PENERAPAN *DATA MINING* DALAM MENGELOMPOKKAN DATA RIWAYAT AKADEMIK SEBELUM KULIAH DAN DATA KELULUSAN MAHASISWA MENGGUNAKAN METODE *AGGLOMERATIVE HIERARCHICAL CLUSTERING*

(*Implementation Of Data Mining In Grouping Academic History Data Before Students And Student Graduation Data Using Method Agglomerative Hierarchical Clustering***)**

Banu Harli Trimulya Suandi As*, Lisna Zahrotun
Dept Informatics Engineering, Ahmad Dahlan University
Jl. Ring Road Selatan, Tamanan, Banguntapan, Bantul Yogyakarta, INDONESIA
Email: banuharlitrimulya@gmail.com, lisna.zahrotun@tif.uad.ac.id

Abstract

The process of admitting new students at the Faculty of Industrial Technology, Ahmad Dahlan University, which has a very high number of students entering at the time of admission and graduating students who graduate on time is still low, causing an imbalance between the ratio of lecturers and students to be large. The number of students using campus facilities exceeds the capacity, and teaching and learning activities become ineffective. This study uses the hierarchical clustering method. The stages in this research are from Load Data, Data Cleaning, Data Transformation with One Hot Encoding method, Euclidean Distance, grouping Agglomerative Hierarchical Clustering. Results testing Cluster using the Silhouette Coefficient, was also carried out evaluation of patterns, and representation of knowledge. The study resulted in 158 recommended student data and all of them came from the island of Java and the average math score was >= 80 on the dataset Informatics, Industry and Electrical, and >= 67 for Chemistry. Obtained the recommended data with the number of data, respectively 43, 24, 19, and 72 data. The test method Silhouette Coefficient obtained very good results with the value Silhouette Coefficient according to the study program respectively of 0.868, 0.883, 0.879, and 0.873.

Keywords: *Agglomerative Hierarchical Clustering, Clustering, Student Data, Data mining, Silhouette Coefficient*

**Penulis Korespondensi*

1. PENDAHULUAN

Data akademik mahasiswa merupakan data yang dihimpun dari hasil kegiatan proses belajar mengajar selama mengikuti studi di suatu perguruan tinggi. Data tersebut antara lain: data pribadi mahasiswa, data rencana studi, dan data hasil studi (nilai dan indeks prestasi) [1]. Data akademik mahasiswa ini juga berupa data penerimaan mahasiswa baru yang rutin dilakukan seluruh perguruan tinggi di Indonesia pada setiap tahunnya.

Universitas Ahmad Dahlan merupakan salah satu perguruan tinggi swasta di Indonesia yang terletak di Yogyakarta, yang terdiri dari 11 fakultas dan 34 program studi. Fakultas Teknologi Indutri (FTI) merupakan salah satunya yang memiliki 4 program studi yaitu teknik informatika, teknik industri, teknik elektro, dan teknik kimia. Penerimaan mahasiswa baru (PMB) yang dilakukan oleh FTI pada tahun 2019 terdapat 3.198 pendaftar dengan jumlah mahasiswa yang diterima berjumlah 950 mahasiswa yang sudah registrasi. Tabel 1 menunjukkan jumlah data statistik kelulusan mahasiswa FTI yang menyelesaikan kuliah tepat waktu yang bersumber dari *website* fakultas [2].

TABEL I. MAHASISWA LULUS TEPAT WAKTU

No	Jumlah Mahasiswa	Tahun	Lulus Tepat Waktu	%
1	360	2012	14	4
2	340	2013	22	6
3	591	2014	55	9
4	837	2015	181	22

Dengan jumlah data pada Tabel I dapat dilihat dari tahun ke tahun memiliki perkembangan yang sangat cepat. Dengan jumlah mahasiswa yang masuk pada saat penerimaan mahasiswa baru yang sangat

tinggi dan kelulusan mahasiswa tepat waktu masih rendah, menyebabkan tidak seimbang menyebabkan rasio dosen dan mahasiswa menjadi besar. Jumlah mahasiswa yang menggunakan fasilitas kampus melebihi kapasitas, dan kegiatan belajar mengajar menjadi tidak efektif.

Pengolahan data mahasiswa yang dilakukan hanya sebatas analisis statistik pada lulusan tanpa melihat data akademik sebelum kuliah. Hal ini menyebabkan belum dapat mengetahui penyebab mahasiswa lulus tidak tepat waktu. Evaluasi standar mahasiswa dan lulusan, ditentukan oleh rekrutmen mahasiswa baru dan lama studi [3].

Penelitian untuk mengetahui kelompok-kelompok data penerimaan mahasiswa baru, di mana dalam penelitian ini berhasil memetakan data mahasiswa baru menggunakan metode *K-Means* [4]. Penelitian terkait pengelompokan yang lain yaitu menggunakan metode *AHC* untuk pengelompokan skripsi mahasiswa [5]. Penelitian memperkenalkan algoritma *clustering* berbasis hierarki yang menawarkan beberapa kelebihan diantaranya tidak terpengaruh adanya *outlier* dan dapat mendeteksi jumlah *cluster* yang tepat [6].

2. TINJAUAN PUSTAKA

Penelitian yang dilakukan sebelumnya oleh Helilintar dkk tahun 2018. Penelitian tentang penerapan metode *k-means clustering* pada data penerimaan mahasiswa baru. Dengan menggunakan variabel : Nilai UAN, Asal sekolah dan Prodi. *Cluster* yang dibuat berjumlah dua *cluster* (k=2). Hasil K-*Means clustering* yang diperoleh ada dua kelompok atau *cluster*, *cluster* pertama jika calon mahasiswa berasal dari sekolah SMA maka rata-rata yang diambil adalah prodi sistem informasi, dan *cluster* kedua jika calon mahasiswa berasal dari sekolah SMK maka rata-rata yang diambil adalah teknik informatika [7].

