membuang duplikasi data, menambahkan data pendukung dan memperbaiki kesalahan data. Teknik ini dilakukan menggunakan Microsoft Excel 2007. Berikut merupakan penjelasan dari proses-proses tersebut:

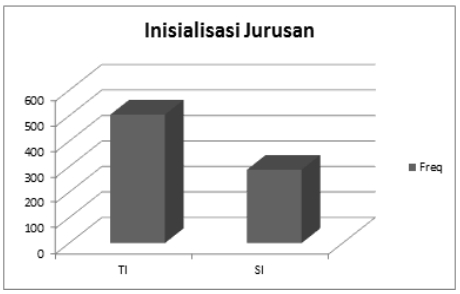
1) Menggabungkan data, tahap menggabungkan data adalah menggabungkan data dari database mahasiswa, data_pribadi dan KHS mulai dari tahun 2006 sampai 2010 yaitu sebanyak 1.519 record.
2) Membuang duplikasi data, tahap ini adalah merupakan salah satu proses cleaning data dengan cara membuang data yang sama yang tidak diperlukan dan hanya memisahkan mahasiswa jurusan SI dan TI serta memisahkan status mahasiswa yang AT (aktif), CT (cuti), NA (not available) dan LS (lulus) saja.
3) Memperbaiki kesalahan data, tahap terakhir ini adalah mengecek kembali data yang sudah di cleaning dan digabung bilamana ada kesalahan dalam penulisan, data masih kosong ataupun tata letak maka penulis memperbaiknya.
4) Dataset akan disimpan dalam bentuk file .csv dan akan diukur menggunakan aplikasi WEKA 3.6.11 dengan menentukan 3 buah klaster.

Data yang tidak ada nilainya dan duplikasi dihilangkan atau di validasi sehingga dari 1.519 record diperoleh 792 record. Dari data tersebut jumlah data mahasiswa dengan status Aktif (AT) sebanyak 95 data, status Cuti (CT) sebanyak 39 data, status Non Aktif (NA) sebanyak 325 data dan yang Lulus (LS) sebanyak 333 data.

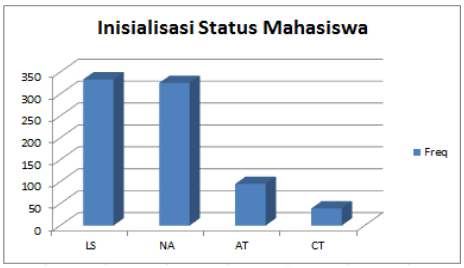
Tabel 2. Dataset setelah dilakukan teknik validasi dan diskretisasi

Nama	Tempat Lahir	Jurusan	Status	Tahun Lulus	Jenis Kelamin	Lama Studi	Jurusan Asal
ABDUL	Leuwimunding	TI	LS	2010	L	4	IPS
ADE SARIF MUDAIN	Keburnen	TI	NA	2006	L	0	IPS
AGUS MUSLIMIN	Garut	TI	NA	2006	L	0	IPA
ANDHI DJOKO PURNOMO	Semarang	TI	NA	2006	L	0	Mesin
ANDRI PRASETYO	Wonogiri	TI	LS	2011	L	5	IPA
Aproni	Brebes	SI	LS	2011	L	5	IPS
ARI SANTOSO	Ciamis	TI	LS	2012	L	6	IPA
AZMIL FAUZI	Bekasi	SI	NA	2006	L	0	IPA
DEDI SUPENO	Kertasemaya	TI	LS	2011	L	5	IPS
ELI SUNARYA	Bekasi	TI	NA	2006	L	0	IPS
ERIX SUTRIADI	Bekasi	TI	NA	2006	L	0	IPS
IN WINARDI	Sagalaherang	TI	NA	2006	L	0	Elektro
Irman Kadori	Palimanan	TI	LS	2010	L	4	IPA
Jaenal Nurohman	Banten	TI	LS	2010	L	4	IPA
JAFAR AMIRUDIN	Ngawi	TI	LS	2012	L	6	IPS
JONES DAVID R. SIMATUPANG	Siborongborong	TI	LS	2013	L	7	IPA
JUWENI RACHMAT	Pandeglang	SI	NA	2006	L	0	Mesin
M. ISROFIL	Pekalongan	TI	LS	2011	L	5	IPA
Mohamad Arwani	Bulumanis Kidul	TI	NA	2006	L	0	IPA

