IMPLEMENTASI ALGORITMA IMPROVED K-MEANS PADA PORTAL JURNAL INTERNASIONAL

SKRIPSI

Untuk memenuhi sebagian persyaratan
memperoleh gelar Sarjana Komputer

Disusun oleh:
Amelia Handana Putri
NIM: 105090607111031

PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS BRAWIJAYA
MALANG
2017

PENGESAHAN

IMPLEMENTASI ALGORITMA IMPROVED K-MEANS PADA PORTAL JURNAL INTERNASIONAL

SKRIPSI

Diajukan untuk memenuhi sebagian persyaratan
memperoleh gelar Sarjana Komputer

Disusun Oleh :
Amelia Handana Putri
NIM:105090607111031

Skripsi ini telah diuji dan dinyatakan lulus pada
2 Februari 2017
Telah diperiksa dan disetujui oleh:

Dosen Pembimbing I	Dosen Pembimbing II
Randy Cahya W, S.ST., M.Kom	M. Ali Fauzi, S.Kom, M.Kom
NIK: 201405 880206 1 001	NIK: 201502 890101 1 001

Mengetahui
Ketua Jurusan Teknik Informatika

Tri Astoto Kurniawan, S.T, M.T, Ph.D
NIP: 19710518 200312 1 001

PERNYATAAN ORISINALITAS

Saya menyatakan dengan sebenar-benarnya bahwa sepanjang pengetahuan saya, di dalam naskah skripsi ini tidak terdapat karya ilmiah yang pernah diajukan oleh orang lain untuk memperoleh gelar akademik di suatu perguruan tinggi, dan tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis disitasi dalam naskah ini dan disebutkan dalam daftar pustaka.

Apabila ternyata didalam naskah skripsi ini dapat dibuktikan terdapat unsur-unsur plagiasi, saya bersedia skripsi ini digugurkan dan gelar akademik yang telah saya peroleh (sarjana) dibatalkan, serta diproses sesuai dengan peraturan perundang-undangan yang berlaku (UU No. 20 Tahun 2003, Pasal 25 ayat 2 dan Pasal 70).

Malang, 27 Januari 2017

Amelia Handana Putri

NIM: 105090607111031

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah SWT, karena dengan rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul **"Implementasi Algoritma Improved K-Means pada Portal Jurnal Internasional"** dengan baik.

Shalawat serta salam selalu tercurahkan kepada Nabi Muhammad SAW, yang dengan semangat juangnya terus memberikan tauladan untuk tetap bersemangat dan tidak berputus ada dengan rahmat-Nya. Maka tanpa itu semua, niscaya Penulis tidak akan dapat menyelesaikan skripsi ini dengan baik dan tepat waktu.

Penulisan dan penyusunan laporan skripsi ini dapat terlaksana dengan baik karena adanya bantuan secara langsung maupun tidak langsung dari pihak tertentu diantaranya:

1. Randy Cahya W, S.ST., M.Kom. selaku dosen pembimbing I yang telah bijaksana dan sabar dalam membimbing dan menyalurkan ilmu kepada penulis serta semua waktu dan nasehat yang telah diberikan dalam proses penyelesaian skripsi ini.
2. M. Ali Fauzi, S.Kom, M.Kom. selaku dosen pembimbing II yang telah bijaksana dan sabar dalam membimbing dan menyalurkan ilmu kepada penulis serta semua waktu dan nasehat yang telah diberikan dalam proses penyelesaian skripsi ini.
3. Dian Eka Ratnawati, S.Si, M.Kom. selaku dosen penasehat akademik yang telah memberikan nasehat, bimbingan, saran, dan dukungan selama penulis menuntut ilmu.
4. Tri Astoto Kurniawan, S.T, M.T, Ph.D selaku Ketua Jurusan Teknik Informatika di Fakultas Ilmu Komputer Universitas Brawijaya.
5. Agus Wahyu Widodo, S.T, M.Cs selaku Ketua Program Studi Teknik Informatika di Fakultas Ilmu Komputer Universitas Brawijaya.
6. Segenap bapak dan ibu dosen yang telah mendidik dan mengenalkan ilmunya kepada penulis.
7. Segenap staff dan karyawan Fakultas Ilmu Komputer Universitas Brawijaya yang telah membantu kelancaran pengerjaan skripsi.
8. Bapak, Ibu dan seluruh keluarga tercinta yang selalu memberikan doa, kasih sayang dan motivasi baik moral mapun materi sehingga penulis dapat menyelesaikan pendidikan dengan baik.
9. Ryo Prayoga Purnama Putra, yang selalu ada untuk mendukung, menemani, dan memberi semangat.

10. Teman-teman seperjuangan Ilmu Komputer angkatan 2010 dan seluruh warga Fakultas Ilmu Komputer Universitas Brawijaya yang telah selalu bersama dalam perjalanan mencari ilmu.
11. Dan semua pihak lain yang telah membantu terselesaikannya skripsi ini yang tidak bisa penulis sebutkan satu per satu.

Semoga jasa dan amal baiknya mendapatkan balasan dari Allah SWT. Dengan segala kerendahan hati, penulis menyadari sepenuhnya bahwa skripsi ini masih jauh dari sempurna karena keterbatasan materi dan pengetahuan yang dimiliki penulis. Akhirnya semoga skripsi ini dapat bermanfaat dan berguna bagi pembaca terutama mahasiswa Fakultas Ilmu Komputer Universitas Brawijaya.

Malang, 27 Januari 2017

Penulis

ameliahandana@gmail.com

ABSTRAK

Amelia Handana Putri. 2017. Implementasi Algoritma Improved K-Means pada Portal Jurnal Internasional. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang. Dosen Pembimbing: Randy Cahya W. S.ST., M.Kom. dan M. Ali Fauzi, S.Kom, M.Kom.

Penelitian adalah hal utama yang membawa pengaruh cukup besar dalam perkembangan teknologi saat ini. Hasil riset yang telah dilakukan oleh peneliti mampu menjadi bahan referensi oleh peneliti lainnya. Pada umumnya, referensi tersebut berasal dari publikasi jurnal yang telah diterbitkan secara internasional. Saat ini belum banyak dilakukan pengelompokan jurnal berdasarkan kualitasnya. Adanya sistem pengelompokan portal jurnal diharapkan dapat meminimalisasi tingkat kesalahan dalam publikasi jurnal sehingga tidak terjebak dalam jurnal predator termasuk dalam beberapa tindakan yang kurang elegan dalam publikasi jurnal ilmiah, misalnya *citation cartel* dan *junk science.* Sehingga diperlukan suatu sistem yang mampu bekerja secara otomatis untuk mengelompokkan portal jurnal internasional yang dapat membantu peneliti tersebut. Teknik yang digunakan adalah *clustering* menggunakan metode Improved K-Means. Metode Improved K-Means melakukan pengembangan pada inisialisasi *centroid* awal dengan perhitungan jarak antar data. Pengujian dilakukan dengan *silhouette coefficient* dengan range nilai antara -1 sampai 1 dimana semakin tinggi nilai dari *silhouette coefficient* maka semakin bagus *cluster* bentukan dari metode ini. Pada implementasi ini dibandingkan pula hasil pengujian pada Improved K-Means terhadap hasil dari pengujian pada K-Means Standar sehingga dapat diketahui keefektifan dari kedua metode. Pada pengujian menggunakan portal jurnal internasional, penggunaan 2 *cluster* dan 198 data menghasilkan nilai *silhouette coefficient* paling tinggi yaitu 0,869. Penggunaan metode Improved K-Means pada beberapa kasus memiliki kestabilan pada nilai *silhouette* dibandingkan dengan penggunaan metode K-Means Standar dan nilai *silhouette coefficient* yang lebih tinggi.

Kata kunci: Jurnal, *Clustering*, K-Means, Improved K-Means, *Silhouette Coefficient.*

ABSTRACT

Amelia Handana Putri. 2017. Implementation of* Improved K-Means *Algorithm in International Journal Portal. *Faculty of Computer, Brawijaya University, Malang. Advisor: Randy Cahya W. S.ST., M.Kom. dan M. Ali Fauzi, S.Kom, M.Kom.*

Research is the main thing that brings considerable influence in the development of today's technology. Results of research conducted by the researchers were able to be a reference by other researchers. In general, the reference is derived from the publication of journals have been published internationally. Little has been done grouping journals based on its quality. A system of grouping journal portal is expected to minimize the error rate in the publication of journals that do not get stuck in predatory journals included in some of the actions that are less elegant in the publication of scientific journals, such as citation cartel and junk science. So, we need a system that can work automatically to classify international journal portal that can assist researchers. The technique used is clustering using K-Means Improved methods. Improved K-Means method to develop the initialization of initial centroid by calculating the distance between the data. Testing is done with a silhouette coefficient with a value range between -1 to 1 where the higher the value of the coefficient, the better silhouette cluster formation of these methods. In this implementation also compared results of tests on Improved K-Means of the results of testing on K-Means standards so it can know the effectiveness of both methods. In testing using the portal international journals, the use of two cluster and the data yielded values 198 silhouette highest coefficient is 0.869. Use of Improved K-Means method in some cases have stability in silhouette value compared with the use of K-Means method Standards and silhouette coefficient value is higher.

Keywords*: Journal, Clustering,* K-Means*,* Improved K-Means*, Silhouette Coefficient.*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR *SOURCECODE*

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Dewasa ini, perkembangan teknologi semakin luar biasa seperti halnya perkembangan ilmu pengetahuan dan internet. Dengan adanya perkembangan tersebut, membawa dampak yang cukup besar dalam berbagai bidang salah satunya bidang penelitian. Berbagai macam hasil penelitian yang dilakukan oleh peneliti banyak dipublikasikan dalam bentuk artikel ilmiah. Publikasi hasil penelitian secara digital ini mulai berkembang pesat seiring dengan meningkatnya penggunaan teknologi komputer, tak terkecuali peneliti yang mempublikasikan hasil penelitiannya. Hal ini menyebabkan banyak dokumen-dokumen secara digital tersebar luas di internet terutama di portal-portal jurnal internasional.

Berdasarkan data dari sebuah lembaga riset internasional (SCImago), publikasi ilmiah para peneliti (khususnya dosen) di Indonesia masih sedikit dibanding dengan negara lain. Hingga tahun 2014, Indonesia mencatat sebanyak 32.355 judul penelitian yang dipublikasikan di jurnal Internasional terakreditasi dan diindeks di Scopus. Sedangkan untuk Negara tetangga Malaysia 153.378 judul paper telah dipublikasikan di jurnal Internasional dan terkareditasi dan diindeks di Scopus (Scimagojr, 2016). Sedangkan pada data lainnya, negara Indonesia jika dibandingkan dengan negara tetangga Malaysia, jumlah dosen di Malaysia yang bergelar doktor 8.000 orang, ternyata bisa menghasilkan 14.103 publikasi setiap tahunnya. Sedangkan di Indonesia, jumlah total dosen di seluruh perguruan tinggi sebanyak 270.000 orang (23.000 diantaranya telah bergelar doktor) hanya menghasilkan publikasi sebanyak 1.975 buah, setiap tahunnya (M. Lutfi, 2012).