Penelitian Suprawoto tahun 2016, tentang klasifikasi data mahasiswa menggunakan metode *k-means* untuk menunjang pemilihan strategi pemasaran. Dengan menggunakan variabel : nama mahasiswa, jurusan SLTA, nilai UAN, kota asal mahasiswa, IPK dan program studi yang dipilih. Jumlah *cluster* yang digunakan dalam penelitian ini yaitu tiga *cluster* (k=3). Berdasarkan penelitian untuk mengelompokkan mahasiswa berdasarkan nilai UN dan IPK. Menghasilkan pengelompokan jenjang pendidikan D3 dan jenjang S1. Dan dari data yang dilatih, didapatkan tiga kelompok/*cluster* di setiap jenjangnya [1].

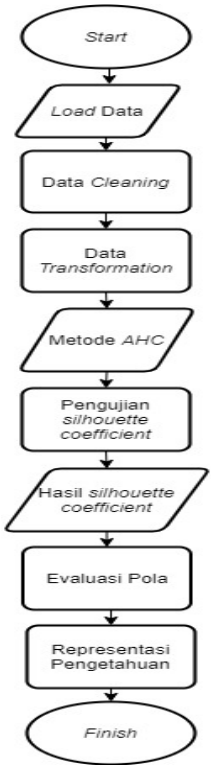
Penelitian Palumpun & Alam tahun 2017, tentang algoritma *K-Means* untuk pengelompokan tingkat kelulusan mahasiswa di Sains dan Teknologi Jayapura. Pengelompokan data kelulusan berdasarkan atribut IPK, lama studi, dan program studi dengan 3 *cluster*. Dari *cluster-cluster* tersebut terdapat lulusan yang memiliki lama studi kurang dari tepat waktu (5-6 tahun) memiliki IPK dan kelompok baik (3.01-3.50). Lulusan yang memiliki lama studi tidak tepat waktu (6-7 tahun) memiliki IPK dalam kelompok cukup (2.00-2.75) sampai dengan cukup baik (2.76-3.00). Sedangkan lulusan yang memiliki lama studi tepat waktu (3-4,5 tahun) memiliki IPK kelompok baik (3.01-3.50) dan sangat baik (>3.50) [8].

Penelitian Rosmini dkk tahun 2018, tentang implementasi metode *k-means* dalam pemetaan kelompok mahasiswa melalui data aktivitas kuliah. Dengan menggunakan variabel : IPK, Presensi, Tanggungan Biaya Kuliah, Organisasi Kampus, Pekerjaan, dan Status. Jumlah *cluster* yang digunakan dalam penelitian ini yaitu dua *cluster* (k=2) menggunakan metode *k-means clustering, cluster* A adalah mahasiswa yang lulus tepat waktu sedangkan *cluster* B adalah mahasiswa yang lulusnya tidak tepat waktu. Data pengelompokan mahasiswa ini merupakan masukan bagi dosen wali dalam membimbing dan mengawasi proses belajar mahasiswa agar bisa lulus tepat waktu [9].

Penelitian Februariyanti & Santoso tahun 2017, tentang *hierarchical agglomerative clustering* untuk pengelompokan skripsi mahasiswa dengan menggunakan judul skripsi sebagai variable. Jumlah *cluster* yang dihasilkan yaitu lima *cluster* (k=5). Hasil pengelompokan judul skripsi mendapatkan lima topik yang paling banyak diambil oleh mahasiswa untuk skripsi. Untuk skripsi di asumsikan topiknya adalah irisan dari masing-masing data pada setiap *cluster* diantaranya : Sistem Informasi, Sistem Informasi Semarang, Sistem pakar berbasis web, Sistem pakar diagnosa penyakit pada tanaman, dan Rancang bangun perangkat ajar anak berbasis [5].

3. METODE PENELITIAN

Dengan menggunakan data yang dihimpun melalui studi pustaka dan wawancara dengan narasumber, maka akan dilakukan pengelompokan mahasiswa lulus tepat waktu pada mahasiswa Fakultas Teknologi Industri dengan menggunakan metode *agglomerative hierarchical clustering.* Adapun metodologi yang digunakan dalam penelitian seperti terdapat pada Gambar 1.



Gambar 1. Metodologi Penelitian

3.1. *Load* Data

Proses *load* data awal pada tahapan ini, data dikumpulkan berupa *file Ms. excel* yang kemudian di *load* ke dalam program yang akan diolah.

3.2. Data *Cleaning*

Melakukan pembersihan data riwayat sebelum kuliah dan data kelulusan, yaitu menghilangkan *noise* dan data yang tidak relevan. Misalnya menghapus data yang belum memiliki kelulusan atau data yang bukan mendaftar melalui jalur PMDK-Raport dan tidak berhubungan dengan variabel yang akan digunakan.

3.3. Data *Transformation*

Transformasi data dapat berupa mengonversikan tipe data, membersihkan data dengan menghapus data nol atau duplikat, memperkaya data, atau melakukan agregasi. Transformasi data yang dilakukan yaitu asal sekolah, rata-rata Matematika, provinsi sekolah, lama studi, IPK, dan *TOEFL*.

3.4. Proses Pengelompokan Metode *AHC*

Melakukan tahap *AHC* yaitu mengelompokkan setiap obyek ke dalam kelompok/ *cluster*. Sehingga menemukan pasangan data paling mirip dimasukkan ke dalam *cluster* yang sama untuk melihat kemiripan data. Selanjutnya menggabungkan kedua obyek ke dalam satu *cluster*. Ulangi proses ini sampai tersisa hanya satu *cluster*.

3.5. Pengujian *Silhouette Coefficient*

Setelah sistem selesai dibuat maka tahapan selanjutnya adalah pengujian terhadap sistem. Pengujian ini bertujuan untuk menguji apakah sistem sudah berjalan sesuai keinginan atau belum. Pengujian terhadap sistem, dilakukan dengan metode *Agglomerative hierarchical clustering.* Pengujian dilakukan dengan menggunakan *silhouette coefficient.*

3.6. Hasil Pengujian *Silhouette Coefficient*

Dari hasil pengujian akan menghasilkan kelompok *cluster* terbaik berdasarkan masing-masing data *cluster* yang telah diuji. Data *cluster* dengan hasil nilai data tertinggi yang akan diterapkan menjadi kelompok data mahasiswa yang terbaik berdasarkan hasil pengujian.

