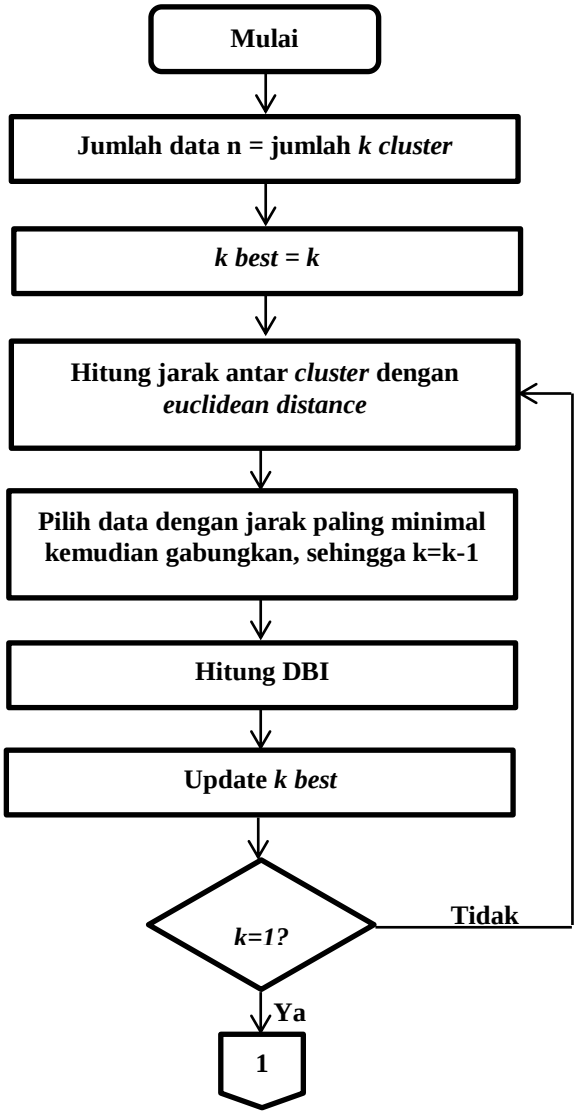
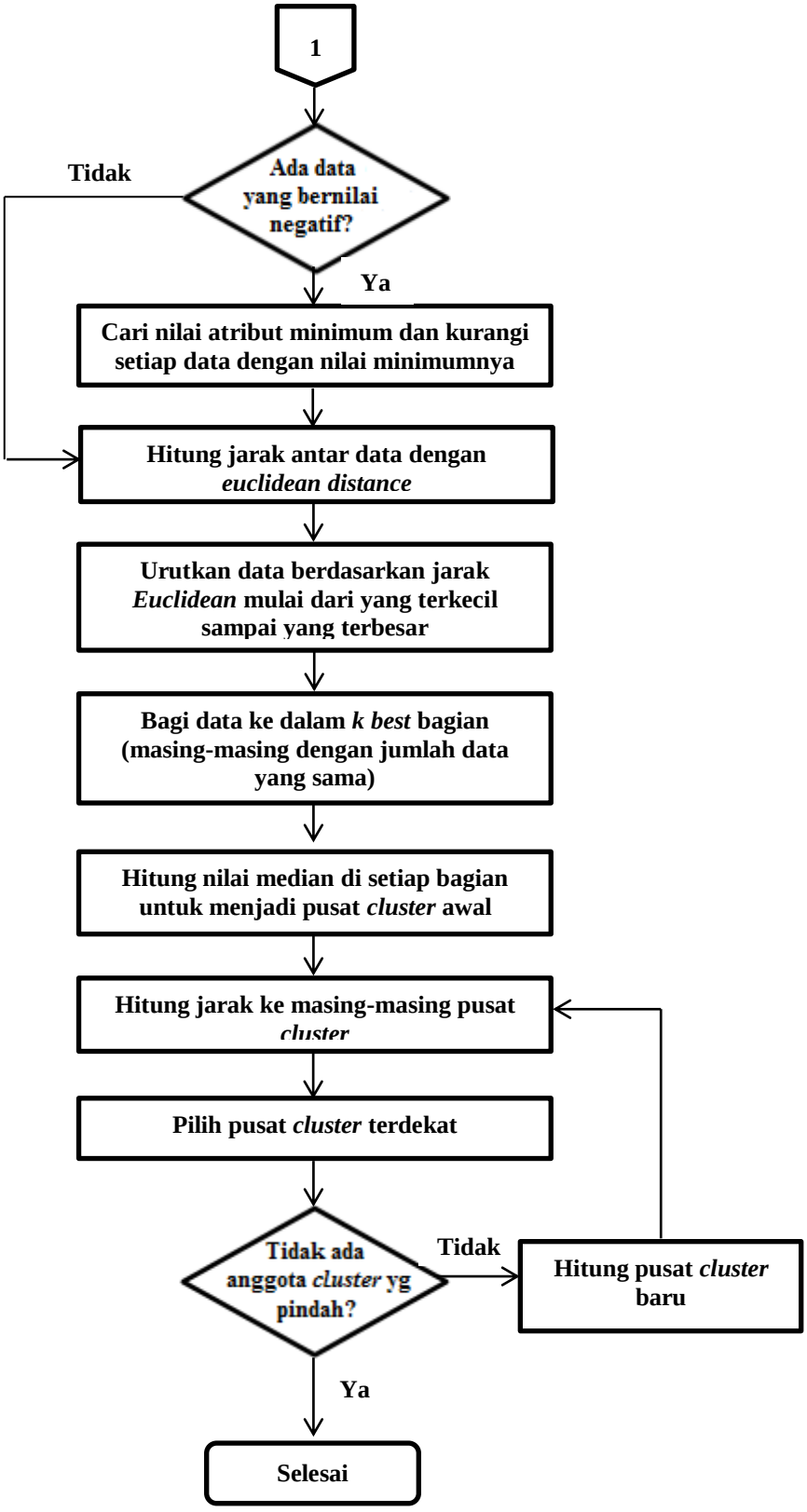
BAB 3
METODE PENELITIAN