Dari data tersebut tidak banyak hasil riset akademis Indonesia yang dipublikasikan dan masuk jurnal internasional yang terpercaya. Para peneliti tersebut memiliki peran besar dalam publikasi informasi ilmiah dalam bentuk jurnal di portal jurnal internasional yang diharapkan dapat meningkatkan kuantitas hasil riset akademis di Indonesia. Sehingga hasil riset tersebut mampu menjadi bahan referensi atau rujukan oleh peneliti lainnya. Selain memilih portal jurnal internasional yang terpercaya sebagai tempat publikasi, peneliti juga harus selektif dalam memilih bahan referensi jurnal yang berkualitas sehingga penelitian yang dihasilkan juga optimal dan terpercaya.

Pada beberapa kasus, pengelompokan jurnal internasional sering kali mengelompokkan konten jurnal untuk memperoleh informasi dari jurnal tersebut. Pada saat ini belum banyak ditemukan pengelompokan portal jurnal berdasarkan kualitasnya seperti jumlah sitasi, jumlah dokumen dan sebagainya yang dapat memudahkan peneliti untuk mempublikasikan jurnal dan memilih referensi jurnal. Maka dari itu, diperlukan suatu sistem yang mampu bekerja secara otomatis untuk mengelompokkan portal jurnal yang dapat membantu para peneliti dalam mengembangkan riset baru, memilih referensi atau rujukan

hasil riset dan mempublikasikan hasil riset mereka. Dengan adanya sistem portal jurnal ini diharapkan dapat meminimalisir tingkat kesalahan dalam publikasi jurnal sehingga tidak terjebak dalam jurnal predator termasuk dalam beberapa tindakan yang kurang elegan dalam publikasi jurnal ilmiah, misalnya *Citation Cartel* dan *Junk Science*. Salah satu teknik yang dapat diimplementasikan ke dalam sistem tersebut adalah teknik pengelompokan (*clustering*) yang menggunakan metode Improved K-Means.

Clustering adalah salah satu teknik yang dikenal dalam data mining. *Clustering* merupakan suatu *unsupervised learning*, dimana sekelompok data langsung dikelompokkan berdasarkan tingkat kemiripannya tanpa dilakukan supervisi. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu *cluster* dan meminimumkan kesamaan antar anggota *cluster* yang berbeda. *Clustering* juga dapat mengelompokkan data berdasarkan tingkat kemiripannya dan juga berdasarkan tingkat akurasinya (Han, 2001). Secara umum, metode *clustering* dapat diklasifikasikan menjadi *partitioning methods* dan *hierarchical methods* (Oded, 2010). Pada proses pengelompokan ini dilakukan dengan memanfaatkan metode K-Means *Clustering.*

K-Means *Clustering* merupakan metode untuk mengklasifikasikan atau mengelompokkan objek-objek (data) ke dalam K-group *(cluster*) berdasarkan atribut tertentu. Pengelompokan data dilakukan dengan memperhitungkan jarak terdekat antara data dengan pusat *cluster* . Prinsip utama dari metode ini adalah menyusun nilai K buah *centroid* atau rata-rata *(mean)* dari sekumpulan data berdimensi N, dimana metode ini mensyaratkan nilai K sudah diketahui sebelumnya. Algoritma K-Means dimulai dengan pembentukan prototype *cluster* diawal kemudian secara iterative prototype *cluster* tersebut diperbaiki sehingga tercapai kondisi konvergen, yaitu kondisi dimana tidak terjadi perubahan yang signifikan pada prototype *cluster*. Perubahan ini diukur dengan menggunakan fungsi objektif D yang umumnya di defenisikan sebagai jumlah atau rata-rata jarak tiap item data dengan *centroid* groupnya.

Namun, K-Means mempunyai mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode K-Means ini sangatlah tergantung pada inisialisasi nilai pusat awal *cluster* yang diberikan. Hal ini menyebabkan hasil *cluster*nya berupa solusi yang sifatnya *local optimal*. Beberapa upaya dilakukan oleh peneliti untuk meningkatkan akurasi dan efisiensi dari algoritma K-Means. Improved K-Means adalah salah satu metode yang dapat diterapkan untuk permasalahan *clustering* (Carlisle, 2001). Dibandingkan dengan K-Means, Improved K-Means memiliki kelebihan yaitu lebih stabil dan kuat dibandingkan dengan metode K-Means Standar (Frans, 2009).

Berdasarkan latar belakang yang telah dikemukakan, pada skripsi ini mengangkat judul "Implementasi Algoritma Improved K-Means pada Portal Jurnal Internasional" yang diharapkan dapat membantu peneliti untuk mengembangkan riset baru, memilih referensi atau rujukan hasil riset dan mempublikasikan hasil riset mereka.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka perumusan permasalahan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan algoritma Improved K-Means pada Portal Jurnal Internasional.
2. Bagaimana hasil pengujian dari implementasi algoritma Improved K-Means pada Portal Jurnal Internasional.

1.3 Tujuan

Adapun tujuan dari penelitian ini adalah:

1. Mengimplementasikan algoritma Improved K-Means pada Portal Jurnal Internasional.
2. Mendapatkan hasil pengujian dari implementasi algoritma Improved K-Means pada Portal Jurnal Internasional.

1.4 Manfaat

Adapun manfaat dari penelitian ini adalah:

1. Meningkatkan kualitas jurnal yang akan diterbitkan peneliti serta memberikan dampak positif pada *H-index* dan *citation* dari setiap publikasi jurnal yang dibuat oleh seorang peneliti.
2. Meningkatkan kepercayaan dan nilai *impact* factor yang tinggi pada portal jurnal internasional.
3. Menunjang integritas terhadap penilaian evaluasi kinerja peneliti dari nilai *impact factor* jurnal yang layak, legal dan dapat dipertanggung jawabkan sesuai dengan etika akademis.
4. Meminimalisir tingkat kesalahan dalam melakukan publikasi jurnal, termasuk dalam beberapa tindakan yang kurang elegan dalam publikasi jurnal ilmiah, misalnya *Citation Cartel* dan *Junk Science.*
5. Meningkatkan kualitas jurnal yang akan diterbitkan oleh peneliti, karena akan diketahui secara jelas bahwa portal jurnal yang akan di-*submit* tersebut benar-benar berkualitas tinggi dan terpercaya.

1.5 Batasan Masalah

Dari permasalahan yang dirumuskan di atas, maka batasan permasalahan yang digunakan untuk merancang dan membuat sistem ini adalah sebagai berikut:

1. Data yang digunakan untuk analisis pada skripsi ini merupakan *dataset SCImagojr Journal* yang diambil dari http://www.scimagojr.com/journalrank.php dalam format *spreadsheet* yaitu *Microsoft Excel*.
2. *Dataset SCIMagojr Journal,* fitur yang diambil antara lain total jumlah dokumen jurnal (Total Docs. (3 years)), total semua dokumen yang disitasi

(buku, *conference,* jurnal, *trade* jurnal dan *book series*) atau Total Cites (3 years), dan total dokumen yang disitasi (khusus artikel, *review, conference paper*) atau Citable Docs. (3 years).

3. Data *SCImagojr Journal* yang digunakan adalah data *journal* ranking pada tahun 2012.
4. Metode *clustering* menggunakan Improved K-Means.
5. Pengujian untuk Improved K-Means menggunakan *silhouette coefficient.*

1.6 Sistematika pembahasan

Sistematika pembahasan dalam skripsi ini antara lain:

Bab 1 Pendahuluan

Bab ini membahas tentang latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah, dan sistematika pembahasan.

Bab 2 Landasan Kepustakaan

Bab ini berisi pustaka yang dikaji dan teori-teori dari berbagai pustaka yang menunjang penelitian ini. Berisi tentang teori-teori yang berkaitan dan menunjang penyelesaian proyek akhir ini.

Bab 3 Metodologi

Bab ini membahas tentang gambaran umum metode yang dipakai dalam penelitian yang terdiri dari studi pustaka, analisis kebutuhan, metode perancangan, metode implementasi, metode pengujian dan analisis, serta penulisan laporan penelitian.

Bab 4 Perancangan

Bab ini berisi rancangan arsitektur sistem yang lebih mendetil dari sistem *clustering* Portal Jurnal Internasional menggunakan Improved K-Means.

Bab 5 Implementasi

Bab ini membahas tentang implementasi dari sistem *clustering* Portal Jurnal Internasional menggunakan Improved K-Means.

Bab 6 Pengujian dan Evaluasi

Bab ini membahas tentang pengujian sistem *clustering* Portal Jurnal Internasional menggunakan Improved K-Means dan evaluasi dari hasil pengujian sistem tersebut.

Bab 7 Penutup

Bab ini membahas tentang kesimpulan yang diperoleh dari implementasi dan pengujian serta saran-saran untuk pengembangan lebih lanjut.

BAB 2 LANDASAN KEPUSTAKAAN

Pada bab ini, dideskripsikan mengenai teori yang berhubungan dengan *clustering* portal jurnal internasional menggunakan Improved K-Means. Dasar teori yang dibahas pada bab ini diantaranya adalah dasar teori tentang *SCImagojr Journal and Country Rank,* Jurnal Ilmiah, *Impact Factor, Data Mining, Clustering,* K-Means *Clustering,* Improved K-Means, Analisis *Cluster* dan *Silhouette Coefficient.*

2.1 SCImagojr Journal & Country Rank

The SCImagojr Journal & Country Rank merupakan sebuah portal yang mencakup indikator jurnal ilmiah dan negara maju dari informasi yang terdapat dalam basis data *Scopus*® (http://www.scopus.com/) yaitu Elsevier BV (http://www.elsevier.com/). Indikator ini dapat digunakan untuk menilai dan menganalisis *domain* ilmiah.

Nama untuk platform ini diambil dari *Scimago Journal Rank (SJR) indicator,* dikembangkan oleh Scimago dari algoritma yang dikenal secara luas yaitu *Google PageRank*™. Indikator tersebut menunjukkan visibilitas jurnal yang terdapat dalam database *Scopus*® mulai tahun 1996.

Scimago ialah sebuah kelompok riset dari *Consejo Superior de Investigaciones Cientificas (CSIC), University of Granada, Extremadura, Carlos III (Madrid)* dan *Alcala de Henares* yang didedikasikan untuk analisis informasi, representasi dan pengambilan melalui teknik visualisasi.

Sama halnya dengan *SJR Portal*, *Scimago* telah mengembangkan *The Atlas of Science Project,* yang mengusulkan pembentukan suatu sistem informasi dengan tujuan utama adalah untuk mencapai representasi grafis dari *IberoAmerican Science Research.* Perwakilan tersebut diartikan sebagai kumpulan peta interaktif, yang memungkinkan fungsi navigasi seluruh ruang semantik yang dibentuk oleh peta (Scimagojr, 2014).

Adapun beberapa *journal indicator* yang digunakan dalam penelitian ini yaitu (Scimagojr, 2014):

1. *H-Index*

   H-Index merupakan jumlah jurnal dari artikel (*h*) yang telah menerima setidaknya *h* kutipan. Hal itu diperoleh dengan mengkuantifikasi atau menjumlahkan antara produktivitas jurnal ilmiah dan dampak ilmiah dan juga berlaku untuk para ilmuwan, negara, dan sebagainya.

2. *Total Docs. (Total Documents)*

   Output pada jangka waktu atau periode yang dipilih. Semua jenis dokumen termasuk di dalamnya, baik dokumen yang bisa dan tidak bisa dikutip.

