BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam mempersiapkan penerimaan mahasiswa baru, universitas swasta memerlukan strategi khusus untuk menarik minat calon mahasiswa. Salah satunya adalah dengan melakukan promosi. 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 data lulusan mahasiswa berdasarkan kemiripan atribut datanya. Dengan begitu dapat dihasilkan pengelompokkan yang tepat sehingga dapat diketahui karakteristik setiap kelompok mahasiswa yang terbentuk untuk menentukan strategi promosi pada masing-masing kelompok. Hal ini dapat dilakukan melalui *clustering*. *Clustering* merupakan metode *data mining* untuk mengelompokkan sejumlah data atau obyek ke dalam *cluster* sehingga setiap anggota dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan obyek dalam *cluster* yang lainnya (Wirawan, Indwiarti, & Sibaroni, 2015). Kualitas hasil *clustering* akan sangat bergantung pada data dan algoritma yang digunakan sebab setiap algoritma memiliki karakteristikya masing-masing yang cocok dengan data yang berbeda. Sehingga diperlukan kecocokan antara algoritma dan data yang akan diproses untuk membentuk hasil *clustering* yang baik. Kualitas hasil *clustering* dapat diukur salah satunya dengan menggunakan Davies-Bouldin Index (DBI). Nilai DBI yang semakin mendekati 0 (nol) menandakan kualitas hasil *clustering* yang terbentuk adalah semakin baik.

Pemanfaatan *clustering* untuk strategi promosi universitas, salah satunya dilakukan oleh Ong (Ong, 2013) yang memanfaatkan data lulusan mahasiswa dengan atribut yang terdiri dari jurusan, kota asal dan IPK. Algoritma yang

digunakan pada penelitian tersebut adalah algoritma K-Means. Algoritma K-Means merupakan algortima berbasis partisi yang mengelompokkan data ke sejumlah *cluster* yang telah ditentukan sebelumnya. Penentuan jumlah *cluster* ini dapat disesuaikan dengan kebutuhan pihak marketing. Akan tetapi, jumlah *cluster* tersebut tidak dapat dipastikan merupakan jumlah *cluster* terbaik yang sesuai dengan karakteristik data lulusan mahasiswa yang ada. Hal tersebut merupakan kekurangan dari algoritma partisi. Selain itu kekurangan algoritma partisi terletak pada inisialisasi *centroid* (pusat *cluster*) awal (Yedla, Pathakota, & Srinivasa, 2010), seperti pada algoritma K-Means. Hal ini menyebabkan hasil *clustering* berupa solusi yang sifatnya *local minimum*, artinya hasil *clustering* yang terbentuk tidak dapat dijamin merupakan hasil yang terbaik, bergantung pada pemilihan pusat *cluster* awal. Pemilihan pusat *cluster* awal yang tidak tepat pada data lulusan mahasiswa dapat menyebabkan hasil pengelompokan data mahasiswa menjadi salah sehingga strategi promosi menjadi tidak tepat sasaran. Selain algoritma partisi, algoritma hierarki juga dapat digunakan untuk menyelesaikan kasus *clustering*. *Clustering* berbasis hierarki tidak memerlukan penentuan jumlah *cluster* di awal proses (Paithankar & Tidke, 2014). Sebaliknya, algoritma hierarki dapat menduga jumlah *cluster* terbaik dari suatu kumpulan data (*dataset*) untuk digunakan dalam *clustering* berbasis partisi (Jung, Park, Du, & Drake, 2002). Melalui algoritma hierarki dapat diketahui jumlah pengelompokkan terbaik dari data lulusan mahasiswa yang sesuai dengan karakteristik data yang ada. Akan tetapi, penambahan jumlah data dan jumlah atribut data dapat menyebabkan akurasi hasil *clustering* algoritma hierarki semakin buruk (Paithankar & Tidke, 2014).