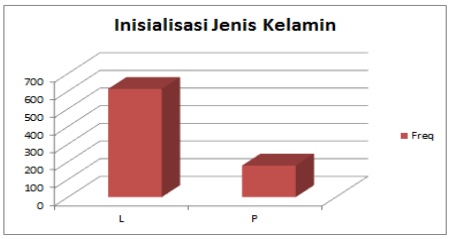
Agar data di atas dapat diolah dengan menggunakan metode *K-Means clustering*, maka data yang berjenis nominal (diskrit) seperti tempat lahir, jurusan, status, jenis kelamin dan lama studi harus di inisialisasikan terlebih dahulu dalam bentuk angka. Berikut grafik hasil inisialisasi seperti pada gambar berikut.



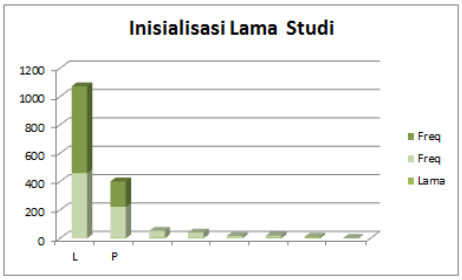
Gambar 2. Hasil inisialisasi Jurusan



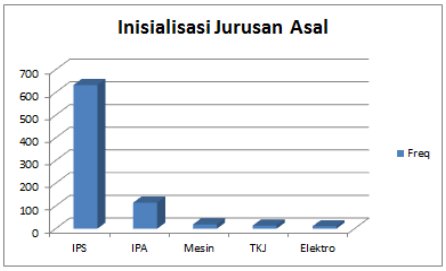
Gambar 3. Hasil inisialisasi Status Mahasiswa



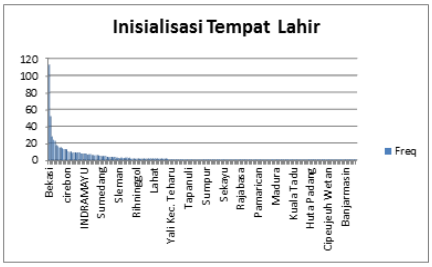
Gambar 4. Hasil inisialisasi Jenis Kelamin



Gambar 5. Hasil inisialisasi Lama Studi

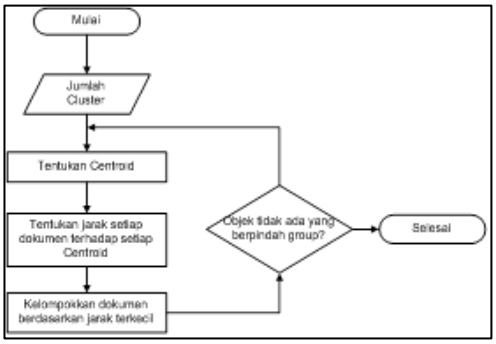


Gambar 6. Hasil inisialisasi Jurusan Asal



Gambar 7. Hasil inisialisasi Tempat Lahir

Setelah semua data ditransformasikan kedalam bentuk angka dan di inisialisasi, maka data–data tersebut telah dapat dikelompokan dengan menggunakan algoritma *K-Means Clustering*. Berikut flowchart dari langkah-langkah K-Means.



Gambar 8. Flowchart *K-Means Clustering*

Untuk dapat melakukan pengelompokan data – data tersebut menjadi beberapa cluster perlu dilakukan beberapa langkah, yaitu :

1) Tentukan jumlah cluster yang diinginkan. Dalam penelitian ini data akan dikelompokan menjadi tiga cluster.
2) Tentukan titik pusat awal dari setiap cluster. Dalam penelitian ini pusat awal ditentukan secara acak untuk mendapatkan titik pusat dari setiap cluster. Data acak diambil dari record ke 13, 353 dan 786. Tabel 3 menunjukkan penentuan titik awal pusat cluster.

Tabel 3. Titik awal pusat cluster acak

Nama	Tempat Lahir	Initial TTL	Jurusan	Initial Jurusan	Status	Initial Status	Jenis Kelamin	Initial Jenkel	Jurusan Asal	Initial Jurusan Asal	Lama Studi	Record ke
Imam Kzdori	Palmaran	66	TI	1	LS	1	L	1	IPA	2	4	13
AHALDIN JAMIL	Bekasi	1	TI	1	MA	2	L	1	IPS	1	0	353
TUTI HERWIG	Pemalang	12	TI	1	KT	3	P	2	IPS	1	0	786

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, dengan teori jarak Euclidean yang dirumuskan sebagai berikut.