3. *Total Docs. (3 years)*

Dikumen yang diterbitkan 3 tahun terakhir (bukan termasuk tahun yang dipilih). Misalnya ketika tahun X dipilih, maka dokumen yang diterbitkan pada tahun X-1, X-2, X-3 akan diambil. Semua jenis dokumen termasuk di dalamnya, baik dokumen yang bisa dan tidak bisa dikutip.

4. *Total References*

   Mencakup semua referensi bibliografi dalam jurnal pada periode atau jangka waktu yang dipilih.

5. *Total Cites (3 years)*

   Jumlah kutipan yang diterima pada tahun yang terpilih oleh jurnal untuk dokumen yang diterbitkan dalam 3 tahun sebelumnya.

6. *Citable Documents*

   Jumlah kutipan dokumen yang dipublikasikan oleh sebuah jurnal dalam 3 tahun sebelumnya (bukan termasuk tahun dokumen yang dipilih). Secara eksklusif termasuk artikel, ulasan (*review*), dan makalah konferensi.

7. *Cites per Documents (2 years)*

   Rata-rata kutipan dokumen dalam jangka waktu 2 tahun. Hal ini dihitung dengan cara mempertimbangkan jumlah kutipan yang diterima oleh jurnal pada tahun saat itu dengan dokumen yang diterbitkan dalam 2 tahun sebelumnya. Contohnya kutipan yang diterima pada tahun X untuk dokumen yang dipublikasikan pada tahun X-1 dan X-2.

8. *References/ Documents*

   Jumlah rata-rata referensi tiap dokumen pada tahun yang dipilih.

2.2 Jurnal Ilmiah

Jurnal ilmiah dianggap sebagai sumber informasi yang paling penting di dunia ilmu pengetahuan dan teknologi. Jurnal ilmiah berisi kumpulan artikel yang dipublikasikan secara periodik, ditulis oleh ilmuwan peneliti untuk melaporkan hasil-hasil penelitian terbarunya. Maka dari itu, keberadaan jurnal ilmiah merupakan hal yang penting untuk memajukan ilmu pengetahuan dan teknologi. Tulisan yang dimuat dalam jurnal ilmiah, sudah mengalami proses *perr-review* dan seleksi ketat dari para pakar di bidangnya masing-masing. Proses *perr-review* dijalankan untuk menjamin kualitas dan validitas ilmiah artikel yang dimuat.

Publikasi jurnal merupakan bagian penting dari metode ilmiah. Tulisan dalam jurnal ilmiah ditujukan untuk para peneliti dan para ahli lainnya di bidang yang sama, Artikel dalam sebuah jurnal harus sedemikian jelas sehingga seorang peneliti independen dapat mengulangi percobaan atau perhitungannya untuk memverifikasi dhasil penelitiannya. Artikel jurnal akan menjadi bagian dari rekam ilmiah untuk selamanya (*permanent scientific record*).

Jurnal ilmiah mempunyai 3 (tiga)peran dalam proses komunikasi ilmiah:

1. Peran sosial.

Untuk membangun dan memelihara kekayaan intelektual, sehingga karya kreatif dan inovatif seorang ilmuwan akan mendapatkan pengakuan dari dunia disiplin ilmu terkait.

2. Peran arsip

   Untuk memberikan pengakuan ilmiah bahwa artikel yang diterbitkan itu sudah di evaluasi dan dinyatakan dapat diterima oleh dunia ilmu pengetahuan. Artikel yang dikirim ke jurnal ilmiah akan mengalami proses *peer review* yaitu proses seleksi dan review oleh para ahli di bidang tersebut untuk menuntun apakah karya tersebut memenuhi syarat keakuratan, reliabilitas dan layak untuk dipublikasikan. Proses ini ditujukan untuk menjaga kualitas literatur ilmiah sehingga hanya karya yang memenuhi syarat ilmiah yang dipublikasikan. Dengan demikian, peneliti lain akan mendapatkan keyakinan ketika menggunakan artikel dalam jurnal ilmiah sebagai dasar untuk mengembangkan karya yang lainnya.

3. Peran diseminasi

   Untuk penyebaran informasi dengan cepat dan meluas di kalangan publik. Kegiatannya dilakukan melalui pelatihan atau workshop, seminar, dan komunikasi. Penyebaran tersebut termasuk membuat tulisan tentang suatu topik untuk dimuat dalam sebuah jurnal ilmiah atau buletin yang diterbitkan sendiri atau instansi, lembaga, organisasi lain, atau dikirim ke redaksi suatu penerbitan media cetak (Manadokota, 2013). Peran diseminasi informasi sangat esensial kare sifat dari ilmu pengetahuan yang kumulatif (terus bertambah). Apalagi dengan kemajuan publikasi *online,* maka diseminasi dari publikasi ilmiah berpotensi untuk dapat dilakukan dengan semakin cepat.

Maka dapat disimpulkan bahwa jurnal ilmiah memiliki peranan yang sangat penting untuk perkembangan ilmu pengetahuan, sebagai berikut (Pustakaristek, 2016):

1. Bagi peneliti yang karyanya dimuat di sebuah jurnal ilmiah internasional,hal itu merupakan pengakuan tertinggi dari dunia ilmiah bahwa karyanya memamng berkualitas, ememnuhi syarat keakuratan, reliabilitas, validitas dan originalitas.
2. Untuk peneliti lain, jurnal ilmiah adalah referensi terkini dari kemajuan ilmu pengetahuan dan teknolohi di bidang keilmuannya. Agar penelitian tetap tersambung dengan kemajuan terkini, maka setiap peneliti harus mengetahui publikasi jurnal ilmiah untuk mencegah jangan sampai penelitiannya itu merupakan duplikasi penelitian yang telah dilakukan orang lain atau merupakan penelitian yang sudah *out of date,* sehingga bisa menjaga bahwa penelitiannya tetap sejalan dengan perkembangan terkini.

2.3 Impact Factor

Impact Factor (IF) oleh sebagian besar peneliti di dunia, dijadikan ukuran kualitas suatu jurnal. Semakin tinggi IF-nya maka jurnal tersebut akan semakin bergengsi. IF diperkenalkan oleh ISI *Web of Science*, sebuah perusahaan di bidang pendidikan.

Penghitungan IF untuk sebuah jurnal (misalnya untuk tahun 2010) dilakukan dengan cara menjumlahkan rata-rata *citation*/rujukan setiap karya yang diterbitkan pada 2 tahun sebelumnya (yaitu tahun 2009 dan 2008) dibandingkan dengan jumlah seluruh karya (yang bisa dirujuk) yang terbit pada tahun 2006 dan 2007. Misal suatu jurnal yang mempunyai nilai IF = 3 berarti setiap paper yang terbit pada tahun 2007 dan 2008 dirujuk oleh rata-rata 3 buah karya ilmiah lainnya. IF tahun 2010 akan keluar pada tahun 2011, dan seterusnya (Rossner, 2007).

Jurnal-jurnal yang memiliki IF tinggi, diantaranya adalah *Nature* (30,98) dan *Science* (29,78) dengan data pada tahun 2012. Perlu diketahui pula bahwa jumlah jurnal yang mempunyai IF lebih besar dari 5,5 sangatlah sedikit, kebanyakan jurnal (>80%) mempunyai IF pada rentang 0,3 sampai dengan 5,5. Sistem perangkingan jurnal dengan metode IF ini sudah digunakan selama lebih dari 40 tahun (Garfield, 2006).

Selain IF, ada juga metode perhitungan tandingan yang dibuat oleh lembaga lain, contoh yang paling popular adalah SJR indicator (http://www.scimagojr.com/) yang dikembangkan oleh Scimago Laboratory. Indikator SJR (*Scimago Journal and Country Rank*) merupakan ukuran suatu jurnal dilihat dari 2 sisi yang berbeda, yaitu (1) jumlah *citation* / rujukan yang diterima oleh jurnal tersebut, dan (2) tingkat *prestise* / gengsi dari jurnal yang mengutip tersebut. Semakin tinggi nilai SJR, maka jurnal tersebut semakin bagus, berkualitas dan bereputasi. Berbeda dengan *Impact Factor*, indikator SJR merupakan sistem perangkingan jurnal yang bersifat *open access*, sehingga lebih cocok digunakan di negara-negara berkembang seperti Indonesia. Selain itu, masih banyak kelebihan lain dari Indikator SJR dari pada IF. Karena Indikator SJR baru dikembangkan pada tahun 2007, maka teknik perhitungan untuk merangking suatu jurnalnya pun lebih canggih dan komprehensif. Algoritma perhitungan indikator SJR berdasarkan pada *Google Pagerank* ditambah dengan *database* dari Scopus. Indikator SJR juga menyediakan beberapa teknik perhitungan yang berbeda, diantaranya adalah '*Cites per Doc (2y)*' yang metode perhitungannya sama dengan IF. Indikator SJR juga unggul dalam berbagai karakteristik penilaian lainnya (Falagas, 2008). Oleh karena itu, dianjurkan agar mulai menggunakan Indikator SJR sebagai pengganti *Impact Factor* (M. Lutfi, 2012).

2.4 Data Mining

2.4.1 Pengertian Data Mining

Data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang

tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Kusrini, 2009). Definisi lain data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, 2005)

Berdasarkan definisi di atas dapat disimpulkan bahwa data mining adalah suatu proses mengekstraksi dan mengidentifikasi informasi dari database yang besar menggunakan teknik statistik, kecerasan buatan, dan *machine learning.*

2.4.2 Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Kusrini, 2009):

1. Deskripsi

   Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variable target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variable target sebagai nilai prediksi. Selanjutnya pada peninjauan berikutnya estimasi nilai dari variable target dibuat berdasarkan nilai variable prediksi.

3. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi. Kecuali bahwa dalam prediksi nilai dari hasil aka nada di masa mendatang.

4. Klasifikasi

   Dalam klasifikasi terdapat target variable kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, sedang, dan rendah.

5. Pengklusteran

   Pengklusteran merupakan pengelompokkan record, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variable target dalam pngklusteran. Pengklusteran tidak mencoba melakukan klasifikasi, estimasi, atau memprediksi nilai dari variable target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang

mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

   Tugas asosiasi adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.4.3 Proses Data Mining

Proses dari *data mining* mempunyai prosedur umum dengan langkah-langkah sebagai berikut (Kantardzic, 2003):

1. Merumuskan permasalahan dan hipotesis

   Pada langkah ini dispesifikasikan sekumpulan variabel yang tidak diketahui hubungannya dan jika memungkinkan dispesifikasikan bentuk umum dari keterkaitan variabel sebagai hipotesis awal.

2. Mengoleksi data

   Langkah ini menitikberatkan pada cara bagaimana data dihasilkan dan dikoleksi. Secara umum ada dua kemungkinan yang berbeda. Yang pertama adalah ketika proses pembangkitan data dibawah kendali dari ahli. Pendekatan ini disebut juga dengan percobaan yang dirancang (*designed experiment*). Kemungkinan yang kedua adalah ketika ahli tidak memiliki pengaruh pada proses pembangkitan data, dikenal sebagai pendekatan observasional.