3.7. Evaluasi Pola

Pola yang dihasilkan dari proses *data mining* menggunakan metode *AHC* perlu menampilkan sebuah bentuk informasi yang mudah dimengerti oleh pihak terkait agar menjadi pengetahuan dasar yang ditemukan.

3.8. Representasi Pengetahuan

Setelah didapatkan hasil visualisasi kemudian dilakukan analisis untuk dapat dimengerti oleh pihak terkait sebagai pengetahuan baru.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data mahasiswa Fakultas Teknologi Industri Universitas Ahmad Dahlan tahun angkatan 2014 dan 2015 yang diterima melalui jalur PMDK-Raport dan telah dinyatakan lulus. Data diperoleh dari Kantor Tata Usaha FTI UAD dengan jumlah data 283 data. Berikut dataset mahasiswa FTI angkatan 2014 dan 2015 dapat dilihat pada Tabel II.

Variabel data mahasiswa yang akan digunakan dalam proses data mining adalah sebagai berikut:

a. Asal Sekolah

Variabel Asal Sekolah berisi jenis sekolah yang ditempuh mahasiswa dengan kategori SMA, SMK dan MA.

TABEL II. *DATASET* MAHASISWA FTI

NIM	Sekolah	Asal Provinsi	Rata MTK	Lama Studi (Hari)	IPK	TOEFL
1400018012	SMA	DI Yogyakarta	83,33	1135	3,46	460
1400018017	SMA	Jawa Barat	89,33	1463	3,57	406
1400018026	SMK	Jawa Barat	78,33	1825	2,68	403
1400018041	SMA	Sulawesi Tengah	79	1463	3,63	400
1400018051	SMA	Riau	78	1478	3,56	433
1400018067	SMA	Nusa Tenggara Barat	89,33	1492	3,06	400
1400018070	SMK	DI Yogyakarta	83	1460	3,65	416
1400018082	SMK	DI Yogyakarta	87,33	1478	3,00	400
….	….	….	….	….	….	….
1500022083	SMK	Jawa Tengah	76	1135	3,65	406

b. Asal Provinsi

Variabel Asal provinsi berisi asal provinsi mahasiswa. Provinsi – provinsi tersebut dibagi menjadi 3 Wilayah yang berdasarkan pada perhitungan kualitas pendidikan di wilayah Indonesia menggunakan metode *K-Means* [10].

c. Rata MTK

Variabel Rata MTK adalah nilai rata – rata matematika mahasiswa saat melakukan pendaftaran masuk Universitas dari jalur PMDK-Raport.

d. Lama masa studi

Variabel lama masa studi menentukan apakah mahasiswa lulus dengan Tepat Waktu atau Tidak. Setelah melakukan wawancara dengan pihak Tata Usaha Fakultas menyatakan bahwa, Mahasiswa dikatakan lulus tepat waktu jika lama masa studi yang ditempuh adalah <= 4 tahun dan jika > 4 tahun maka tidak tepat waktu.

e. IPK

IPK adalah Indeks Prestasi Komulatif yang didapat setelah lulus kuliah.

f. TOEFL

Merupakan salah satu syarat kelulusan bagi mahasiswa tingkat akhir.

4.2. Pengujian Algoritma

Pengujian algoritma dilakukan dengan mengambil 10 data sampel dengan menggunakan metode *Agglomerative hierarchical clustering. Dataset* yang digunakan yaitu Prodi Informatika dapat dilihat pada Tabel III.

Pada tahap transformasi data, berupa proses konversi dari data ke sumber tujuan yang sudah ditetapkan. Transformasi yang akan dilakukan adalah mentransformasikan asal sekolah, dan asal provinsi menggunakan metode *One Hot Encoding*. Metode transformasi ini menggunakan angka biner sebagai *value* dari atribut atau data yang akan di transformasi [11].

TABEL III. DATASET PENGUJIAN

NIM	Sekolah	Asal Provinsi	Rata MTK	Lama Studi (Hari)	IPK	TOEFL
1400018012	SMA	DI Yogyakarta	83,33	1135	3,46	460
1400018017	SMA	Jawa Barat	89,33	1463	3,57	406
1400018026	SMK	Jawa Barat	78,33	1825	2,68	403
1400018041	SMA	Sulawesi Tengah	79	1463	3,63	400
1400018051	SMA	Riau	78	1478	3,56	433
1400018067	SMA	Nusa Tenggara Barat	89,33	1492	3,06	400
1400018070	SMK	DI Yogyakarta	83	1460	3,65	416
1400018082	SMK	DI Yogyakarta	87,33	1478	3,00	400
1400018084	SMA	Sumatera Selatan	84,33	1478	3,20	446
1400018092	SMK	Jawa Timur	76,66	1478	3,16	430

1. Asal Sekolah

TABEL IV. TRANSFORMASI ASAL SEKOLAH

NIM	Sekolah	MA	SMA	SMK
1400018012	SMA	0	1	0
1400018017	SMA	0	1	0
1400018026	SMK	0	0	1
1400018041	SMA	0	1	0
1400018051	SMA	0	1	0
1400018067	SMA	0	1	0
1400018070	SMK	0	0	1
1400018082	SMK	0	0	1
1400018084	SMA	0	1	0
1400018092	SMK	0	0	1

2. Asal Provinsi

Gambar 2. Pemetaan Kualitas Pendidikan di Indonesia [10]

Pengelompokan ini berdasarkan tingkat kualitas pendidikan dengan parameter tertentu yang sudah diteliti [10].