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

dimana:

$D(i,j)$ = Jarak data ke i ke pusat cluster j

X_{ki} = Data ke i pada atribut data ke k

X_{kj} = Titik pusat ke j pada atribut ke k

4) Dari perhitungan jarak dengan rumus diatas maka didapatkan hasil pengukuran untuk setiap titik pusat cluster sebagai berikut.

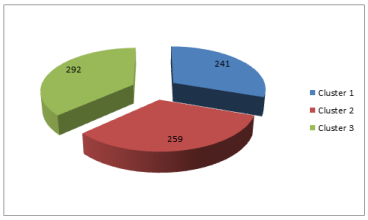
Initial TTL	Initial Jurusan	Initial Status	Initial Jenkel	Initial Jur. Asal	Lama Studi	Cluster 1	Cluster 2	Cluster 3	Pusat Cluster
66	1	1	1	2	4	0	65	54	1
1	1	2	1	1	0	65	0	11,09	2
12	1	3	2	1	0	54	11,09	0	3

Gambar 9. Hasil penentuan titik pusat cluster

3.2. Evaluasi Hasil Pengukuran

Setelah semua dataset dihitung jaraknya dengan titik pusat cluster maka bisa diketahui masing-masing masuk ke dalam cluster yang mana dengan menghitung jarak terdekat dengan titik pusat cluster. Dari setiap dataset di setiap cluster kita bisa hitung frekuensi terbesar baik dari tempat lahir, jurusan, status, jenis kelamin, jurusan asal SMA dan lama lulusan. Dari data tersebut kemudian bisa kita

olah dengan mengambil nilai max, min, standar deviasi dan average. Dari data 3 cluster didapatkan nilai sebanyak 241 pada cluster pertama, 259 pada cluster kedua dan 292 pada cluster ketiga



Gambar 10. Nilai pada setiap cluster

Secara acak akan dipilih data sejumlah dengan nilai k yaitu 3. Maka dipilih data pada record ke 13, 353 dan 786 sebagai pusat cluster. Nilai Euclidean diperhitungkan untuk setiap data kedalam lima titik tersebut.
Pehitungan nilai D dari data-1 ke kelima pusat segmentasi adalah sebagai berikut:

$$D(1,1) = \sqrt{(66-66)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (2-2)^2 + (4-4)^2} = 0$$

$$D(1,2) = \sqrt{(66-1)^2 + (1-1)^2 + (1-2)^2 + (1-1)^2 + (2-1)^2 + (4-0)^2} = 65{,}14$$

$$D(1,3) = \sqrt{(66-12)^2 + (1-1)^2 + (1-3)^2 + (1-2)^2 + (2-1)^2 + (4-0)^2} = 54{,}20$$

Pehitungan nilai D dari data-2 ke kelima pusat segmentasi adalah sebagai berikut:

$$D(2,1) = \sqrt{(1-66)^2 + (1-1)^2 + (2-1)^2 + (1-1)^2 + (1-2)^2 + (0-4)^2} = 65{,}14$$

$$D(2,2) = \sqrt{(1-1)^2 + (1-1)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2} = 0$$

$$D(2,3) = \sqrt{(1-12)^2 + (1-1)^2 + (2-3)^2 + (1-2)^2 + (1-1)^2 + (0-0)^2} = 11{,}09$$

Pehitungan nilai D dari data-3 ke kelima pusat segmentasi adalah sebagai berikut:

$$D(3,1) = \sqrt{(12-66)^2 + (1-1)^2 + (3-1)^2 + (2-1)^2 + (1-2)^2 + (0-4)^2} = 54{,}20$$

$$D(3,2) = \sqrt{(12-1)^2 + (1-1)^2 + (3-2)^2 + (2-1)^2 + (1-1)^2 + (0-0)^2} = 11{,}09$$

$$D(3,3) = \sqrt{(12-12)^2 + (1-1)^2 + (3-3)^2 + (2-2)^2 + (1-1)^2 + (0-0)^2} = 0$$