3. Pra pengolahan data

   Pra pengolahan data memiliki dua tugas utama yaitu:

   a. Deteksi dan pembuangan data asing (*outlier*)

      Data asing merupakan data dengan nilai yang tidak dibutuhkan karena tidak konsisten pada sebagian pengamatan. Biasanya data asing dihasilkan dari kesalahan pengukuran, kesalahan pengkodean, dan pencatatan beberapa nilai abnormal yang wajar. Ada dua strategi untuk menangani data asing, yang pertama mendeteksi dan berikutnya membuang data asing sebagai bagian dari fase pra pengolahan. Yang kedua adalah mengembangkan metode permodelan yang kuat yang tidak merespon data asing.

   b. Pemberian skala, pengkodean, dan seleksi fitur

      Pra pengolahan data menyangkut beberapa langkah seperti memberikan skala variabel dan beberapa jenis pengkodean. Sebagai contoh, satu fitur dengan *range* [0, 1] dan yang lain dengan *range* [-100, 100] tidak akan memiliki bobot yang sama pada teknik yang diaplikasikan dan akan berpengaruh pada hasil akhir *data mining*. Oleh karena itu, disarankan

untuk pemberian skala dan membawa fitur-fitur tersebut ke bobot yang sama untuk analisis lebih lanjut.

4. Mengestimasi model
   Pemilihan dan implementasi dari tehnik *data mining* yang sesuai merupakan tugas utama dari fase ini. Proses ini tidak mudah, biasanya dalam pelatihan, implementasi berdasarkan pada beberapa model dan pemilihan model yang terbaik merupakan tugas tambahan.
5. Menginterprestasikan model dan menarik kesimpulan
   Pada banyak kasus, model *data mining* akan membantu dalam pengambilan keputusan. Metode *data mining* modern diharapkan akan menghasilkan hasil akurasi yang tinggi dengan menggunakan model dimensi-tinggi.

Pengetahuan yang baik pada keseluruhan proses sangat penting untuk kesuksesan aplikasi. Tidak perduli seberapa kuat metode *data mining* yang digunakan, hasil dari model tidak akan valid jika pra pengolahan dan pengkoleksian data tidak benar atau jika rumusan masalah tidak berarti.

2.5 Clustering

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah *dataset* ke dalam *subset*, sehingga data dalam setiap *subset* memiliki arti yang bermanfaat. Dimana sebuah *cluster* terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada *cluster* lainnya. Algoritma *clustering* terdiri dari dua bagian yaitu secara *hirarkis* dan secara *partitional*. Algoritma *hirarkis* menemukan *cluster* secara berurutan dimana *cluster* ditetapkan sebelumnya, sedangkan algoritma *partitional* menentukan semua kelompok pada waktu tertentu. *Clustering* juga bisa dikatakan suatu proses dimana mengelompokan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokan pada *cluster* yang sama dan memisahkan diri dengan membentuk pola yang berbeda ke *cluster* yang berbeda (Berkhin, 2002).

Clustering dapat memainkan peran penting dalam kehidupan sehari-hari, karena tidak bisa lepas dengan sejumlah data yang menghasilkan informasi untuk memenuhi kebutuhan hidup. Salah satu sarana yang paling penting dalam hubungan dengan data adalah untuk mengklasifikasikan atau mengelompokkan data tersebut ke dalam seperangkat kategori atau *cluster*. *Clustering* dapat ditemukan dibeberapa aplikasi yang ada di berbagai bidang. Sebagai contoh pengelompokan data yang digunakan untuk menganalisa data statistik seperti pengelompokan untuk pembelajaran mesin, data mining, pengenalan pola, analisis citra dan *bioinformatika* (Berkhin, 2002).

Teknik pengelompokan saat ini dapat diklasifikasikan menjadi tiga kategori yaitu partitional, hirarkis dan berbasis lokalitas algoritma. Terdapat satu set objek dan kriteria *clustering* atau pengelompokan, pengelompokan partitional mememperoleh partisi objek ke dalam *cluster* sehingga objek dalam *cluster* akan lebih mirip dengan benda-benda yang ada di dalam *cluster* dari pada objek yang terdapat pada *cluster* yang berbeda. Partitional mencoba untuk

menguraikan *dataset* ke satu set *cluster* dengan menentukan jumlah *cluster* awal yang diinginkan (Osmar, 1999).

2.6 K-Means *Clustering*

2.6.1 Definisi K-Means

K-Means merupakan salah satu metode data klustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/ kelompok. Metode ini mempartisi ke dalam *cluster*/ kelompok sehingga data yang memiliki karakteristik yang sama (*High intra class similarity*) dikelompokkan ke dalam satu *cluster* yang sama dan yang memiliki karakteristik yang berbeda (*Low inter class similarity*) dikelompokkan pada kelompok yang lain (Giyanto, 2008). Proses *clustering* dimulai dengan mengidentifikasi data yang akan dikluster, X_{ij} (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan di *cluster* dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap *cluster* ditetapkan secara bebas (sembarang), C_{kj} (k=1,...,k+1; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan setiap pusat *cluster*. Untuk melakukan penghitungan jarak data ke-i (x_i) pada pusat *cluster* ke-k (c_k), diberi nama (di_k), dapat digunakan formula *Euclidean* seperti pada persamaan (2.1), yaitu:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - {c_{kj}}^2)} \qquad (2.1)$$

Suatu data akan menjadi anggota dari *cluster* ke-*k* apabila jarak data tersebut ke pusat *cluster* ke-*k* bernilai paling kecil jika dibandingkan dengan jarak ke pusat *cluster* lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (2.2). Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap *cluster*.

$$\text{Min } \sum_{k=1}^{k+1} d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - {c_{kj}}^2)} \qquad (2.2)$$

Nilai pusat *cluster* yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada *cluster* tersebut, dengan menggunakan rumus pada persamaan (2.3):

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; \qquad (2.3)$$

Dimana x_{ij} $\in$ kluster ke-*k*

P = banyaknya anggota *cluster* ke-*k*

Algoritma dasar dalam K-Means adalah:

1. Tentukan jumlah *cluster* (*k*), tetapkan pusat *cluster* sembarang .
2. Hitung jarak setiap data ke pusat *cluster* menggunakan persamaan (2.1).
3. Kelompokkan data ke dalam *cluster* yang dengan jarak yang paling pendek menggunakan persamaan (2.2).

4. Hitung pusat *cluster* yang baru menggunakan persamaan (2.3)

Ulangi langkah 2 sampai dengan 4 hingga sudah tidak ada lagi data yang berpindah ke kluster yang lain.

2.6.2 Keuntungan Algoritma K-Means

Algoritma K-Means juga memiliki keuntungan yaitu :

1. Dalam implementasi menyelesaikan masalah, algoritma K-Means sangat *simple* serta *fleksibel*. Artinya perhitungan komputasinya tidak terlalu rumit dan algoritma ini dapat diimplementasikan pada segala bidang.
2. Algoritma K-Means sangat mudah untuk dipahami, terutama dalam implementasi data yang sangat besar serta dapat mengurangi kompleksitas data yang dimiliki (Berkhin, 2002).

2.6.3 Kelemahan Algoritma K-Means

Kelemahan yang dimiliki oleh algoritma K-Means yaitu :

1. Di Algoritma K-Means user memerlukan angka yang tepat dalam menentukan jumlah *cluster* sebanyak k karena terkadang pusat *cluster* awal dapat berubah sehingga kejadian ini bisa mengakibatkan pengelompokan data menjadi tidak stabil (Joshi, 2013).
2. Algoritma K-Means tidak bisa maksimal dalam menentukan atau menginisialkan nilai *centroid* awalnya, karena pada pengelompokan data dengan algoritma K-Means sangat bergantung pada nilai *centroid*. (Ahmed, 2011).
3. Output dari K-Means tergantung pada nilai - nilai pusat yang dipilih pada *clustering*. Sehingga pada algoritma ini nilai awal titik pusat *cluster* menjadi dasar dalam penentuan *cluster*. Pemilihan *centroid cluster* awal secara acak akan memberikan pengaruh terhadap kinerja *cluster* tersebut (Ahmed, 2011).

2.7 Improved K-Means

Improved K-Means merupakan modifikasi algoritma K-Means Standar untuk meningkatkan akurasi dan efisiensi. Metode modifikasi ini diuraikan sebagai berikut (Nazeer, 2009):

Masukan:

D = {d1, d2,...,dn} //kumpulan dari *n* data

k //banyaknya *cluster*

Keluaran:

Satu set *k clusters*

Langkah-langkah:

Fase 1: Menentukan pusat awal *centroid* dari *cluster.*

Fase 2: Menempatkan tiap *record* pada *cluster.*

Tahap pertama Improved K-Means, *centroid* awal ditentukan sistematis sehingga menghasilkan *cluster* dengan akurasi yang lebih baik. Tahap kedua memanfaatkan variasi dari metode pengelompokan dimulai dengan membentuk *cluster* awal berdasarkan jarak relatif setiap *record* dari *centroid* awal. *Cluster* ini kemudikan diperbaiki dengan menggunakan pendekatan heuristik, dengan demikian dapat meningkatkan efisiensi.

Dua fase algoritma Improved K-Means diuraikan sebagai berikut (Nazeer, 2009):

1. Fase menentukan pusat awal *centroid.*
    1. Inputkan *dataset* dan k *cluster* yang diinginkan.
    2. Set m=1.
    3. Hitung jarak tiap data dengan seluruh himpunan di D.
    4. Temukan pasangan data terdekat dan bentuk himpunan data-point Am ($1 \leq m \leq k$) yang berisi dua data-point tersebut. Hapus baris data dari D.
    5. Temukan data di D yang dekat dengan data-point Am. Tambahkan pada Am dan hapus dari D. Ulangi langkah ini hingga jumlah data-point mencapai *threshold**(n/k).
    6. Jika m<k, maka m=m+1, tentukan pasangan data-point yang lain, kembali ke langkah 4.
    7. Untuk setiap data-point Am, temukan rata-rata dari tiap vektor data-point pada Am, rata-rata ini akan menjadi inisial *centroid* (*centroid* awal).
2. Fase menempatkan tiap *record* pada *cluster*
    1. Input himpunan *dataset* D, himpunan *cluster* C.
    2. Hitung jarak tiap data-point di($1 \leq i \leq n$) dengan semua *centroid* cj($1 \leq j \leq k$) sebagai d(di, cj).
    3. Untuk setiap data-point di, temukan *centroid* yang terdekat cj dan masukan di ke *cluster* j.
    4. Set *Cluster*ID[i]=j;
    5. Set NearestDist[i]=d(di, cj);
    6. Untuk tiap *cluster* j($1 \leq j \leq k$), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.
    7. ULANG.
    8. Untuk tiap data-point di:

        8.1 Hitung jarak data dengan *cluster* akhir yang terdekat

        8.2 IF jarak sama atau kurang dari *cluster* akhir yang terdekat, data tetap berada dalam *cluster*

        ELSE

Data berpindah *cluster*, lalu ulangi langkah 2 hingga 5.

9. UNTIL kriteria konvergen ditemui, yaitu ketika tidak ada data yang berpindah *cluster*.

2.8 Analisis *Cluster*

Analisis *cluster* merupakan teknik statistik multivariat yang pada awalnya dikembangkan klasifikasi biologis. Dalam penelitian untuk *clustering* ini, digunakan ukuran kualitas atau *goodness*. Jenis ukuran ini memungkinkan untuk membandingkan suatu kelompok yang berbeda dari kelompok lain berdasarkan kesamaan berpasangan dokumen dalam *cluster*. Jenis lain dari ukuran kualitas ini memungkinkan mengevaluasi seberapa baik pengelompokan tersebut dengan membandingkan kelompok yang dihasilkan oleh teknik *clustering* untuk kelas yang sudah ada (Saracli, 2013).