TABEL V. TRANSFORMASI WILAYAH

Wilayah	Warna Peta	Provinsi
Wilayah 1	Abu-abu	Maluku Utara dan Kalimantan Tengah.
Wilayah 2	Coklat	Aceh, Sumatra Barat, Sumatra Utara, Sumatra Selatan, Riau, Kepulauan Riau, Jambi, Bengkulu, Kepulauan Bangka Belitung, Lampung, Banten, DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa

		Timur, DI Yogyakarta, Bali, NTB, NTT, Kalimantan Barat, Kalimantan Selatan, Gorontalo, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tenggara, Sulawesi Tenggara, dan Sulawesi Utara.
Wilayah 3	Hijau	Papua, Papua Barat, Kalimantan Timur, dan Maluku.

TABEL VI. TRANSFORMASI ASAL PROVINSI

NIM	Asal Provinsi	Wilayah 1	Wilayah 2	Wilayah 3
1400018012	DI Yogyakarta	0	1	0
1400018017	Jawa Barat	0	1	0
1400018026	Jawa Barat	0	1	0
1400018041	Sulawesi Tengah	0	1	0
1400018051	Riau	0	1	0
1400018067	Nusa Tenggara Barat	0	1	0
1400018070	DI Yogyakarta	0	1	0
1400018082	DI Yogyakarta	0	1	0
1400018084	Sumatera Selatan	0	1	0
1400018092	Jawa Timur	0	1	0

Berikut merupakan transformasi keseluruhan *dataset* pengujian.

TABEL VII. TRANSFORMASI DATASET PENGUJIAN

NIM	SMA	SMK	Wilayah 2	Rata MTK	Lama Studi (Hari)	IPK	TOEFL
1400018012	1	0	1	83,33	1135	3,46	460
1400018017	1	0	1	89,33	1463	3,57	406
1400018026	0	1	1	78,33	1825	2,68	403
1400018041	1	0	1	79	1463	3,63	400
1400018051	1	0	1	78	1478	3,56	433
1400018067	1	0	1	89,33	1492	3,06	400
1400018070	0	1	1	83	1460	3,65	416
1400018082	0	1	1	87,33	1478	3,00	400
1400018084	1	0	1	84,33	1478	3,20	446
1400018092	0	1	1	76,66	1478	3,16	430

Tahapan selanjutnya menghitung *Euclidean Distance* digunakan untuk mengukur jarak terdekat pada setiap data mahasiswa. Dikatakan jarak terdekat adalah nilai jarak *Euclidean* yang paling kecil atau nilai yang mendekati 0.

Rumus *Euclidean Distance*

$$||U - V|| = \sqrt{\sum i\ (Ui - Vi)^2} \quad (1)$$

dimana:

U_i = nilai U pada data latih

V_i = nilai V pada data uji

Menghitung jarak *euclidean* pada setiap mahasiswa, seperti pada 2 sampel perhitungan berikut :

a. Menghitung MHS 1 – MHS 2 :

$$\sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (83{,}33 - 89{,}33)^2 + (1135 - 1463)^2 + (3{,}46 - 3{,}57)^2 + (460 - 406)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2}$$

$$= \sqrt{0 + 0 + 0 + 36 + 107584 + 0{,}012 + 2916 + 0 + 0 + 0}$$

$$= \sqrt{110.536{,}012} = 332{,}469$$

b. Menghitung MHS 1 – MHS 3 :

$$\sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (83{,}33 - 78{,}33)^2 + (1135 - 1825)^2 + (3{,}46 - 2{,}68)^2 + (460 - 403)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2}$$

$$= \sqrt{0 + 0 + 0 + 25 + 476100 + 0{,}608 + 3249 + 0 + 1 + 1}$$

$$= \sqrt{479.376{,}608} = 692.37$$

Jarak *euclidean* dihitung untuk seluruh data, dilakukan seperti perhitungan pada rumus Persamaan (1) seperti pada contoh.

TABEL VIII. HASIL PERHITUNGAN EUCLIDEAN DISTANCE

EUCLIDEAN DISTANCE										
Data	mhs 1	mhs 2	mhs 3	mhs 4	mhs 5	mhs 6	mhs 7	mhs 8	mhs 9	mhs 10
mhs 1	0	332,47	692,37	333,47	344,1	362,05	327,97	348,23	342,29	344,38
mhs 2	332,47	0	362,18	11,95	32,9	29,62	12,29	16,35	44,12	31,04
mhs 3	692,37	362,18	0	362,02	348,3	362,2	365,27	362,12	349,71	348,05
mhs 4	333,47	11,95	362,02	0	32,26	30,79	16,82	17,17	48,68	33,65
mhs 5	344,1	32,9	348,3	32,26	0	37,6	25,3	34,32	14,46	3,6
mhs 6	362,05	29,62	362,2	30,79	37,6	0	36,37	14,21	48,34	35,47
mhs 7	327,97	12,29	365,27	16,82	25,3	36,37	0	24,48	35,04	23,67
mhs 8	348,23	16,35	362,12	17,17	34,32	14,21	24,48	0	46,11	10,66
mhs 9	342,29	44,12	349,71	48,68	14,46	48,34	35,04	46,11	0	17,8
mhs 10	344,38	31,04	348,05	33,65	3,6	35,47	23,67	10,66	17,8	0

Berdasarkan matrik jarak, selanjutnya dilakukan pengelompokan data dengan *Agglomerative Hierarchical Clustering (AHC)* menggunakan metode yaitu *single linkage* pada Persamaan (2).

$$d_{data} = min\ \{ d_{data} \}\ ,\ d_{data} \in D \quad (2)$$

dimana :

d_{data} = jarak antara tetangga terdekat/terkecil dari kelompok data

D = *matriks* kedekatan jarak *Euclidean*

Dengan menganggap data sebagai kelompok, selanjutnya memilih jarak dua kelompok yang terkecil terdapat pada Tabel IX.