Tabel 4. Nilai D pada data acak

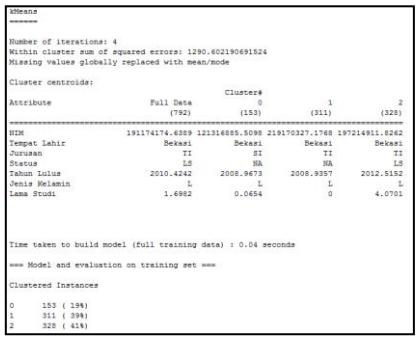
Data	Centroid		
Cluster	13	353	786
1	0	65,14	54,2
2	65,14	0	11,09
3	54,2	11,09	0

3.3. Pengukuran Menggunakan WEKA 3.6.11

Data yang digunakan dalam penelitian ini masih sama dengan data yang digunakan dengan metode *K-Means cluster* yaitu berasal dari database Kampus STMIK MIC CIKARANG yang terdiri dari tabel mahasiswa, tabel data_pribadi dan tabel KHS mulai dari tahun 2006 sampai tahun 2010. Semua data di gabungkan menjadi satu yang terdiri dari beberapa field antara lain nama, tempat lahir, jurusan, status, jenis kelamin dan lama studi. Jumlah data set sebanyak 792 record.

Berbeda dengan metode teoritis, data yang digunakan pada metode aplikatif ini harus di konvert dahulu menjadi bentuk file yang bisa di proses oleh WEKA 3.6.11 yaitu file .csv dengan nama file data_mahasiswa.csv.

Bedasarkan hasil aplikasi WEKA 3.6.11 dalam pengujian penelitian ini untuk data kelulusan mahasiswa dengan metode Sample K-Means maka di dapati hasil seperti Gambar di bawah ini:



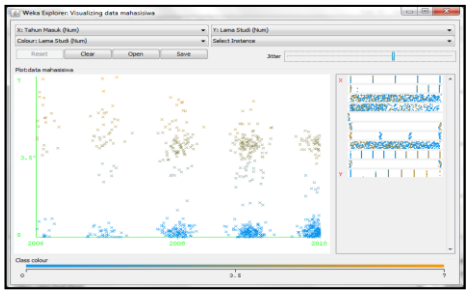
```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1290.602190691524
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                          (792)        (153)      (311)      (328)
=============================================================================
NIM                 191174174.6389 121316885.5098 219170327.1768 197214911.8262
Tempat Lahir                Bekasi     Bekasi     Bekasi     Bekasi
Jurusan                         TI         SI         TI         TI
Status                          LS         NA         NA         LS
Tahun Lulus            2010.4242  2008.9673  2008.9357  2012.5152
Jenis Kelamin                   L          L          L          L
Lama Studi                 1.6982     0.0654          0     4.0701



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       153 ( 19%)
1       311 ( 39%)
2       328 ( 41%)
```

Gambar 11. Hasil Sample K-Means Pada WEKA 3.6.11

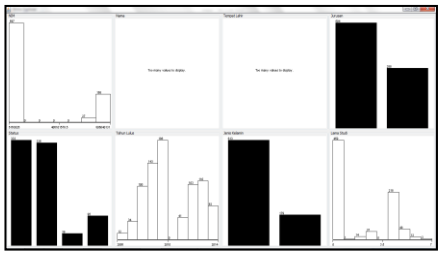
Dari hasil di atas dengan menggunakan WEKA 3.6.11 menggunakan Sample K-Means terlihat bahwa paling banyak data kelulusan tepat waktu di cluster 3 dengan lama studi 4.071 berjumlah 328 atau 41% dengan komposisi terbanyak adalah mahasiswa dari Bekasi dengan jurusan paling banyak TI, berstatus lulus dengan jenis kelamin laki-laki. Rata-rata lulusan adalah mahasiswa yang mendaftar pada tahun 2008 dan terprediksi lulus tepat waktu yaitu 2012.

Sedangkan pada cluster 1, dapat dilihat bahwa terdapat 153 atau 19% belum lulus meskipun ada nilai lama studi sekitar 0,0654 dan berstatus non aktif (NA). pada cluster 1 rata-rata tahun masuk di 2008 juga dengan jumlah terbesar dari daerah Bekasi jurusan TI dan jenis kelamin laki-laki.