2.9 Silhouette Coefficient

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation*. Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut (Al Zhoubi, 2008):

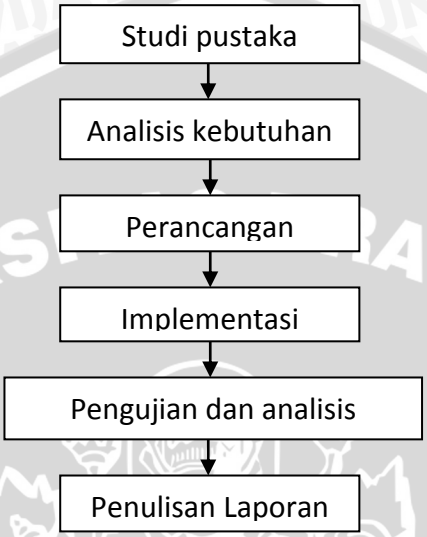
1. Untuk setiap obyek *i*, hitung rata-rata jarak antar obyek pada *cluster* tersebut. Beri nama α(i).
2. Untuk setiap obyek *i,* hitung rata-rata jarak ke semua obyek pada *cluster* lain. Kemudian diambil nilai mininummya. Beri nama *b*(i).
3. Nilai *Silhouette* Coefficient nya adalah :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (2.4)$$

Nilai *silhouette coefficient* dapat bervariasi antara -1 dan 1. Nilai -1 tidak diinginkan dimana α(i) lebih besar dari *b*(i). Hal yang diinginkan adalah nilai *silhouette* positif (α(i)< *b*(i)), dan untuk α(i). Ukuran keseluruhan dari pengelompokan dapat diperoleh dengan menghitung *silhouette coefficient* rata-rata semua titik. Semakin tinggi nilai *silhouette* semakin bagus juga kekuatan dan kualitas *cluster.*

BAB 3 METODOLOGI

Pada bab metodologi ini mengulas metode yang dipakai dalam penelitian yang terdiri dari studi pustaka, analisis kebutuhan, metode perancangan, metode implementasi, metode pengujian dan analisis serta penulisan laporan. Diagram alir dari penelitian dapat diilustrasikan pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

3.1 Studi Pustaka

Studi pustaka dilakukan untuk mempelajari dan literatur-literatur yang menunjang atau mendukung dalam proses perancangan, pembuatan dan pengujian pengelompokan portal jurnal internasional menggunakan Improved K-Means. Sumber dapat berasal dari skripsi, makalah, paper, buku cetak, e-book, jurnal, dan juga berbagai macam tutorial pemrograman yang berasal dari internet. Terdapat beberapa teori yang dipelajari guna pembuatan laporan ini diantaranya:

- *SCImagojr Journal and Country Rank*
- Jurnal Ilmiah
- *Impact Factor*
- *Data Mining*
- *Clustering*
- K-Means *Clustering*
- Improved K-Means
- Analisis *cluster*

- Pengujian kualitas *clustering* menggunakan *Silhouette Coefficient*.

3.2 Analisis Kebutuhan

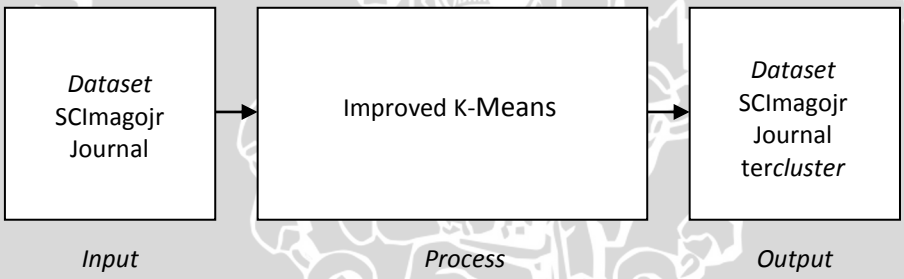
Analisis kebutuhan bertujuan untuk mendapatkan semua kebutuhan dalam membangun aplikasi pengelompokan portal jurnal internasional menggunakan metode Improved K-Means.

3.3 Perancangan

Perancangan dilakukan sebagai dasar untuk proses implementasi. Adapun tahap-tahap dalam perancangan adalah perancangan diagram alir algoritma, perhitungan manual dan perancangan antarmuka.

3.3.1 Deskripsi Umum Sistem

Pada bagian ini dijelaskan mengenai langkah-langkah yang akan dilakukan untuk membuat sistem. Sistem yang dibuat merupakan implementasi Algoritma Improved K-Means. Tujuan dari sistem ini adalah untuk pengelompokan portal jurnal internasional. Desain sistem secara umum dapat dilihat pada Gambar 3.2.

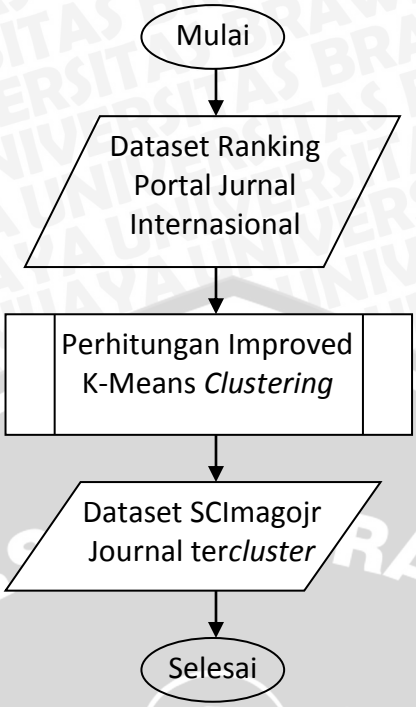


Gambar 3.2 Gambaran Umum Sistem

Secara umum proses pertama sistem ini yakni user memasukkan *dataset* ranking Jurnal Internasional. Kemudian sistem akan mengelompokkan *dataset* tersebut dengan algoritma Improved K-Means. Hasil dari proses *clustering* yakni *dataset* yang telah di*cluster.*

3.3.2 Diagram Alir Sistem

Pada subbab ini akan dijelaskan bagaimana proses-proses yang dilakukan dalam sistem untuk mengelompokkan portal jurnal internasional menggunakan Improved K-Means. Adapun alur proses yang digunakan dalam sistem digambarkan dalam diagram alir (*flowchart*) pada Gambar 3.3.



Gambar 3.3 Diagram Alir Sistem

Gambar 3.3 merupakan diagram alir sistem. Proses diawali dari masukan *dataset ranking* Jurnal Internasional yang diambil dari *The SCImago Journal & Country Rank.* Dari *dataset* tersebut dilakukan proses *clustering* atau pengelompokan dengan Algoritma Improved K-Means.

3.3.3 Pengelompokan Data dengan Algoritma Improved K-Means

Algoritma pengelompokan yang digunakan pada skripsi ini adalah Algoritma Improved K-Means. Algoritma Improved K-Means terdiri dari dua langkah utama yaitu menemukan inisial *centroid* dan memasukkan *record* ke dalam *cluster* (Nazeer, 2009). Berikut ini adalah proses menemukan inisial *centroid*:

1. Inputkan *dataset* dan k *cluster* yang diinginkan.
2. Set m=1.
3. Hitung jarak tiap data dengan seluruh himpunan di D.
4. Temukan pasangan data terdekat dan bentuk himpunan data-point Am ($1 \leq m \leq k$) yang berisi dua data-point tersebut. Hapus baris data dari D.
5. Temukan data di D yang dekat dengan data-point Am. Tambahkan pada Am dan hapus dari D. Ulangi langkah ini hingga jumlah data-point mencapai *threshold**(n/k).
6. Jika m<k, maka m=m+1, tentukan pasangan data-point yang lain, kembali ke langkah 4.

7. Untuk setiap data-point Am, temukan rata-rata dari tiap vektor data-point pada Am, rata-rata ini akan menjadi inisial *centroid* (*centroid* awal).

Langkah selanjutnya adalah menempatkan data pada *cluster*. Pada langkah ini akan dilakukan dengan cara sebagai berikut:

1. Input himpunan *dataset* D, himpunan *cluster* C.
2. Hitung jarak tiap data-point di($1 \leq i \leq n$) dengan semua *centroid* cj($1 \leq j \leq k$) sebagai d(di, cj).
3. Untuk setiap data-point di, temukan *centroid* yang terdekat cj dan masukan di ke *cluster* j.
4. Set *Cluster*ID[i]=j;
5. Set NearestDist[i]=d(di, cj);
6. Untuk tiap *cluster* j($1 \leq j \leq k$), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.
7. ULANG.
8. Untuk tiap data-point di:
   8.1 Hitung jarak data dengan *cluster* akhir yang terdekat
   8.2 IF jarak sama atau kurang dari *cluster* akhir yang terdekat, data tetap berada dalam *cluster*

   ELSE

   Data berpindah *cluster*, lalu ulangi langkah 2 hingga 5.
9. UNTIL kriteria konvergen ditemui, yaitu ketika tidak ada data yang berpindah *cluster*.

3.4 Implementasi

Implementasi aplikasi pengelompokan *dataset* menggunakan metode Improved K-Means dilakukan dengan mengacu pada perancangan sistem. Implementasi perangkat lunak dilakukan dengan bahasa pemrograman Java. Implementasi program ini meliputi:

- Spesifikasi perangkat keras dan lunak
- Implementasi algoritma
- Implementasi antarmuka

3.5 Pengujian

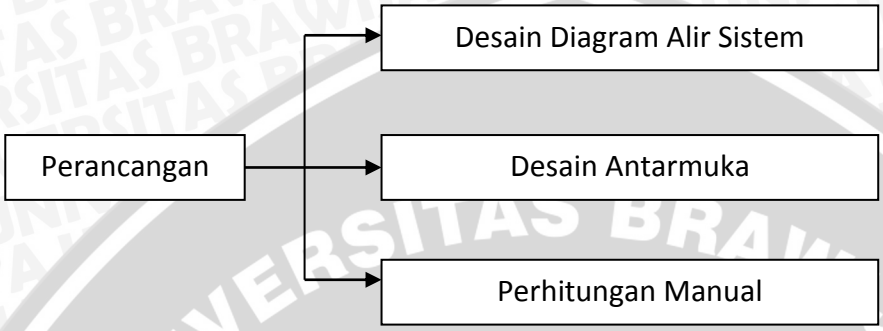
Pengujian metode dilakukan untuk memastikan bahwa aplikasi yang telah dibuat sudah benar. Pengujian yang dilakukan dalam penelitian ini adalah pengujian dengan menggunakan *Silhouette Coefficient.*

3.6 Kesimpulan

Kesimpulan berisi rangkuman dari hasil penelitian yang dibutuhkan untuk pengelompokan portal jurnal internasional menggunakan metode Improved K-Means.