$$\min(d_{data}) = \min(d_{mhs5,mhs10}) = 3{,}6$$

TABEL IX. TAHAPAN 1 SINGLE LINKAGE

Data	mhs 1	mhs 2	mhs 3	mhs 4	mhs 5	mhs 6	mhs 7	mhs 8	mhs 9	mhs 10
mhs 1	0	332,47	692,37	333,47	344,1	362,05	327,97	348,23	342,29	344,38
mhs 2	332,47	0	362,18	11,95	32,9	29,62	12,29	16,35	44,12	31,04
mhs 3	692,37	362,18	0	362,02	348,3	362,2	365,27	362,12	349,71	348,05
mhs 4	333,47	11,95	362,02	0	32,26	30,79	16,82	17,17	48,68	33,65
mhs 5	344,1	32,9	348,3	32,26	0	37,6	25,3	34,32	14,46	3,6
mhs 6	362,05	29,62	362,2	30,79	37,6	0	36,37	14,21	48,34	35,47
mhs 7	327,97	12,29	365,27	16,82	25,3	36,37	0	24,48	35,04	23,67
mhs 8	348,23	16,35	362,12	17,17	34,32	14,21	24,48	0	46,11	10,66
mhs 9	342,29	44,12	349,71	48,68	14,46	48,34	35,04	46,11	0	17,8
mhs 10	344,38	31,04	348,05	33,65	3,6	35,47	23,67	10,66	17,8	0

Tahapan selanjutnya dipilih jarak terkecil dari kelompok untuk menghitung jarak antara kelompok mhs 5 dan mhs 10 dengan kelompok lain yang masih tersisa yaitu, mhs 1, mhs 2, mhs 3, mhs 4, mhs 6, mhs 7, mhs 8, dan mhs 9. Dengan menghapus baris-baris dan kolom-kolom matrik pada kelompok mhs 5 dan mhs 10, serta menambahkan baris dan kolom untuk kelompok (mhs 5, mhs 10). Hasil proses pengelompokan tahap 1 dapat dilihat pada Tabel IX. Dengan menganggap data sebagai kelompok, selanjutnya memilih jarak dua kelompok yang terkecil terdapat pada Tabel X.

$$\min(d_{data}) = \min(d_{mhs5,mhs10,mhs8}) = 10{,}66$$

TABEL X. TAHAPAN 2 SINGLE LINKAGE

Data	mhs 5, mhs 10	mhs 1	mhs 2	mhs 3	mhs 4	mhs 6	mhs 7	mhs 8	mhs 9
mhs 5, mhs 10	0	344,1	31,04	348,05	32,26	35,47	23,67	10,66	14,46
mhs 1	344,1	0	332,47	692,37	333,47	362,05	327,97	348,23	342,29
mhs 2	31,04	332,47	0	362,18	11,95	29,62	12,29	16,35	44,12
mhs 3	348,05	692,37	362,18	0	362,02	362,2	365,27	362,12	349,71
mhs 4	32,26	333,47	11,95	362,02	0	30,79	16,82	17,17	48,68
mhs 6	35,47	362,05	29,62	362,2	30,79	0	36,37	14,21	48,34
mhs 7	23,67	327,97	12,29	365,27	16,82	36,37	0	24,48	35,04
mhs 8	10,66	348,23	16,35	362,12	17,17	14,21	24,48	0	46,11
mhs 9	14,46	342,29	44,12	349,71	48,68	48,34	35,04	46,11	0

Tahapan single linkage ini dilakukan sampai tersisa hanya satu cluster atau pengelompokan. Sehingga diperoleh hasil cluster atau pengelompokan yang dapat dilihat pada Tabel XI.

TABEL XI. HASIL CLUSTER ATAU PENGELOMPOKAN

Tahapan	Jumlah *Cluster*	Anggota	Jarak Terkecil
1	9	mhs 5, mhs 10	3,6
2	8	mhs 5, mhs 10, mhs 8	10,66
3	7	mhs 2, mhs 4	11,98
4	6	mhs 2, mhs 4, mhs 7	12,29
5	5	mhs 5, mhs 10, mhs 8, mhs 6	14,21
6	4	mhs 5, mhs 10, mhs 8, mhs 6, mhs 9	14,46
7	3	mhs 5, mhs 10, mhs 8, mhs 6, mhs 9, mhs 2, mhs 4, mhs 7	17,17
8	2	mhs 5, mhs 10, mhs 8, mhs 6, mhs 9, mhs 2, mhs 4, mhs 7, mhs 1	327,97
9	1	mhs 5, mhs 10, mhs 8, mhs 6, mhs 9, mhs 2, mhs 4, mhs 7, mhs 1, mhs 3	348,05

Selanjutnya menguji hasil *cluster* dengan menggunakan *silhouette coefficient.*

a. Menghitung nilai rata-rata jarak objek dengan dokumen yang berada dalam satu cluster dengan menggunakan Persamaan (3).

$$a(i) = \frac{1}{[A]-1} \quad \sum j \in C\, d\,(i\,,\, j) \qquad (3)$$

dimana :
a(i) =Perbedaan rata-rata objek (*i*) ke semua objek lain pada A.
d(i,j) =Jarak antara data *i* ke *j.*
A =*Cluster*.

TABEL XII. HASIL PERHITUNGAN A(I)

Data	*a(i)*	*Cluster*
mhs 1	341,87	0
mhs 2	63,84	0
mhs 3	444,03	1
mhs 4	65,60	0
mhs 5	109,11	0
mhs 6	65,57	0
mhs 7	62,74	0
mhs 8	63,94	0
mhs 9	74,61	0
mhs 10	62,53	0

b. Menghitung jarak objek dengan dokumen antara *cluster* dengan menggunakan Persamaan (4) dan Persamaan (5).

$$d(i, C) = \frac{1}{[A]} \quad \sum j \in C\, d\,(i\,,\, j) \qquad (4)$$

$$b(i) = \min C \neq A \;\; d(i,C) \qquad (5)$$

dimana :
d(i, C) = Perbedaan rata-rata objek (*i*) ke semua objek lain pada C.
C = *Cluster* lain selain *cluster* A atau *cluster* C tidak sama dengan *cluster* A.

b(i) = Rara-rata jarak objek dengan semua objek lain yang berbeda pada *cluster* lainnya.