Begitu juga pada cluster 2 sejumlah 311 atau 39% sama sekali belum ada yang lulus dan kemungkinan masih tetap aktif sampai tahun 2014 ini. Rata-rata tahun masuk 2009 padahal secara prediksi semestinya 2013 lalu sudah lulus. Jumlah ini di dominasi juga dari daerah Bekasi jurusan TI dan jenis kelamin laki-laki.



Gambar 12. Hasil sebaran cluster, x = tahun masuk, y = lama studi



Gambar 13. Visualisasi atribut keseluruhan

3.4. Pengukuran Menggunakan SPSS 19.0

Proses clustering yang dilakukan melalui 12 tahapan iterasi sampai mendapatkan konvergensi yang tepat. Dari Gambar 14 berikut disebutkan bahwa jarak minimum antar pusat cluster yang terjadi dari hasil iterasi adalah 5.010.

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	1.679	1.833	1.743
2	.274	.296	.455
3	.278	.089	.229
4	.197	.168	.355
5	.218	.011	.070
6	.063	.012	.009
7	.060	.011	.011
8	.042	.009	.007
9	.036	.006	.007
10	.018	.005	.000
11	.031	.008	.000
12	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 12. The minimum distance between initial centers is 5.010.

Gambar 14. Proses clustering dengan hasil iterasi 12 kali

Hasil Output Final Cluster Centers telah dihasilkan yang masih terkait dengan proses standardisasi data sebelumnya, yang mengacu pada z-score dengan ketentuan sebagai berikut:

- Nilai negatif (-) berarti data berada di bawah rata-rata total.
- Nilai positif (+) berarti data berada di atas rata-rata total

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(tempat_lahir)	2.21949	-.37282	-.30727
Zscore(jurusan)	.21534	-.09993	.06641
Zscore(status)	-.25049	.69820	-.95552
Zscore(jenis_kalamin)	.06290	-.05867	.06389
Zscore(lama_studi)	.23666	-.82240	1.14838

Gambar 15. Hasil Output Final Cluster Centers

Rumus yang digunakan untuk mengetahui rata-rata nilai masing-masing variabel dalam tiap cluster sebagai berikut:

$$\chi = \mu + Z.\sigma$$

Di mana:
X = Rata-rata sampel dalam cluster
μ = Rata-rata populasi
Z = Nilai standardisasi
σ = Standar Deviasi

Hitung rata-rata nilai lama studi dalam setiap cluster.

Cluster 1:
(rata-rata lama studi) + (0,2366 x Standar Deviasi rata-rata lama studi)
= 2,1869 + (0,2366 x 2,0472)
= 2,6713

Cluster 2:
(rata-rata lama studi) + (-0,8224 x Standar Deviasi rata-rata lama studi)
= 0 + (-0,8224 x 0)
= 0

Cluster 3:
(rata-rata lama studi) + (1,1483 x Standar Deviasi rata-rata lama studi)
= 4,0695 + (1,1483 x 0,8127)
= 5,0027

Dari hasil diatas dapat didefinisikan sebagai berikut:

Cluster 1
Dalam cluster 1 terlihat bahwa rata-rata status berada di bawah rata-rata, tempat lahir terlihat nilainya paling besar diatas rata-rata dibanding jurusan, jenis kelamin dan lama studi.

Dengan demikian dapat diduga bahwa pada cluster ini terdapat mahasiswa yang sudah lulus dengan lama studi rendah, banyak mahasiswa non aktif yang berasal dari kota tertentu yang jumlahnya banyak dan didominasi oleh mereka yang jurusan asal SMA/SMK dari IPS.

Cluster 2
Dalam cluster 2 terlihat bahwa rata-rata tempat_lahir, jurusan, jenis_kelamin dan lama studi berada di bawah rata-rata, sedang status berada diatas rata-rata. Dengan demikian dapat diduga bahwa pada cluster ini terdapat mahasiswa masih

aktif kuliah dan belum ada kelulusan didominasi oleh mereka yang jurusan asal SMA/SMK dari IPS.

Cluster 3

Dalam cluster 3 terlihat bahwa rata-rata tempat_lahir dan status berada di bawah rata-rata, sedang jurusan, jenis_kelamin dan lama_studi berada diatas rata-rata. Dengan demikian dapat diduga bahwa pada cluster ini terdapat kelompok mahasiswa yang sudah lulus dalam jumlah relatif banyak.