3.1 Hierarchical Agglomerative - Enhanced K-Means *Clustering*

Hierarchical Agglomerative - Enhanced K-Means (HA-EKM) *Clustering* merupakan penggabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means dengan menggabungkan kelebihan masing-masing algoritma untuk menduga jumlah *cluster* terbaik sekaligus menentukan pusat *cluster* awal dari setiap *cluster*. Algoritma HA-EKM dapat dilihat di Gambar 3.1 dan Gambar 3.2.



Gambar 3.1 Algoritma HA – EKM Bagian 1



Gambar 3.2 Algoritma HA – EKM Bagian 2

Proses *clustering* algoritma HA-EKM diawali dengan algoritma Hierarchical Agglomerative Clustering yang bertujuan untuk mencari jumlah *cluster* terbaik yang ditunjukkan di Gambar 3.1 dan dilanjutkan dengan algoritma Enhanced K-Means untuk mencari pusat *cluster* awal dan sekaligus membentuk hasil *cluster* akhir. Proses ini ditunjukkan di Gambar 3.2.

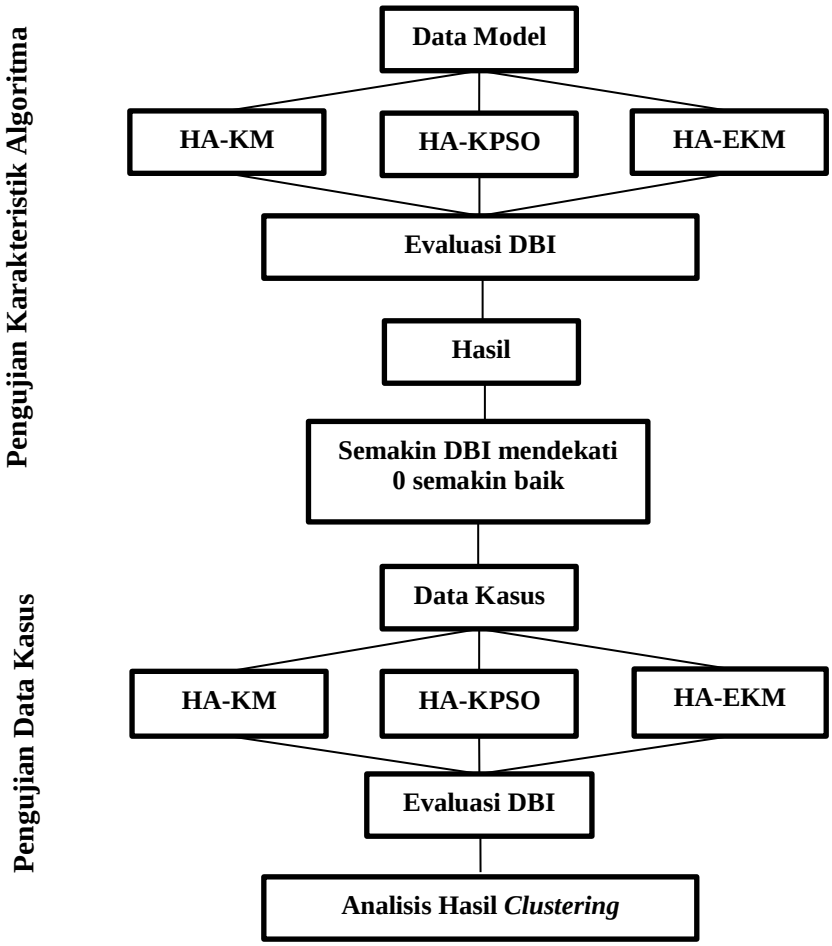
Pada Hierarchical Agglomerative Clustering, sebelum dilakukan pengelompokan, setiap data diasumsikan sebagai *cluster*. Jika terdapat data sebanyak *n* dan *k* dianggap sebagai jumlah *cluster*, maka *n* = *k*. Di awal tahap diasumsikan juga bahwa *k* terbaik adalah *k*. Selanjutnya dilakukan perhitungan jarak antar *cluster* dengan menggunakan jarak *Euclidean* dan berdasarkan jarak tersebut dipilih data yang memiliki jarak paling minimal kemudian gabungkan sehingga $k = k - 1$. Proses berlanjut hingga $k = 1$. Algoritma Hierarchical Agglomerative Clustering selengkapnya dijelaskan di BAB 2 subbab 2.2.3. Pada setiap proses penggabungan, dilakukan evaluasi *cluster* menggunakan teknik evaluasi DBI. Teknik evaluasi ini selengkapnya dijelaskan di BAB 2 subbab 2.3. Berdasarkan hasil evaluasi tersebut