BAB 4 PERANCANGAN

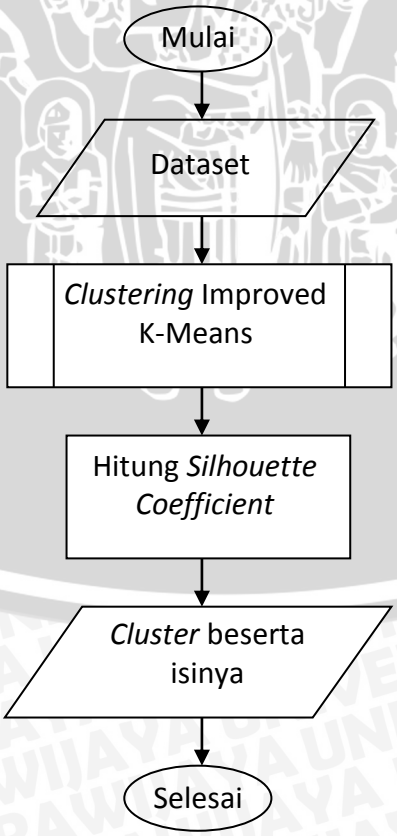
Perancangan perangkat lunak pada skripsi ini terdiri dari lima tahap, yaitu desain diagram alir sistem (*flowchart*), desain antarmuka, dan perhitungan manual. Gambar 4.1 merupakan diagram perancangan yang diterapkan pada skripsi ini.



Gambar 4.1 Diagram Perancangan

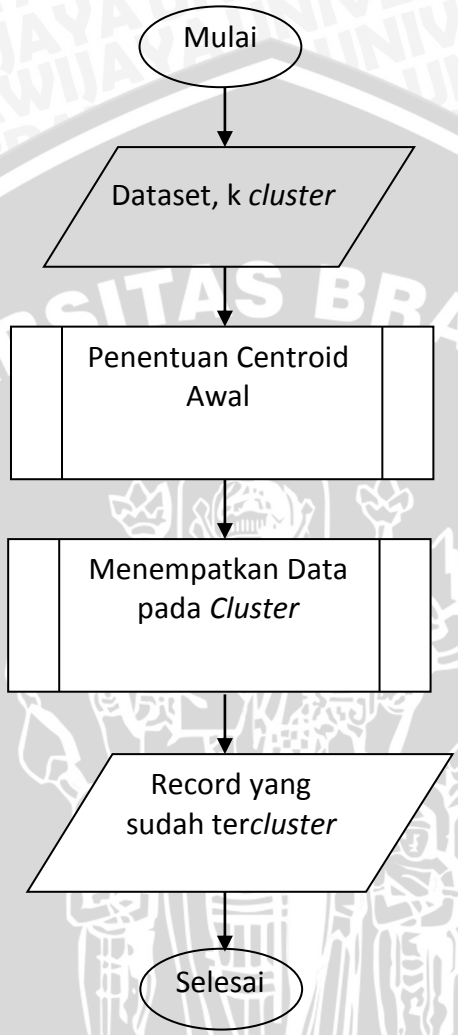
4.1 Desain Diagram Alir

Sistem yang dibangun memiliki tiga proses utama yaitu proses pengelompokan data, pengujian dengan *Silhouette Coefficient,* dan *cluster* beserta isinya. Alur proses dari sistem ditampilkan oleh Gambar 4.2.



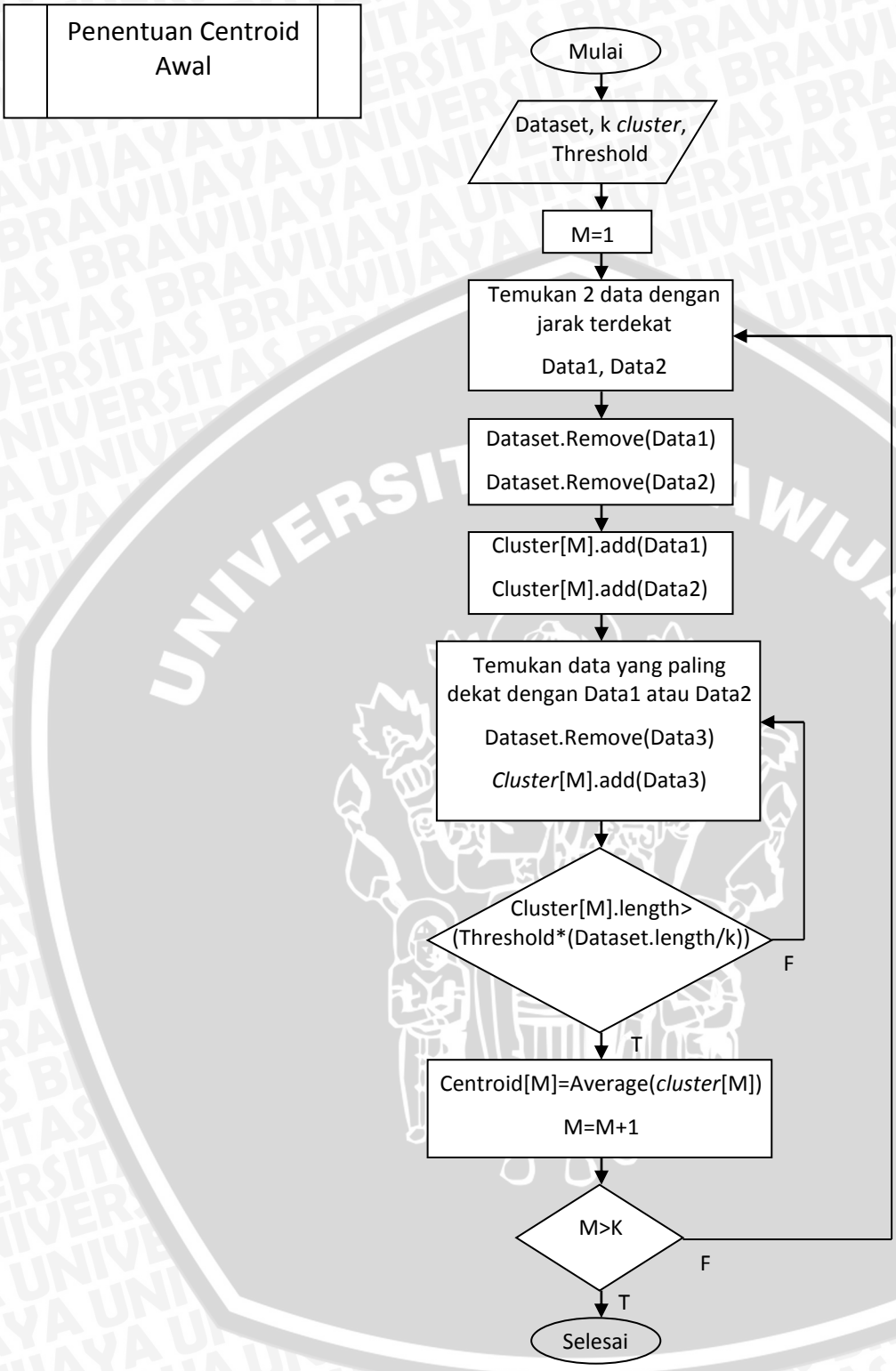
Gambar 4.2 Alur Proses Sistem Secara Umum

Algoritma *clustering* pada penelitian ini adalah algoritma Improved K-Means. Langkah-langkah pada Improved K-Means ada dua, yaitu menemukan *centroid* awal dan menaruh data pada *cluster*. Gambar 4.3 adalah *flowchart clustering* Improved K-Means.

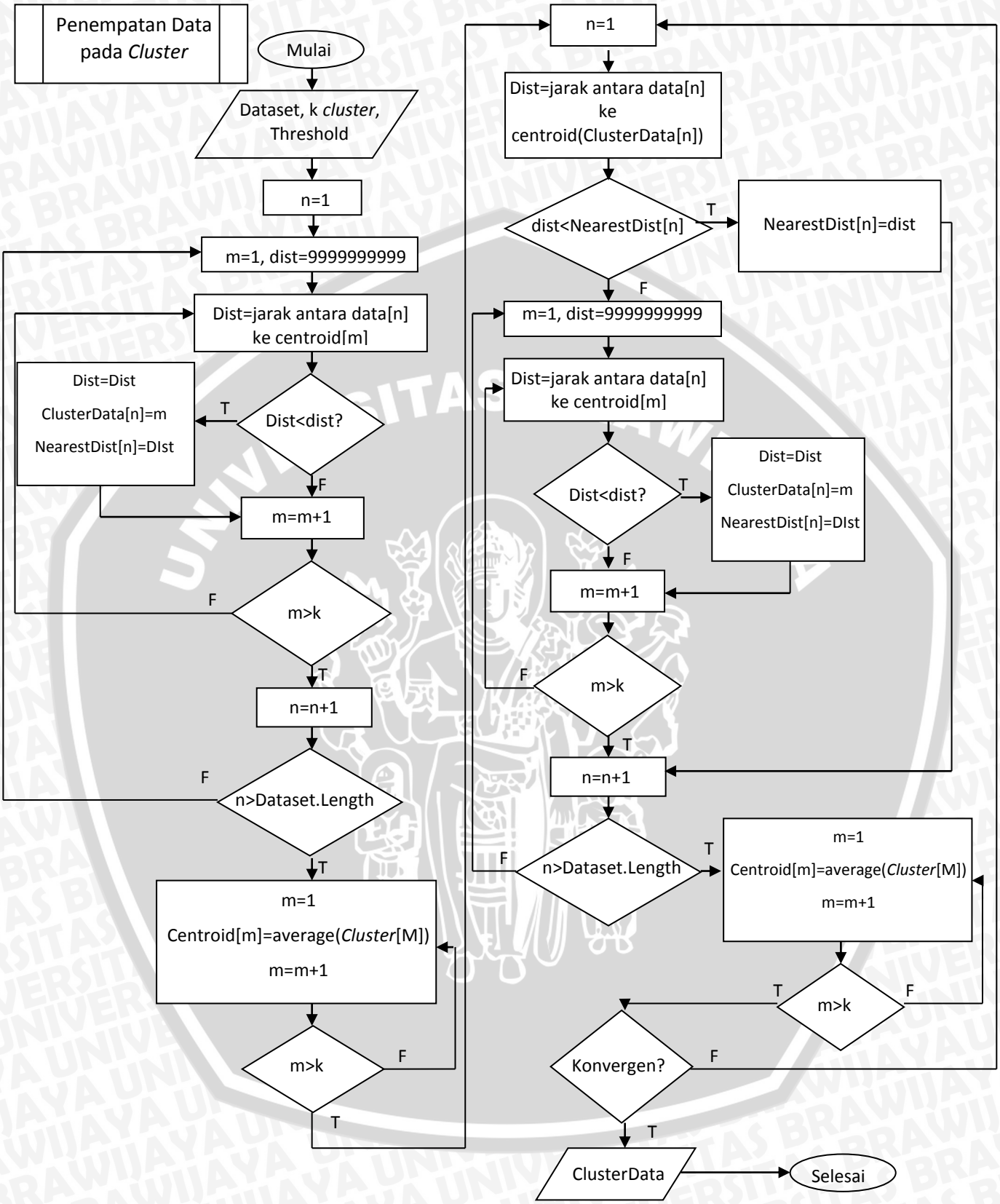


Gambar 4.3 *Flowchart Clustering* Improved K-Means

Gambar 4.4 merupakan *flowchart* untuk proses awal dari Improved K-Means yaitu *flowchart* menemukan *centroid* awal. Gambar 4.5 merupakan *flowchart* proses kedua dari Improved K-Means yaitu flowchart menempatkan data pada *cluster.*