Selanjutnya untuk mengetahui jumlah anggota masing-masing cluster yang terbentuk dapat dilihat pada Tabel 5 berikut:

Tabel 5. Jumlah anggota tiap cluster

Number of Cases in each Cluster

Cluster	1	107.000
	2	412.000
	3	273.000
Valid		792.000
Missing		.000

Hasil perbandingan komparasi WEKA 3.6.11 dengan SPSS 19.0 dapat dilihat pada Tabel 6 berikut:

Tabel 6. Tabel perbandingan hasil pengukuran

			CLUSTER								
NO	PENGUKURAN	ITERASI	1	%	Lama Studi	2	%	Lama Studi	3	%	Lama Studi
1	WEKA 3.6.11	4 kali	153	19%	0,0654	311	39%	0	328	41%	4,0701
2	SPSS 19.0	12 kali	107	13%	2,6713	412	52%	0	273	34%	5,0027

Dari penelitian yang telah dilakukan untuk memprediksi kelulusan menggunakan teknik data mining metode algoritma *K-Means Clustering* yaitu dengan mengukur jarak terdekat dengan titik pusat cluster, hasilnya di dapat 3 buah cluster dari 792 dataset yang diukur. Pada cluster pertama terdapat 153 data yang jarak titik pusat clusternya dekat dengan titik pusat cluster pertama, cluster kedua terdapat 311 data yang jarak titik pusat clusternya dekat dengan titik pusat cluster kedua dan cluster ketiga terdapat 328 data yang jarak titik pusat clusternya dekat dengan titik pusat cluster ketiga. Pengukuran menggunakan metode simple K-Means pada aplikasi WEKA 3.6.11 menunjukkan paling banyak data kelulusan tepat waktu di cluster 3 dengan lama studi tepat 4 tahun berjumlah 328 atau 41% dengan komposisi terbanyak adalah mahasiswa dari Bekasi dengan jurusan paling banyak TI, berstatus lulus (LS) dengan jenis kelamin laki-laki. Rata-rata lulusan adalah mahasiswa yang mendaftar pada tahun 2008 dan terprediksi lulus tepat waktu yaitu 2012. Sedangkan pada cluster 1, terlihat masih ada nilai lama studi 0,00654 dan sekitar 19% mahasiswa yang seharusnya sudah lulus 4 tahun tetapi sampai lebih dari 4 tahun masih aktif, dan berstatus non aktif (NA). Pada cluster 1 rata-rata tahun masuk di 2008 dengan jumlah terbesar dari daerah Bekasi jurusan TI dan jenis kelamin laki-laki didominasi oleh mereka yang jurusan asal SMA/SMK dari IPS. Begitu juga pada cluster 2 sejumlah 123 atau 16% masih tetap aktif sampai tahun 2014 ini. Rata-rata tahun masuk 2009 padahal secara prediksi semestinya 2013 lalu sudah lulus.

Jumlah ini di dominasi juga dari daerah Bekasi jurusan TI dan jenis kelamin laki-laki. Pada cluster 2 sejumlah 311 atau 39% sama sekali belum ada yang lulus dan kemungkinan masih tetap aktif sampai tahun 2014 ini. Rata-rata tahun masuk 2009 padahal secara prediksi semestinya 2013 lalu sudah lulus. Jumlah ini di dominasi juga dari daerah Bekasi jurusan TI dan jenis kelamin laki-laki didominasi oleh mereka yang jurusan asal SMA/SMK dari IPS.

Dari hasil pengukuran SPSS 19.0 dapat dilihat bahwa paling banyak data kelulusan tepat waktu sama dengan WEKA yaitu di cluster 3 dengan lama studi 5 tahun berjumlah 273 atau 34%.

Dari hasil diatas menunjukkan bahwa jumlah mahasiswa yang non aktif (NA) menunjukkan angka terbesar yang rata-rata di dominasi oleh laki-laki yang sebagian besar berasal dari daerah Bekasi dengan jurusan TI didominasi oleh mereka yang jurusan asal SMA/SMK dari IPS.