dilakukan peng-*update*-an *k* terbaik dengan aturan jika hasil evaluasi *cluster* lebih baik dari sebelum penggabungan maka *k best* = *k*. Sedangkan jika tidak lebih baik, maka *k best* tidak berubah. Karena jumlah *cluster* yang dibentuk biasanya berjumlah lebih kecil dari jumlah data, maka tentukan jumlah *k* maksimal. Perhitungan DBI dan peng-*update*-an *k* terbaik dilakukan pada saat jumlah *k* kurang dari atau sama dengan *k* maksimal.

Setelah didapatkan nilai *k best*, langkah selanjutnya adalah melakukan pengelompokan dengan algoritma Enhanced K-Means. Tahapannya dimulai dengan pengecekan apakah *dataset* yang digunakan mengandung atribut dengan nilai negatif atau tidak. Jika *dataset* mengandung atribut yang bernilai negatif maka dilakukan pengubahan seluruh data di dalam *dataset* ke dalam bentuk positif dengan cara mengurangi setiap data dengan nilai atribut minimum dalam *dataset*. Setelah itu, hitung jarak antar data dengan menggunakan jarak *Euclidean* lalu berdasarkan jarak tersebut dilakukan pengurutan data mulai dari jarak terkecil hingga yang terbesar. Selanjutnya data dibagi ke dalam *k best* bagian dengan

masing-masing bagian memiliki jumlah data yang sama banyak. Lalu pada masing-masing bagian dihitung nilai mediannya. Nilai median inilah yang akan menjadi pusat *cluster* awal. Selanjutnya dilakukan perhitungan jarak setiap data ke setiap pusat *cluster* awal. Berdasarkan jarak tersebut data dimasukkan ke dalam *cluster* yang memiliki jarak terdekat dengan pusat *cluster*. Proses ini akan berulang hingga mencapai kondisi tidak ada anggota *cluster* yang berpindah dan proses *clustering* pun selesai. Algoritma Enhanced K-Means selengkapnya dijelaskan di BAB 2 subbab 2.2.4.