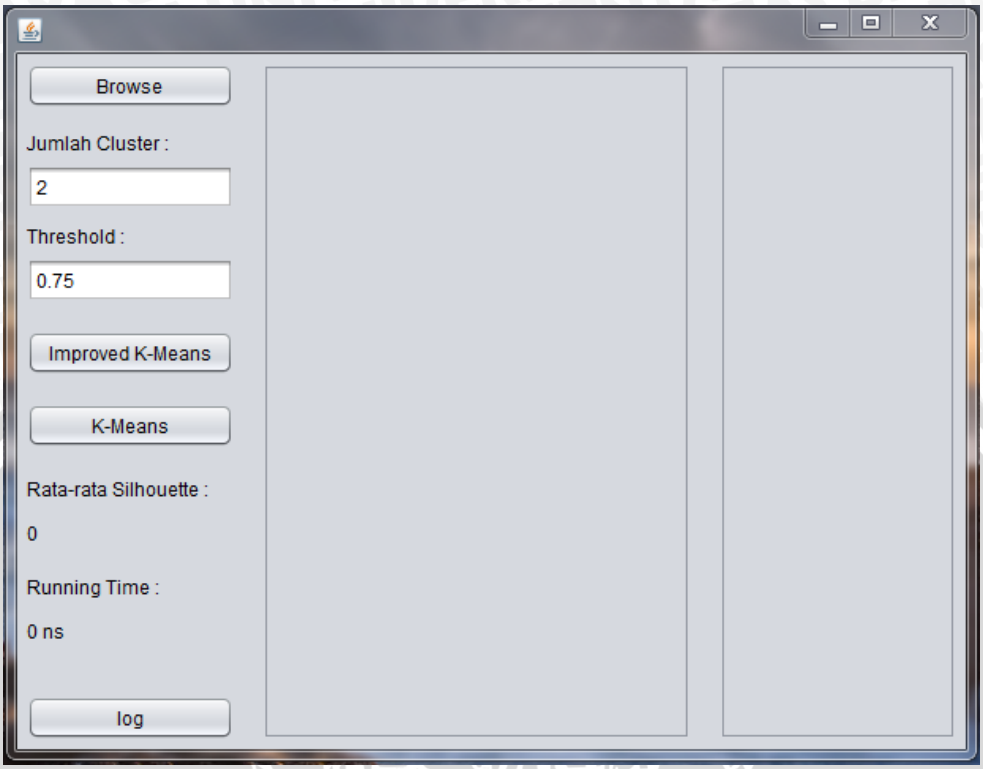
Gambar 4.4 *Flowchart* Penemuan *Centroid* Awal



Gambar 4.5 *Flowchart* Penempatan Data pada *Cluster*

4.2 Desain Antarmuka

Subbab ini menjelaskan antarmuka sistem yang digunakan dalam penelitian ini. Adapun desain antarmuka sistem digambarkan pada Gambar 4.6.



Gambar 4.6 Rancangan Antarmuka Implementasi Improved K-Means

Terdapat tujuh bagian yang berfungsi dalam jalannya aplikasi, antara lain *Button Browse, Combo Box K, Combo Box threshold, Datagridview Dataset, Button Execute by* K-Means*, Button Execute by* Improved K-Means*,* dan *RichTextBox Result.*

Keterangan:

1. *Button Browse* adalah button yang berfungsi untuk menampilkan *dataset* yang akan digunakan dalam perhitungan. *Dataset* yang akan ditampilkan muncul di *Datagridview Dataset*.
2. *Datagridview Dataset* adalah data *gridview* tempat untuk menampilkan *dataset* yang akan digunakan.
3. *ComboBox K* adalah tempat untuk menginputkan nilai K yang dibutuhkan.
4. *ComboBox Threshold* adalah tempat untuk menginputkan nilai *threshold* yang dibutuhkan.
5. *Button execute by* K-Means adalah button dimana fungsi k-means dijalankan, hasil dari *button execute by* K-Means akan ditampilkan pada *Rich Text Box Result.*

6. *Button execute by* Improved K-Means adalah button dimana fungsi utama dijalankan, hasil dari *button execute by* Improved K-Means akan ditampilkan pada *Rich Text Box Result.*
7. *Rich Text Box Result* adalah tempat dimana hasil perhitungan berupa *dataset* yang telah ter*cluster*.

4.3 Perhitungan Manual

4.3.1 Perhitungan Manual Improved K-Means

Perhitungan manual berfungsi sebagai gambaran umum perancangan sistem. Dengan melakukan perhitungan manual dapat diketahui apakah perhitungan yang nantinya akan dilakukan sistem benar atau tidak. Pada proses perhitungan manual Improved K-Means, data diambil dari *SCImagojr Journal* dengan subject area Computer Vision dan *subject category Artificial Intelligence.* Pada perhitungan manual atau manualisasi ini, digunakan *dataset* sejumlah 10 data dengan 3 fitur yaitu Total Docs (3 years), Total Cites (3 years), dan Citable Docs (3 years). *Dataset* yang digunakan dapat dilihat pada Tabel 4.1.

Tabel 4.1 *Dataset* Perhitungan Manual

No	Title	Total Docs (3 year)	Total Cites (3 year)	Citable Docs (3 year)
1	IEEE Transactions on Pattern Analysis and Machine Intelligence	577	6110	542
2	International Journal of Computer Vision	301	1893	285
3	International Journal of Machine Learning and Cybernetics	35	244	33
4	Foundations and Trends in Computer Graphics and Vision	11	63	11
5	2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009	379	1896	377
6	Medical Image Analysis	210	1106	197
7	Pattern Recognition	909	4305	895

8	Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	901	3126	899
9	Journal of Mathematical Imaging and Vision	169	450	164
10	Signal Processing	853	2560	844

Langkah 1 : input k=2

Tabel 4.2 Tabel *Dataset* Perhitungan Manual 1

Record	A1	A2	A3
1	577	6110	542
2	301	1893	285
3	35	244	33
4	11	63	11
5	379	1896	377
6	210	1106	197
7	909	4305	895
8	901	3126	899
9	169	450	164
10	853	2560	844

Langkah 2 : m=1

Langkah 3 : Hitung jarak tiap baris data dengan seluruh himpunan data di D.

Jarak = $\sqrt{\sum_{i=0}^{n}\sum_{j=0}^{n} Ri^2 - Rj^2}$ → Ri = record ke-1

Rj = record ke-2

Misalkan :

Record 1 dan 2 : √(577-301)^2 + (6110-1893)^2 + (542-285)^2 = 4233,83

Record 1 dan 3 : √(577-35)^2 + (6110-244)^2 + (542-33)^2 = 5912,93

Record 1 dan 4 : √(577-11)^2 + (6110-63)^2 + (542-11)^2 = 6096,60

Record 1 dan 5 : $\sqrt{(577-379)^2 + (6110-1896)^2 + (542-377)^2} = 4221,87$

Record 1 dan 6 : $\sqrt{(577-210)^2 + (6110-1106)^2 + (542-197)^2} = 5029,29$

Record 1 dan 7 : $\sqrt{(577-909)^2 + (6110-4305)^2 + (542-895)^2} = 1868,92$

Record 1 dan 8 : $\sqrt{(577-901)^2 + (6110-3126)^2 + (542-899)^2} = 3022,69$

Record 1 dan 9 : $\sqrt{(577-169)^2 + (6110-450)^2 + (542-164)^2} = 5687,26$

Record 1 dan 10 : $\sqrt{(577-853)^2 + (6110-2560)^2 + (542-844)^2} = 3573,50$

Record 2 dan 3 : $\sqrt{(301-35)^2 + (1893-244)^2 + (285-33)^2} = 1689,22$

Record 2 dan 4 : $\sqrt{(301-11)^2 + (1893-63)^2 + (285-11)^2} = 1872,99$

Record 2 dan 5 : $\sqrt{(301-379)^2 + (1893-1896)^2 + (285-377)^2} = 120,65$

Record 2 dan 6 : $\sqrt{(301-210)^2 + (1893-1106)^2 + (285-197)^2} = 797,12$

Record 2 dan 7 : $\sqrt{(301-909)^2 + (1893-4305)^2 + (285-895)^2} = 2561,15$

Record 2 dan 8 : $\sqrt{(301-901)^2 + (1893-3126)^2 + (285-899)^2} = 1502,43$

Record 2 dan 9 : $\sqrt{(301-169)^2 + (1893-450)^2 + (285-164)^2} = 1454,07$

Record 2 dan 10 : $\sqrt{(301-853)^2 + (1893-2560)^2 + (285-844)^2} = 1030,57$

Record 3 dan 4 : $\sqrt{(35-11)^2 + (244-63)^2 + (33-11)^2} = 183,90$

Record 3 dan 5 : $\sqrt{(35-379)^2 + (244-1896)^2 + (33-377)^2} = 1722,14$

Record 3 dan 6 : $\sqrt{(35-210)^2 + (244-1106)^2 + (33-197)^2} = 894,74$

Record 3 dan 7 : $\sqrt{(35-909)^2 + (244-4305)^2 + (33-895)^2} = 4242,48$

Record 3 dan 8 : $\sqrt{(35-901)^2 + (244-3126)^2 + (33-899)^2} = 3131,43$

Record 3 dan 9 : $\sqrt{(35-169)^2 + (244-450)^2 + (33-164)^2} = 278,48$

Record 3 dan 10 : $\sqrt{(35-853)^2 + (244-2560)^2 + (33-844)^2} = 2587,64$

Record 4 dan 5 : $\sqrt{(11-379)^2 + (63-1896)^2 + (11-377)^2} = 1905,06$

Record 4 dan 6 : $\sqrt{(11-210)^2 + (63-1106)^2 + (11-197)^2} = 1077,98$

Record 4 dan 7 : $\sqrt{(11-909)^2 + (63-4305)^2 + (11-895)^2} = 4425,20$

Record 4 dan 8 : $\sqrt{(11-901)^2 + (63-3126)^2 + (11-899)^2} = 3310,98$

Record 4 dan 9 : $\sqrt{(11-169)^2 + (63-450)^2 + (11-164)^2} = 445,13$

Record 4 dan 10 : $\sqrt{(11-853)^2 + (63-2560)^2 + (11-844)^2} = 2763,67$

Record 5 dan 6 : $\sqrt{(379-210)^2 + (1896-1106)^2 + (377-197)^2} = 827,68$

Record 5 dan 7 : $\sqrt{(379-909)^2 + (1896-4305)^2 + (377-895)^2} = 2520,42$

Record 5 dan 8 : $\sqrt{(379-901)^2 + (1896-3126)^2 + (377-899)^2} = 1434,53$

Record 5 dan 9 : $\sqrt{(379-169)^2 + (1896-450)^2 + (377-164)^2} = 1476,61$

Record 5 dan 10 : $\sqrt{(379-853)^2 + (1896-2560)^2 + (377-844)^2} = 940,03$

Record 6 dan 7 : $\sqrt{(210-909)^2 + (1106-4305)^2 + (197-895)^2} = 3348,04$

Record 6 dan 8 : $\sqrt{(210-901)^2 + (1106-3126)^2 + (197-899)^2} = 2247,37$

Record 6 dan 9 : $\sqrt{(210-169)^2 + (1106-450)^2 + (197-164)^2} = 658,11$

Record 6 dan 10 : $\sqrt{(210-853)^2 + (1106-2560)^2 + (197-844)^2} = 1716,44$

Record 7 dan 8 : $\sqrt{(909-901)^2 + (4205-3126)^2 + (895-899)^2} = 1179,03$

Record 7 dan 9 : $\sqrt{(909-169)^2 + (4205-450)^2 + (895-164)^2} = 3992,87$

Record 7 dan 10 : $\sqrt{(909-853)^2 + (4205-2560)^2 + (895-844)^2} = 1746,64$

Record 8 dan 9 : $\sqrt{(901-169)^2 + (3126-450)^2 + (899-164)^2} = 2870,02$

Record 8 dan 10 : $\sqrt{(901-853)^2 + (3126-2560)^2 + (899-844)^2} = 570,69$

Record 9 dan 10 : $\sqrt{(169-853)^2 + (450-2560)^2 + (164-844)^2} = 2319,99$

Langkah 4 : Temukan pasangan data terdekat dari D dan bentuk himpunan data-point Am ($1 \leq m \leq k$) yang berisi dua data point tersebut. Hapus baris data tersebut dari D.