TABEL XIII. HASIL PERHITUNGAN D(I, C) DAN B(I)

Data	*d(i,C)*	*b(i)*	*Cluster*
mhs 1	76,93	692,37	0
mhs 2	40,24	362,18	0
mhs 3	0,00	355,22	1
mhs 4	40,22	362,02	0
mhs 5	38,70	348,30	0
mhs 6	40,24	362,20	0
mhs 7	40,59	365,27	0
mhs 8	40,24	362,12	0
mhs 9	38,86	349,71	0
mhs 10	38,67	348,05	0

c. Kemudian menghitung nilai *Silhouette Coefficient dengan* menggunakan Persamaan (6).

$$s(i) = \frac{b(i)-a(i)}{min(a(i),b(i))} \quad (6)$$

dimana :
S(i) = *Silhouette Coefficient*
a(i) = Rata-rata jarak objek dengan semua objek yang berbeda dalam satu *cluster*.
b(i) = Rara-rata jarak objek dengan semua objek lain yang berbeda pada *cluster* lainnya.

TABEL XIV. Hasil *Silhouette Coefficient s(i)*

Duv	*a(i)*	*b(i)*	*s(i)*	*Cluster*
mhs 1	341,87	692,37	0,51	0
mhs 2	63,84	362,18	0,82	0
mhs 3	444,03	355,22	-0,25	1
mhs 4	65,60	362,02	0,82	0
mhs 5	109,11	348,30	0,69	0
mhs 6	65,57	362,20	0,82	0
mhs 7	62,74	365,27	0,83	0
mhs 8	63,94	362,12	0,82	0
mhs 9	74,61	349,71	0,79	0
mhs 10	62,53	348,05	0,82	0

Berdasarkan hasil pengujian *silhouette coefficient* untuk kualitas *cluster* dengan jumlah sampel 10 data diperoleh dua nilai, nilai *silhouette coefficient* sebesar 0,83 yang artinya kualitas *cluster* baik dan nilai *silhouette coefficient* sebesar -0,25 yang artinya kualitas *cluster* kurang baik. Representasi pengetahuan dari hasil *clustering* yang disajikan pada Tabel XIV sebagai berikut :

1. *Cluster* 0 memiliki anggota sebanyak 9 data dengan rata-rata matematika 83,37 dan masa studi 1436 hari, serta dengan rata-rata IPK 3,37.
2. *Cluster* 1 memiliki anggota sebanyak 1 data dengan rata-rata matematika 78,33, masa studi 1825 hari, serta dengan rata-rata IPK 2,68.

4.3. Pengembangan Sistem

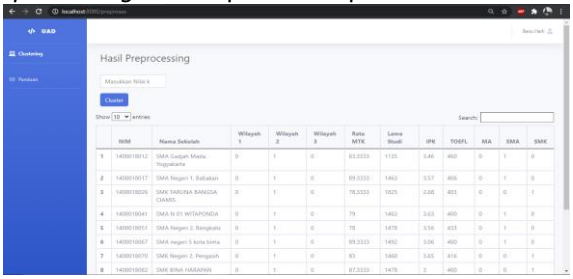
4.3.1. *Load* data

Tahapan *load* data merupakan tahapan awal dalam proses *data mining*. Di mana, pengguna meng-*upload* data mahasiswa FTI dalam format *excel*. *Dataset* berupa *file excel* yang terpisah menjadi 4 buah file dengan nama file "Informatika.xlsx", "Kimia.xlsx", "Industri.xlsx", dan "Elektro.xlsx".
Dengan menggunakan salah satu *dataset* yaitu *dataset* Program Studi Informatika, hasil dari load data dapat dilihat pada Gambar 3.

Gambar 3. *Load Dataset*

4.3.2. *Preprocessing* data

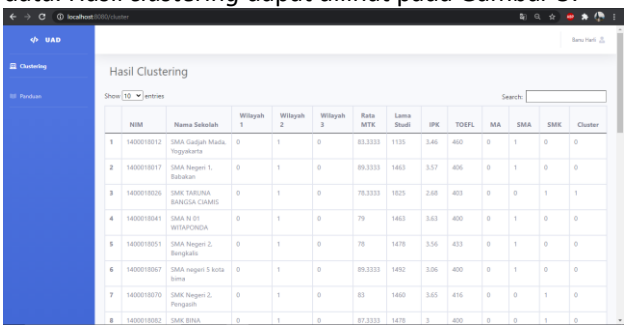
Cleaning atau Pembersihan Data adalah proses di mana data - data mahasiswa yang telah di *load* dibersihkan terlebih dahulu dari data – data *noise* (memiliki nilai aneh) dan data yang tidak digunakan dalam proses *mining*. Selanjutnya tahapan seleksi lagi untuk memutuskan variabel data yang akan digunakan dalam proses *mining* karena hanya data yang relevan yang diambil dari database. Setelah diseleksi data akan melalui proses transformasi. Hasil *preprocessing* data dapat dilihat pada Gambar 4.



Gambar 4. *Preprocessing* Data

4.3.3. *Agglomerative Hierarchical Clustering*

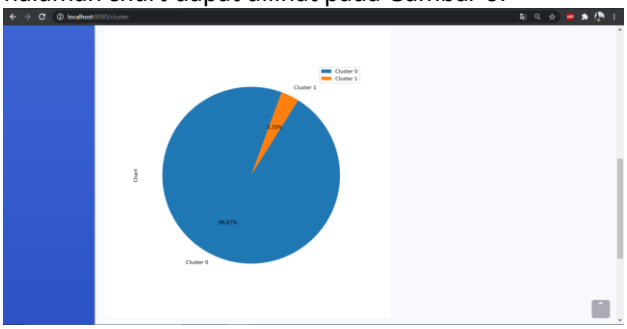
Tahap *AHC* yaitu mengelompokkan setiap obyek ke dalam kelompok atau *cluster*. Sehingga menemukan pasangan data paling mirip dimasukkan ke dalam *cluster* yang sama untuk melihat kemiripan data. Hasil *clustering* dapat dilihat pada Gambar 5.