Berbagai penelitian telah dilakukan untuk menutupi kekurangan pada algoritma *clustering*, misalnya dengan menggabungkan algoritma hierarki dan partisi untuk meningkatkan kualitas hasil *clustering*. Seperti yang dilakukan oleh Chen, dkk (Chen, Tai, Harisson, & Pan, 2005) yang menggabungkan algoritma Hierarchical Agglomerative Clustering (HAC) dengan K-Means dan Jung, dkk. (Jung, Park, Du, & Drake, 2002) yang menerapkan *clustering gain* pada algoritma hierarki untuk menentukan jumlah *cluster* terbaik sebagai inputan jumlah *cluster*

pada algoritma partisi. Milligan, dkk (Milligan & Cooper, 1985) menggunakan validasi *cluster* untuk menentukan jumlah *cluster* terbaik dari suatu *dataset*. Penelitian lain juga dilakukan untuk memperbaiki kekurangan pada *clustering* berbasis partisi, seperti yang dilakukan oleh Yelda, dkk (Yedla, Pathakota, & Srinivasa, 2010) yang mengajukan pendekatan baru yaitu algoritma Enhanced K-Means yang mengkombinasikan metode sistematis untuk mencari pusat *cluster* awal dan memberi cara yang efisien untuk memasukkan data ke dalam *cluster*. Algoritma yang diajukan terbukti dapat meningkatkan akurasi hasil *clustering*. Akan tetapi algoritma ini memiliki keterbatasan yaitu masih memerlukan inputan jumlah *cluster*, sehingga tidak menjamin jumlah *cluster* yang terbentuk adalah jumlah *cluster* yang terbaik.

Dalam menyelesaikan kasus *clustering* untuk strategi promosi universitas diperlukan algoritma yang tepat yang sesuai dengan karakteristik data lulusan mahasiswa, sehingga dapat dihasilkan hasil strategi promosi yang tepat. Berdasarkan hal yang telah diuraikan sebelumnya, dapat diketahui bahwa algoritma Enhanced K-Means dapat mengatasi permasalahan inisialisasi pusat *cluster* awal sedangkan pada penggabungan algoritma HAC dan K-Means diketahui bahwa algoritma HAC mampu menduga jumlah *cluster* terbaik dari suatu *dataset*, akan tetapi penggabungan kedua algoritma ini masih memiliki kekurangan yaitu tidak mampu mengatasi penentuan pusat *cluster* awal yang disebabkan karena algoritma K-Means tidak dapat menentukan pusat *cluster* awal secara otomatis. Berdasarkan hal tersebut, penggabungan algoritma HAC dan Enhanced K-Means dinilai dapat memberikan hasil *clustering* yang lebih baik pada kasus strategi promosi universitas. Hal ini disebabkan karena algoritma HAC dan Enhanced K-Means dapat mengatasi penentuan pusat *cluster* awal dan penentuan jumlah *cluster*. Sehingga gabungan algoritma HAC dan Enhanced K-Means dapat menghasilkan *cluster* yang antar anggota *cluster*-nya memiliki tingkat kemiripan yang tinggi pada *clustering* data lulusan mahasiswa untuk strategi promosi universitas.

1.2 Identifikasi Masalah

Permasalahan yang dikaji pada penelitian ini adalah cara menggabungkan algoritma Hierarchical Agglomerative Clustering (HAC) dan Enhanced K-Means untuk menghasilkan *cluster* yang memiliki nilai DBI terbaik pada data lulusan mahasiswa untuk strategi promosi universitas. Algoritma HAC memiliki keunggulan yaitu dapat menduga jumlah *cluster* terbaik dari suatu *dataset*. Sementara itu, algoritma Enhanced K-Means dapat menentukan pusat *cluster* awal secara otomatis dan memiliki akurasi yang baik dalam pengelompokkan data. Dengan menggabungkan kelebihan masing-masing algoritma, diharapkan gabungan algoritma HAC dan Enhanced K-Means mampu untuk menghasilkan *cluster* yang memiliki nilai DBI terbaik pada data lulusan mahasiswa untuk strategi promosi universitas. Pada penelitian sebelumnya yaitu pada penggabungan algoritma HAC dan K-Means (Chen, Tai, Harisson, & Pan, 2005)` maupun Enhanced K-Means (Yedla, Pathakota, & Srinivasa, 2010), tidak disebutkan bagaimana karakteristik algoritma dalam menangani variasi data.