3.2 Teknik Pengujian

Dalam penelitian ini, pengujian dibagi ke dalam dua bagian yaitu pengujian karakteristik algoritma dan pengujian data kasus untuk strategi promosi universitas. Tahapan pengujian dalam penelitian ini diperlihatkan di Gambar 3.3.



Gambar 3.3 Teknik Pengujian Karakteristik Algoritma

Untuk dapat memberikan gambaran mengenai karakteristik algoritma secara menyeluruh maka diperlukan data model buatan yang sudah diketahui karakteristik bentuk, jumlah *cluster* dan jarak antar *cluster* nya sehingga dapat diketahui karakteristik algoritma dalam mengelompokan data. Dengan begitu dapat diketahui karakteristik data terbaik dimana algoritma dapat diimplementasikan. Setelah itu, untuk menguji kinerja algoritma terhadap data yang sering dipakai dalam penelitian di wilayah *data mining clustering* maka diperlukan juga data testing yang digunakan untuk menguji kesesuaian perilaku algoritma terhadap data. Setelah dilakukan pengujian karakteristik algoritma, langkah selanjutnya adalah melakukan pengujian pada data kasus untuk mengetahui algoritma yang paling baik untuk diimplementasikan pada data kasus. Data yang digunakan dalam penelitian ini selengkapnya dijelaskan pada subbab 3.2.1 sementara alur pengujiannya dijelaskan pada subbab 3.2.2.

3.2.1 Pemilihan Data Penelitian

Terdapat dua macam data yang digunakan dalam penelitian ini yaitu data model dan data kasus. Data model digunakan pada pengujian karakteristik algoritma, sementara data kasus digunakan pada pengujian data kasus.

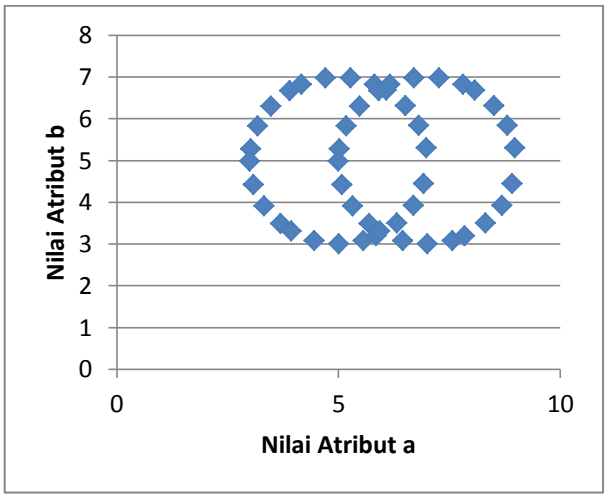
1. Data Model

Data model merupakan data yang digunakan untuk mengetahui karaktersitik algoritma dalam mengelompokkan data. Data model terdiri dari data model buatan dan data testing.

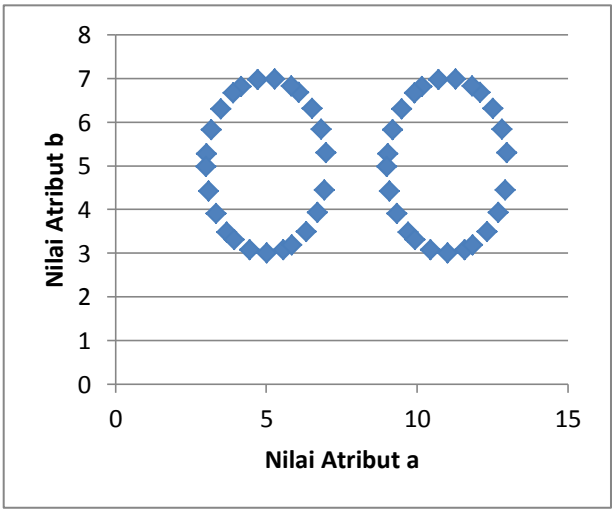
1. Data Model Buatan

Data model buatan merupakan data yang dibuat sendiri oleh peneliti yang dimanipulasi sebaran data, jumlah data, jumlah atribut serta jumlah *cluster*-nya. Data model digunakan untuk mengetahui perilaku algoritma dalam mengelompokkan data. Agar data model sesuai dengan data uji yang sering dipakai dalam penelitian di bidang *data mining clustering*, maka karakteristik data model seperti jumlah *cluster*, jumlah atribut dan jumlah data disesuaikan dengan karakteristik data testing. Jumlah atribut dibuat mulai dari 2 hingga 13 atribut,