4. Kesimpulan

Tujuan akhir dari pembuatan tulisan ini adalah membuat pemetaan faktor apa saja yang membuat faktor kelulusan menjadi lama pada program studi Sistem Informasi dan Teknik Informatika di institusi pendidikan perguruan tinggi dan dari hasil analisis yang dilakukan penulis, maka dapat diambil kesimpulan sebagai berikut :

1. Potensi kelulusan tidak tepat waktu di dominasi oleh mahasiswa yang berasal dari daerah Bekasi dan mengambil jurusan Teknik Informatika (TI) dengan jenis kelamin laki-laki dan jurusan asal SMA IPS. Hal ini dapat menjadi acuan bagi calon mahasiswa yang berasal dari daerah Bekasi terutama yang berjenis kelamin laki-laki dan jurusan asal IPS untuk diarahkan ke jurusan Sistem Informasi (SI).
2. Jumlah mahasiswa yang belum lulus di dominasi juga mahasiswa Not Available (NA) dengan jurusan Teknik Informatika (TI), ini menjadi pekerjaan rumah bagi Kaprodi TI untuk lebih intens membina mahasiswa di jurusannya agar potensi mahasiswa putus di tengah jalan bisa dikurangi.
3. Menjadi tolak ukur bagi dosen pembimbing akademik di jurusan TI untuk lebih membina dan memotivasi mahasiswa terutama yang berjenis kelamin laki-laki agar retensi bisa berkurang.

4. Menjadi alat deteksi dini (early warning) bagi manajemen dalam menyelenggarakan pendidikan.
5. Pada penelitian ini menunjukkan perlunya dan pentingnya peran aktif manajemen dalam memperbaiki tingkat kelulusan mahasiswa yang nantinya berimbas pada kesan di mata stake holder akan kualitas lulusan kampus. Data ini juga bisa dimanfaatkan sebagai peringatan dini (early warning) bagi mahasiswa tertentu, jurusan tertentu, yang berdasarkan hasil prediksi dinyatakan berpotensi lulus dengan melampaui ketentuan lama studi.

Daftar Pustaka

[1] Agusta Y. K-Means-Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: Jurnal Sistem dan Informatika (Februari 2007) Vol. 3: 47-60; 2007.

[2] Dawson, C. W. (2009). Projects in Computing and Information Systems a student's guide (Second Edition ed.). Harlow, UK: Addison-Wesley.

[3] Delen, Dursun. (2010). A Comparative Analysis of Machine Learning Techniques for Student Retention Management. Journal of Decision Support Systems 49 (2010) 498–506.

[4] Tahta Alfina, Budi Santosa, Ali Ridho Barakbah. 2012. Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data. Jurnal Teknik ITS Vol.I, ISSN: 2301-9271, Sept 2012.

[5] Fayyad, U. M. et al. (1996). From data mining to knowledge discovery: an overview. In Fayyad, U. M.et al (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

[6] Han, J., & Kamber, M. (2007). Data Mining Concepts and Techniques. San Fransisco: Mofgan Kaufan Publisher.

[7] Karamouzis, T. S., & Vrettos, A. (2008). An Artificial Neural Network for Predicting Student Graduation Outcomes. Preceeding of World Congress on Engineering and Computer Science,978-988-98671-02.

[8] Kikie Riesky Andini, M. Akbar, Helda Yudiastuti. 2013. Penerapan Data Mining Untuk Mengolah Informasi Konsentrasi Keahlian Dengan Metode Clustering Pada Universitas Bina Darma. Jurnal Ilmiah Vol.X No.10, November 2013:1-15.

[9] Larose, Daniel T. (2007). Data Mining Methods and Models. Wiley-Interscience, ISBN-13 978-0-471-66656-1.

[10] Lillyan Hadjaratie (2009). Prediksi dan Pemetaan Data Mahasiswa Fakultas Teknik Universitas Negeri Gorontalo Menggunakan pendekatan data mining. Information Technology Journal (Jurusan Teknik Informatika Universitas Negeri Gorontalo) 8, 8, 1256-1262.

[11] Murtaugh, P.A., L.D. Burns and J. Schuster, 1999 Predicting the retention of university students. Higher Education, 4: 355-357.

[12] Qudril, M. N., & Kalyankar, N. V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques. Global Journal of Computer Science and Technology, 2-4.

[13] Suhartinah & Ernastuti (2010). Graduation prediction of Gunadarma university students Using algorithm Naive Bayes and C4.5 algorithm.

Vercellis, C. (2009). Business Intelligent: Data Mining and Optimization for Decision Making. Southern Gate: John Willey & Sons Inc.