Gambar 5. Pengelompokan *AHC*

4.3.4. *Chart*

Chart adalah menampilkan jumlah data yang terbentuk dari hasil *cluster* untuk mengetahui *cluster* yang ada dalam pengelompokan data. Tampilan halaman *chart* dapat dilihat pada Gambar 6.



Gambar 6. *Chart*

4.3.5. Pengujian

Pengujian adalah cara untuk mengetahui seberapa baik hasil yang diperoleh dari proses *clustering* menggunakan *silhouette coefficient.* Tampilan halaman pengujian dapat dilihat pada Gambar 7.

Gambar 7. Pengujian

4.3.6. Representasi Pengetahuan

a. *Dataset* Informatika

Pada *dataset* Informatika dengan jumlah data 90, diperoleh pengujian yang disajikan pada Tabel XV.

TABEL XV. HASIL PENGUJIAN DATASET INFORMATIKA

Nilai K (Jumlah Titik Pusat)	Hasil *Silhouette Coefficient*
2	0,593
3	0,868
4	0,878
5	0,861
6	0,786
7	0,791
8	0,782

Berdasarkan hasil pengujian pada *dataset* Informatika dapat ditarik kesimpulan bahwa pada *dataset* Informatika jumlah *cluster* terbaik yaitu 4. Sehingga diperoleh 4 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

TABEL XVI. HASIL CLUSTER INFORMATIKA

Nilai k	Data	Asal Sekolah	Wilayah	Pulau	Rata MTK	TOEFL	IPK	Lama Studi
cluster 0	4	SMA & SMK	2	Kalimantan	63	206	3,44	3 Tahun 10 Bulan
cluster 1	43	SMA	2	Jawa	80	424	3,5	3 Tahun 9 Bulan
cluster 2	40	SMA	2	Jawa	82	428	3,38	4 Tahun 3 Bulan
cluster 3	3	SMA	2	Jawa	79	426	2,96	5 Tahun

Hasil dari evaluasi pola pada *dataset* informatika dengan jumlah *cluster* 4, maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi lulus tepat waktu, jika nilai rata-rata matematika 80, berasal dari pulau Jawa, dan rata-rata IPK sebesar 3,50.

b. *Dataset Kimia*

Pada *dataset* Kimia dengan jumlah data 87, diperoleh pengujian yang disajikan pada Tabel XVII.

TABEL XVII. HASIL PENGUJIAN DATASET KIMIA

Nilai K (Jumlah Titik Pusat)	Hasil *Silhouette Coefficient*
2	0,873
3	0,819
4	0,539
5	0,473
6	0,463
7	0,446
8	0,399

Berdasarkan hasil pengujian pada *dataset* kimia dapat ditarik kesimpulan bahwa pada *dataset* kimia jumlah *cluster* terbaik yaitu 2. Sehingga diperoleh 2 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

TABEL XVIII. HASIL CLUSTER KIMIA

Nilai k	Data	Asal Sekolah	Wilayah	Pulau	Rata MTK	TOEFL	IPK	Lama Studi
cluster 0	15	SMA	2	Jawa	60	420	3,33	4 Tahun 2 Bulan
cluster 1	72	SMA	2	Jawa	67	438	3,53	3 Tahun 9 Bulan

Hasil dari evaluasi pola pada *dataset* kimia dengan jumlah *cluster* 2, maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi lulus

tepat waktu, jika nilai rata-rata matematika 67, berasal dari pulau Jawa, dan rata-rata IPK sebesar 3,53.

c. *Dataset* Industri

Pada *dataset* Industri dengan jumlah data 76, diperoleh pengujian yang disajikan pada Tabel XIX.

TABEL XIX. HASIL PENGUJIAN DATASET INDUSTRI

Nilai K (Jumlah Titik Pusat)	Hasil *Silhouette Coefficient*
2	0,749
3	0,625
4	0,883
5	0,710
6	0,374
7	0,404
8	0,407

Berdasarkan hasil pengujian pada *dataset* Industri dapat ditarik kesimpulan bahwa pada *dataset* Industri jumlah *cluster* terbaik yaitu 4. Sehingga diperoleh 4 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

TABEL XX. Hasil *Cluster* Industri

Nilai k	Data	Asal Sekolah	Wilayah	Pulau	Rata MTK	TOEFL	IPK	Lama Studi
cluster 0	49	SMA	2	Jawa	75	428	3,36	4 Tahun 2 Bulan
cluster 1	24	SMA	2	Jawa	84	423	3,56	3 Tahun 10 Bulan
cluster 2	2	SMA	2	Sumatra	84	455	3,22	5 Tahun
cluster 3	1	SMA	2	Sumatra	78	423	3,31	6 Tahun

Hasil dari evaluasi pola pada *dataset* Industri dengan jumlah *cluster* 4*,* maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi lulus tepat waktu, jika nilai rata-rata matematika 84, berasal dari pulau Jawa, dan rata-rata IPK sebesar 3,56.

d. *Dataset* Elektro

Pada *dataset* Elektro dengan jumlah data 30, diperoleh pengujian yang disajikan pada Tabel XXI.

TABEL XXI. Hasil Pengujian *Dataset* Elektro

Nilai K (Jumlah Titik Pusat)	Hasil *Silhouette Coefficient*
2	0,602
3	0,879
4	0,768
5	0,686
6	0,716
7	0,458
8	0,325

Berdasarkan hasil pengujian pada *dataset* Elektro dapat ditarik kesimpulan bahwa pada *dataset* Industri jumlah *cluster* terbaik yaitu 3. Sehingga diperoleh 3 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

TABEL XXII. HASIL CLUSTER ELEKTRO

Nilai k	Data	Asal Sekolah	Wilayah	Pulau	Rata MTK	TOEFL	IPK	Lama Studi
cluster 0	10	SMA & SMK	2	Sumatra	74	435	3,23	4 Tahun 4 Bulan
cluster 1	1	SMA	2	Jawa	80	443	3,31	5 Tahun
cluster 2	19	SMA & SMK	2	Jawa	80	421	3,56	3 Tahun 10 Bulan