sementara jumlah bentuk *cluster* awal dibuat mulai dari 2 dan 3 *cluster*. Penambahan jumlah data pada *dataset* 2 *cluster* dimulai dari 24, 48, 96, 192, 384 hingga 768, sementara penambahan jumlah data pada *dataset* 3 *cluster* dimulai dari 36, 72, 144, 288, 576 dan 1152. Manipulasi sebaran data dilakukan dengan menjauhkan jarak antar titik pusat *cluster* (jarak inter *cluser*) seperti yang dapat dilihat di Gambar 3.4 dan Gambar 3.5



Gambar 3.4 Contoh Sebaran Data Model 1



Gambar 3.5 Contoh Sebaran Data Model 2

Gambar 3.4 dan Gambar 3.5 memperlihatkan contoh data model dengan jumlah 2 atribut yang memiliki 24 data pada masing-masing *cluster*. Dari kedua gambar tersebut terlihat bahwa data model 1 memiliki jarak *cluster* yang

berdekatan yaitu dengan jarak inter 2 sementara data model 2 memiliki jarak yang lebih jauh yaitu dengan jarak inter 6. Data model ini digunakan untuk melihat karakteristik algoritma dalam menentukan jumlah *cluster*.

2. Data Testing

Data testing merupakan data yang digunakan untuk mengkonfirmasi kesesuaian perilaku algoritma pada data model dengan data pada kasus yang lebih nyata yang digunakan pada penelitian-penelitian sebelumnya pada bidang *clustering*. Data testing diperoleh dari UCI *repository of machine learning databases*, yaitu repositori yang disediakan oleh Nationan Science Foundation yang berisi kumpulan *dataset* untuk keperluan penelitian di bidang *data mining.* Data testing terdiri dari 3 *dataset* uji yaitu Iris, Diabetes dan Wine. Karakteristik ketiga *dataset* tersebut disajikan di Tabel 3.1.

Tabel 3.1 Karakteristik Data Testing

Dataset	**Jumlah Data**	**Jumlah Atribut**	**Nilai Atribut yang Hilang**	**Keterangan**
Iris (UCI *repository*)	150	4	Tidak	Pengelompokan bunga Iris berdasarkan panjang dan lebar kelopak dan daun bunga
Diabetes (UCI *repository*)	768	8	Ya	Kemungkinan seseorang terkena penyakit diabetes berdasarkan kriteria tertentu seperti indeks massa tubuh, usia, dll.
Wine (UCI *repository*)	178	13	Tidak	Penentuan tipe wine (anggur) berdasarkan kriteria tertentu seperti warna, kandungan kimia, dll.

a. *Dataset* Iris

Dataset Iris berisi pengelompokan bunga iris berdasarkan empat atribut yaitu *sepal length*, *sepal width*, *petal length* dan *petal width* ke dalam 3 *cluster* yaitu iris setosa, iris versicolour dan iris virginica.

b. *Dataset* Diabetes

Dataset Diabetes berisi pengelompokan pasien wanita berdasarkan 8 atribut yaitu *number of times pregnant*, *plasma glucose concentration*, *diastolic blood pressure*, *triceps skin fold thickness*, 2-*hour serum insulin*, *body mass index*, *diabetes pedigree function*, dan *age* ke dalam 2 *cluster* yaitu 0 yang

merepresentasikan bahwa pasien tidak memiliki kemungkinan terkena diabetes dan 1 yang merepresentasikan bahwa pasien memiliki kemungkinan terkena diabetes.

c. *Dataset* Wine

Dataset Wine berisi pengelompokan anggur berdasarkan 13 atribut yaitu *alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, Nonflavanoid phenols, proanthocyanins, color intensity, hue,* dan *proline* ke dalam 3 *cluster* yaitu 1, 2 dan 3.