Berdasarkan permasalahan tersebut maka hal berikutnya yang dilakukan dalam penelitian ini adalah meneliti karakteristik algoritma dalam menangani data dengan variasi jumlah data, jumlah atribut dan jumlah *cluster* menggunakan data model. Sehingga dapat diketahui kasus terbaik dan terburuk algoritma dapat diimplementasikan. Selain data model, pengujian dilakukan juga pada data testing yang umum digunakan dalam penelitian di bidang *data mining clustering* yang diperoleh dari UCI *machine learning*, yaitu repositori yang menyediakan *dataset* yang digunakan untuk penelitian di bidang *data mining*. Setelah diketahui karakteristik gabungan algoritma HAC dan Enhanced K-Means, hal yang dilakukan selanjutnya adalah melakukan implementasi algoritma pada model data lulusan mahasiswa untuk strategi promosi universitas. Sehingga dapat diketahui algoritma yang paling tepat untuk menyelesaikan *clustering* pada data lulusan mahasiswa.

1.3 Premis

Berdasarkan pada permasalahan yang diuraikan pada latar belakang dan identifikasi masalah, diformulasikan premis-premis sebagai berikut:

1. Algoritma Hierarchical Agglomerative Clustering dapat menduga jumlah *cluster* terbaik yang dapat dibentuk dari suatu *dataset*. Akan tetapi, penambahan jumlah data dan jumlah atribut data dapat menyebabkan akurasi hasil *clustering* semakin buruk.
2. Algoritma Enhanced K-Means dapat menentukan pusat *cluster* awal secara otomatis dan memiliki akurasi yang baik dalam pengelompokkan data, akan tetapi harus mengetahui jumlah *cluster* yang akan dibentuk di awal proses.

1.4 Hipotesis

Penggabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means (HA-EKM) dapat menghasilkan *cluster* dengan nilai DBI paling baik pada kasus strategi promosi universitas.

1.5 Tujuan Penelitian

Tujuan dilakukannya penelitian ini adalah sebagai berikut:

1. Menggabungkan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means untuk menghasilkan *cluster* yang memiliki nilai DBI terbaik
2. Mengetahui karakteristik gabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means
3. Mengetahui algoritma yang tepat untuk menyelesaikan kasus pengelompokan data mahasiswa untuk strategi promosi universitas

1.6 Batasan Masalah

Pada penelitian ini, permasalahan dibatasi sebagai berikut:

1. *Clustering* dilakukan pada kasus strategi promosi universitas
2. Algoritma yang diuji adalah algoritma Hierarchical Agglomerative Clustering (HAC) dan Enhanced K-Means
3. Pengujian dilakukan dengan menggunakan data model dan data kasus yang selengkapnya dijelaskan pada BAB 3

1.7 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Bagi sisi keilmuan, penelitian ini dapat menghasilkan algoritma yang dapat meningkatkan kualitas hasil *clustering* dengan cara penambahan proses penentuan jumlah *cluster* dan pusat *cluster* awal secara otomatis pada suatu *dataset*.
2. Bagi bidang penelitian di bidang *data mining*, penelitian ini memberikan pengetahuan mengenai karakteristik algoritma dalam melakukan proses *clustering* pada karakteristik data tertentu sehingga dapat diimplementasikan pada kasus nyata yang tepat untuk menghasilkan hasil *clustering* terbaik.

1.8 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan tesis ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, identifikasi masalah, premis, hipotesis, tujuan penelitian, manfaat penelitian serta sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Bab ini menguraikan tentang dasar-dasar teori mengenai *clustering* dalam *data mining*, teknik-teknik dalam *data mining* serta evaluasi hasil *clustering*.

BAB 3 METODE PENELITIAN

Bab ini menguraikan tentang teknik penggabungan algoritma Hierarchical Agglomerative Clustering dan Enhanced K-Means serta langkah-langkah pengujian yang terdiri dari pengujian karakteristik algoritma dengan menggunakan data model serta menguji data kasus untuk mengetahui algoritma yang paling baik dalam menyelesaikan kasus *clustering* untuk strategi promosi universitas.

BAB 4 HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menguraikan tentang hasil penelitian yang disajikan dalam bentuk grafik-grafik serta analisis hasil penelitian.

BAB 5 KESIMPULAN DAN SARAN

Bab ini menguraikan tentang kesimpulan dari penelitian yang telah dilakukan serta saran-saran untuk penelitian selanjutnya.