Hasil dari evaluasi pola pada *dataset* Elektro dengan jumlah *cluster* 3*,* maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi lulus tepat waktu, jika nilai rata-rata matematika 80, berasal dari pulau Jawa, dan rata-rata IPK sebesar 3,56.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian "Penerapan *Data Mining* Dalam Mengelompokkan Data Riwayat Akademik Sebelum Kuliah dan Data Kelulusan Mahasiswa Menggunakan Metode *Agglomerative Hierarchical Clustering"* dapat ditarik kesimpulan sebagai berikut :

a. Penelitian menghasilkan 158 data mahasiswa yang di rekomendasikan dari keseluruhan *dataset* yang terdiri dari beberapa data :
   1. Rekomendasi prodi Informatika jumlah data sebanyak 43 data dan mayoritas berasal dari SMA, Asal wilayah 2, pulau Jawa, serta memiliki rata-rata nilai matematika sebesar 80, *TOEFL* sebesar 424, IPK sebesar 3,50 dengan lama studi yaitu 3 tahun 9 bulan.
   2. Rekomendasi prodi Kimia jumlah data sebanyak 72 data dan mayoritas berasal dari SMA, Asal wilayah 2, pulau Jawa, serta memiliki rata-rata nilai matematika sebesar 67, *TOEFL* sebesar 438, IPK sebesar 3,53 dengan lama studi yaitu 3 tahun 9 bulan.
   3. Rekomendasi prodi Industri jumlah data sebanyak 24 data dan mayoritas berasal dari SMA, Asal wilayah 2, pulau Jawa, serta memiliki rata-rata nilai matematika sebesar 84, *TOEFL* sebesar 423, IPK sebesar 3,56 dengan lama studi yaitu 3 tahun 10 bulan.
   4. Rekomendasi prodi Elektro jumlah data sebanyak 19 data dan berasal dari SMA dan SMK berjumlah sama, Asal wilayah 2, pulau Jawa, serta memiliki rata-rata nilai matematika sebesar 80, *TOEFL* sebesar 421, IPK sebesar 3,56 dengan lama studi 3 tahun 10 bulan.

b. Hasil pengujian pada aplikasi "Penerapan *Data Mining* Dalam Mengelompokkan Data Riwayat

Akademik Sebelum Kuliah dan Data Kelulusan Mahasiswa Menggunakan Metode *Agglomerative Hierarchical Clustering"* dan menggunakan pengujian hasil *cluster* dengan *Silhouette Coefficient* memperoleh hasil yang sangat bagus dengan nilai *silhouette coefficient* untuk *dataset* prodi Informatika sebesar 0,868, untuk *dataset* prodi Kimia sebesar 0,873, untuk *dataset* prodi Industri sebesar 0,883, dan untuk *dataset* prodi Elektro sebesar 0,879.

5.2. Saran

Beberapa saran yang dapat dijadikan landasan untuk pengembangan dan perbaikan untuk menutupi kekurangan dan kelemahan dari penelitian ini, sebagai berikut :

a. Menggunakan *dataset* yang lebih banyak dan menambahkan variabel atau atribut yang dapat menghasilkan pengelompokan yang lebih akurat.
b. Menggunakan beberapa metode pengelompokan *Agglomerative Hierarchical* lainnya seperti : *Complete Linkage* dan *Average Linkage*.

DAFTAR PUSTAKA

[1] T. Suprawoto, "Klasifikasi Data Mahasiswa Menggunakan Metode K-Means Untuk Menunjang Pemilihan Strategi Pemasaran," *JIKO (Jurnal Inform. dan Komputer)*, vol. 1, no. 1, pp. 12–18, 2016, doi: 10.26798/jiko.2016.v1i1.9.

[2] N. Syahid, "Tahun 2019 'Lulusan Tepat Waktu Mahasiswa FTI UAD Angkatan 2015 Sebanyak 22%,'" 2019. https://fti.uad.ac.id/lulusan-tepat-waktu-mahasiswa-fti-uad-angkatan-2015-sebanyak-22/.

[3] J. F. Ulysses, "Data Mining Classification Untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan Dengan Metode Naive Bayes," no. 125301917, pp. 1–8, 2008.

[4] F. Nasari and S. Darma, "Seminar Nasional Teknologi Informasi dan Multimedia 2015 Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru (Studi Kasus : Universitas Potensi Utama)," pp. 6–8, 2015.

[5] H. Februariyanti and D. B. Santoso, "Hierarchical Agglomerative Clustering Untuk Pengelompokan Skripsi Mahasiswa," *Pattern Recognit.*, 2017, doi: 10.1016/0031-3203(79)90049-9.

[6] J. A. S. Almeida, L. M. S. Barbosa, A. A. C. C. Pais, and S. J. Formosinho, "Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering," *Chemom. Intell. Lab. Syst.*, vol. 87, no. 2, pp. 208–217, 2007, doi: 10.1016/j.chemolab.2007.01.005.

[7] R. Helilintar, I. N. Farida, and R. H. Irawan, "Penerapan Metode K-Means Clustering Pada Data Penerimaan Mahasiswa Baru," pp. 14–20, 2018.

[8] Y. Palumpun and S. N. Alam, "Pengelompokan Tingkat Kelulusan Mahasiswa Menggunakan Algoritma K-Means," no. November, pp. 98–102, 2017.

[9] R. Rosmini, A. Fadlil, and S. Sunardi, "Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," *It J. Res. Dev.*, vol. 3, no. 1, p. 22, 2018, doi: 10.25299/itjrd.2018.vol3(1).1773.

[10] G. S. Nugraha and H. Hairani, "Aplikasi Pemetaan Kualitas Pendidikan di Indonesia Menggunakan Metode K-Means," *J. MATRIK*, vol. 17, no. 2, pp. 13–23, 2018, doi: 10.30812/matrik.v17i2.84.

[11] M. DelSole, "What is One Hot Encoding and How to Do It," 2018. https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179 (accessed Dec. 13, 2019).