2. Data Kasus

Penelitian ini berfokus pada *clustering* untuk menghasilkan *cluster* yang memiliki nilai DBI terbaik pada kasus strategi promosi universitas. Supaya promosi dapat dilakukan dengan tepat maka bagian marketing perlu mengetahui karakteristik calon mahasiswa di setiap wilayah. Bagian marketing perlu mengetahui jurusan apa yang paling diminati di setiap wilayah dan seperti apa potensi calon mahasiswa yang dimiliki di setiap wilayah. Untuk mengatahui informasi tersebut, bagian marketing dapat memanfaatkan data lulusan mahasiswa dengan cara mengelompokkan mahasiswa berdasarkan kemiripannya sehingga dapat diketahui karakteristik setiap kelompok mahasiswa yang terbentuk. Hal ini dapat dilakukan melalui *clustering*. Informasi yang diperoleh dari setiap kelompok dapat dijadikan bahan pengambilan keputusan untuk strategi promosi universitas. Data yang digunakan adalah data lulusan mahasiswa yang berjumlah 768 data dengan atribut yang digunakan adalah kota asal, jurusan dan IPK. Contoh model data lulusan mahasiswa dapat dilihat di Tabel 3.2.

Tabel 3.2 Model Data Lulusan Mahasiswa

NIM	Kota Asal	Jurusan	IPK
111	Bandung	Teknik Informatika	3.3
112	Bandung	Sistem Informasi	3.42
113	Jakarta	Teknik Informatika	2.98
114	Surabaya	Desain Komunikasi Visual	2.98
115	Medan	Teknik Informatika	2.73
116	Palembang	Teknik Industri	2.83
117	Makassar	Sistem Informasi	3.54

Sebelum dilakukan pengolahan data menggunakan teknik *clustering*, terlebih dahulu dilakukan proses transformasi data. Agar data siap diolah dengan menggunakan teknik *clustering*, maka data yang berjenis data nominal seperti kota asal dan jurusan harus diubah ke dalam bentuk angka menggunakan langkah-langkah sebagai berikut:

1. Pada data kota asal terlebih dulu dilakukan pembagian wilayah yang menjadi beberapa bagian wilayah. Misalkan wilayah Jawa Barat terdiri dari kota Bandung, Bekasi, Bogor, Ciamis, dll. Pembagian wilayah dibagi menjadi 33 wilayah.
2. Wilayah tersebut diurutkan berdasarkan frekuensi mahasiswa yang berasal dari daerah tersebut dari yang paling banyak hingga yang paling sedikit.
3. Wilayah yang memiliki frekuensi terbanyak diberi inisial dengan angka 1, wilayah dengan frekuensi terbesar kedua diberi inisial angka 2, dan begitu seterusnya hingga wilayah dengan frekuensi paling sedikit.

Contoh hasil tranformasi atribut kota asal dapat dilihat di Tabel 3.3.

Tabel 3.3 Contoh Hasil Transformasi Atribut Kota Asal

Wilayah	Frekuensi	Inisial
Jawa Barat 1	35	1
DKI Jakarta	33	2
Jawa Tengah	32	3
Sumatera Selatan	33	4
Sulawesi Selatan	29	5
Sumatera Utara	27	6
Jawa Timur	28	7

Selain atribut kota asal, atribut jurusan pun diubah ke dalam bentuk angka dengan cara yang sama seperti atribut kota asal. Contoh hasil transformasi atribut jurusan dapat dilihat di Tabel 3.4.

Tabel 3.4 Contoh Hasil Transformasi Atribut Jurusan

Jurusan	Frekuensi	Inisial
Teknik Informatika	49	1
Sistem Informasi	45	2
Akuntansi	43	3
Manajemen	42	4
Teknik Komputer	40	5
Desain Komunikasi Visual	39	6
Teknik Industri	39	7

Setelah dilakukan transformasi pada atribut kota asal dan jurusan maka hasil data yang didapat adalah seperti yang terlihat di Tabel 3.5.

Tabel 3.5 Contoh Hasil Transformasi Data Lulusan Mahasiswa

NIM	Kota Asal	Jurusan	IPK
111	15	2	3.3
112	12	1	3.42
113	29	5	2.98
114	1	20	2.98
115	28	18	2.73
116	25	20	2.83
117	23	18	3.54
118	5	3	3.7
119	21	1	2.96
120	9	18	3.45

Tabel 3.5 memperlihatkan data hasil transformasi yang siap diolah dengan menggunakan algoritma *clustering*.

3.2.2 Alur Pengujian

Pengujian dibagi menjadi dua bagian yaitu pengujian karakteristik algoritma dengan menggunakan data model dan data testing serta pengujian data kasus dengan menggunakan data kasus.

A. Pengujian Karakteristik Algoritma

Pengujian karakteristik algoritma dibagi menjadi dua bagian yaitu pengujian terhadap data model dan pengujian terhadap data testing. Pengujian dilakukan dengan cara mengimplementasikan dan membandingkan kinerja algoritma *clustering* yang diusulkan yaitu gabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means (HA-EKM) dengan gabungan algoritma sebelumnya yaitu Hierarchical Agglomerative Clustering dan K-Means (HA-KM). Sebagai tambahan, dilakukan perbandingan juga terhadap gabungan algoritma Hierarchical Agglomerative Clustering dan K-Means Particle Swarm Optimization (HA-KPSO). Algoritma K-Means dan K-Means Particle Swarm Optimization (KPSO) selengkapnya dapat dilihat di lampiran.

1. Pengujian Data Model

Pengujian data model dilakukan dengan cara mengimplementasikan algoritma *clustering* terhadap data model buatan lalu melakukan evaluasi hasil *clustering* dengan menggunakan Davies-Bouldin Index (DBI). Hal ini dilakukan untuk mengetahui bagaimana kinerja algoritma berdasarkan teknik evaluasi yang berbeda. Sehingga dapat diketahui kecocokan antara algoritma, data, dan teknik evaluasi hasil *clustering* yang digunakan. Tahapan pengujian data model adalah sebagai berikut:

A. Melakukan implementasi algoritma *clustering* dan evaluasi hasil *clustering*

Implementasi algoritma dan evaluasi hasil *clustering* dilakukan berdasarkan jumlah atribut dan jumlah data pada data 2 *cluster* dan 3 *cluster*. Evaluasi hasil *clustering* dilakukan dengan cara menghitung nilai DBI dari setiap hasil *clustering* pada data model untuk setiap jumlah data di setiap jumlah atribut. Selain itu, untuk mengetahui *cost* yang harus dikeluarkan masing-masing algoritma dalam membentuk hasil *clustering* digunakan jumlah iterasi sebagai pembanding tambahan .

B. Melakukan analisis dengan cara membandingkan nilai kinerja dan iterasi

Semakin kecil nilai DBI menunjukkan semakin baik hasil *clustering* tersebut, sebaliknya semakin besar nilai DBI menunjukkan semakin buruk hasil *clustering* yang dihasilkan. Jumlah iterasi menunjukkan *cost* yang diperlukan untuk menghasilkan hasil *clustering*, sehingga semakin sedikit jumlah iterasi menunjukkan semakin baik algoritma tersebut. Hasil pengujian data model dibahas di BAB 4 subbab 4.1.1.

2. Pengujian Data Testing

Pengujian data testing juga dilakukan dengan cara mengimplementasikan algoritma *clustering* terhadap data testing lalu melakukan evaluasi hasil *clustering* dengan menggunakan Davies-Bouldin Index (DBI). Pengujian data testing

dilakukan dengan tahapan yang sama seperti pada pengujian data model. Hasil pengujian data testing dibahas di BAB 4 subbab 4.1.2.

B. Pengujian Data Kasus

Pengujian data kasus dilakukan dengan cara mengimplementasikan dan membandingkan kinerja algoritma yang diusulkan yaitu gabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means (HA-EKM) dengan gabungan algoritma sebelumnya yaitu Hierarchical Agglomerative Clustering dan K-Means (HA-KM) pada data lulusan mahasiswa. Setelah itu, dilakukan evaluasi hasil *clustering* dengan menggunakan Davies-Bouldin Index (DBI). Melalui nilai DBI yang didapat, dapat diketahui algoritma mana yang dapat memberikan hasil *clustering* yang paling baik pada data lulusan mahasiswa. Selanjutnya, dilakukan analisis hasil *clustering* untuk menentukan strategi promosi universitas dan menentukan algoritma mana yang memberikan hasil *clustering* yang paling baik untuk strategi promosi universitas sehingga promosi dapat dilakukan secara tepat. Hasil pengujian data kasus dibahas di BAB 4 subbab 4.2.