Record 2 dan 5 dengan jarak 120,65 adalah yang terdekat. Record 2 dan 5 dihilangkan dari *dataset*.

Langkah 5 : Temukan baris data di D yang dekat dengan datapoint Am. Tambahkan pada Am dan hapus dari d. Ulangi langkah ini hingga jumlah data point mencapai *threshold* = nilai *threshold(n/k).**

Misal nilai *Threshold* = 0,75*(n/k)

= 0,75*(10/2)

= 3,75 = 3

Ambil satu record yang memiliki jarak terdekat dengan record 2 dan 5. Pilihan jatuh pada record 6 dengan nilai 797,12.

Langkah 6 : Jika m<k, maka m=m+1, temukan pasangan data-point yang lain kembali ke langkah 4. Hitung jarak tiap baris data selain pasangan yang sudah membentuk himpunan datapoint Am pada awal perhitungan.

Record 1 dan 3 : $\sqrt{(577-35)^2 + (6110-244)^2 + (542-33)^2} = 5912,93$

Record 1 dan 4 : $\sqrt{(577-11)^2 + (6110-63)^2 + (542-11)^2} = 6096,60$

Record 1 dan 7 : $\sqrt{(577-909)^2 + (6110-4305)^2 + (542-895)^2} = 1868,92$

Record 1 dan 8 : $\sqrt{(577-901)^2 + (6110-3126)^2 + (542-899)^2} = 3022,69$

Record 1 dan 9 : $\sqrt{(577-169)^2 + (6110-450)^2 + (542-164)^2} = 5687,26$

Record 1 dan 10 : $\sqrt{(577-853)^2 + (6110-2560)^2 + (542-844)^2} = 3573,50$

Record 3 dan 4 : $\sqrt{(35-11)^2 + (244-63)^2 + (33-11)^2} = 183,90$

Record 3 dan 7 : $\sqrt{(35-909)^2 + (244-4305)^2 + (33-895)^2} = 4242,48$

Record 3 dan 8 : $\sqrt{(35-901)^2 + (244-3126)^2 + (33-899)^2} = 3131,43$

Record 3 dan 9 : $\sqrt{(35-169)^2 + (244-450)^2 + (33-164)^2} = 278,48$

Record 3 dan 10 : $\sqrt{(35-853)^2 + (244-2560)^2 + (33-844)^2} = 2587,64$

Record 4 dan 7 : $\sqrt{(11-909)^2 + (63-4305)^2 + (11-895)^2} = 4425,20$

Record 4 dan 8 : $\sqrt{(11-901)^2 + (63-3126)^2 + (11-899)^2} = 3310,98$

Record 4 dan 9 : $\sqrt{(11-169)^2 + (63-450)^2 + (11-164)^2} = 445,13$

Record 4 dan 10 : $\sqrt{(11-853)^2 + (63-2560)^2 + (11-844)^2} = 2763,67$

Record 7 dan 8 : $\sqrt{(909-901)^2 + (4205-3126)^2 + (895-899)^2} = 1179,03$

Record 7 dan 9 : $\sqrt{(909-169)^2 + (4205-450)^2 + (895-164)^2} = 3992,87$

Record 7 dan 10 : $\sqrt{(909-853)^2 + (4205-2560)^2 + (895-844)^2} = 1746,64$

Record 8 dan 9 : $\sqrt{(901-169)^2 + (3126-450)^2 + (899-164)^2} = 2870,02$

Record 8 dan 10 : $\sqrt{(901-853)^2 + (3126-2560)^2 + (899-844)^2} = 570,69$

Record 9 dan 10 : $\sqrt{(169-853)^2 + (450-2560)^2 + (164-844)^2} = 2319,99$

Record 3 dan 4 dengan jarak 183,90 adalah yang terdekat. Record 3 dan 4 dihilangkan dari *dataset*.

Ambil satu record yang memiliki jarak mendekati 183,90 pilihan jatuh pada record 9 (278,48).

Langkah 7 : Untuk setiap himpunan data-point Am, temukan rata-rata dari tiap vector data-point pada AM, rata-rata ini akan menjadi *centroid* awal.

Centroid awal 1 : record 2, 5, 6

Centroid awal 2 : record 3, 4, 9

Centroid 1 :

Record	A1	A2	A3
2	301	1893	285
5	379	1896	377
6	210	1106	197
Rataan	296,67	1631,67	286,33

Centroid 2 :

Record	A1	A2	A3
3	35	244	33
4	11	63	11
9	169	450	164
Rataan	71,67	252,33	69,33

Selanjutnya adalah menentukan baris data ke dalam *cluster*.

Langkah 8 : Input himpunan *dataset* D, himpunan *cluster* C

Tabel 4.3 Tabel *Dataset* Perhitungan Manual 2

Record	A1	A2	A3
1	577	6110	542
2	301	1893	285
3	35	244	33
4	11	63	11

5	379	1896	377
6	210	1106	197
7	909	4305	895
8	901	3126	899
9	169	450	164
10	853	2560	844

Langkah 9 : Hitung jarak tiap baris data di (1 ≤ i ≤ n) dengan semua *centroid* (1 ≤ j ≤ k) sebagai d(di,cj)

Misalkan:

(Record 1, *Centroid* 1): $\sqrt{(577-296,67)^2+(6110-1631,67)^2+(542-286,33)^2}=4494,37$

(Record 1, *Centroid* 2) : $\sqrt{(577-71,67)^2+(6110-252,33)^2+(542-69,33)^2}=5898,39$

(Record 2, *Centroid* 1): $\sqrt{(301-296,67)^2+(1893-1631,67)^2+(285-286,33)^2}=261,36$

(Record 2, *Centroid* 2) : $\sqrt{(301-71,67)^2+(1893-252,33)^2+(285-69,33)^2}=1670,6$

(Record 3 *Centroid* 1) : $\sqrt{(35-296,67)^2+(244-1631,67)^2+(33-286,33)^2}=1434,66$

(Record 3 *Centroid* 2) : $\sqrt{(35-71,67)^2+(244-252,33)^2+(33-69,33)^2}=52,28$

(Record 4, *Centroid* 1) : $\sqrt{(11-296,67)^2+(63-1631,67)^2+(11-286,33)^2}=1618$

(Record 4, *Centroid* 2) : $\sqrt{(11-71,67)^2+(63-252,33)^2+(11-69,33)^2}=207,19$

(Record 5, *Centroid* 1): $\sqrt{(379-296,67)^2+(1896-1631,67)^2+(377-286,33)^2}=291,32$

(Record 5, *Centroid* 2) : $\sqrt{(379-71,67)^2+(1896-252,33)^2+(377-69,33)^2}=1700,22$

(Record 6, *Centroid* 1): $\sqrt{(210-296,67)^2+(1106-1631,67)^2+(197-286,33)^2}=540,20$

(Record 6, *Centroid* 2) : $\sqrt{(210-71,67)^2+(1106-252,33)^2+(197-69,33)^2}=874,17$

(Record 7, *Centroid* 1): $\sqrt{(909-296,67)^2+(4305-1631,67)^2+(895-286,33)^2}=2809,29$

(Record 7, *Centroid* 2) : $\sqrt{(909-71,67)^2+(4305-252,33)^2+(895-69,33)^2}=4219,83$

(Record 8, *Centroid* 1): $\sqrt{(901-296,67)^2+(3126-1631,67)^2+(899-286,33)^2}=1724,41$

(Record 8, *Centroid* 2) : $\sqrt{(901-71,67)^2+(3126-252,33)^2+(899-69,33)^2}= 3103,88$

(Record 9, *Centroid* 1) : √(169-296,67)^2+(450-1631,67)^2+(164-286,33)^2=1194,82

(Record 9, *Centroid* 2) : √(169-71,67)^2+(450-252,33)^2+(164-69,33)^2=239,81s

(Record 10, *Centroid* 1):√(853-296,67)^2+(2560-1631,67)^2+(844-286,33)^2=1217,49

(Record 10, *Centroid* 2) : √(853-71,67)^2+(2560-252,33)^2+(844-69,33)^2=2556,647

Langkah 10 : Untuk setiap baris data di, temukan *centroid* yang terdekat cj dan masukkan di ke *cluster* j.

Centroid terdekat untuk tiap record akan ditunjukkan oleh Tabel 4.4.

Tabel 4.4 Tabel *Centroid* Terdekat Tiap Record (Perulangan 1)

Record	*Cluster*
1	1
2	1
3	2
4	2
5	1
6	1
7	1
8	1
9	2
10	1

Langkah 11 : Untuk tiap *cluster* j (1 ≤ j ≤ k), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.

Cluster 1 :

Record	A1	A2	A3
1	577	6110	542
2	301	1893	285
5	379	1896	377
6	210	1106	197
7	909	4305	895

8	901	3126	899
10	853	2560	844
Rataan	590	2999,429	577

Cluster 2 :

Record	**A1**	**A2**	**A3**
3	35	244	33
4	11	63	11
9	169	450	164
Rataan	71,66	252,33	69,33

Langkah 12 : Hitung jarak tiap baris data di (1 ≤ i ≤ n) dengan semua *cluster* (1 ≤ j ≤ k) sebagai d(di,cj)

Misalkan:

(Record 1, *Cluster* 1):$\sqrt{(577—590)^2+(6110-2999,42)^2+(542-577)^2}$=3110,79

(Record1, *Cluster* 2):$\sqrt{(577—71,66)^2+(6110-252,33)^2+(542-69,33)^2}$=5898,39

(Record 2, *Cluster* 1):$\sqrt{(301—590)^2+(1893-2999,42)^2+(285-577)^2}$=2862,92

(Record2, *Cluster* 2):$\sqrt{(301—71,66)^2+(1893-252,33)^2+(285-69,33)^2}$=1670,59

(Record 3, *Cluster* 1):$\sqrt{(35—590)^2+(244-2999,42)^2+(33-577)^2}$=2862,92

(Record 3, *Cluster* 2):$\sqrt{(35—71,66)^2+(244-252,33)^2+(33-69,33)^2}$=52,28

(Record 4, *Cluster* 1):$\sqrt{(11—590)^2+(63-2999,42)^2+(11-577)^2}$=3046,01

(Record 4, *Cluster* 2):$\sqrt{(11—71,66)^2+(63-252,33)^2+(11-69,33)^2}$=207,19

(Record 5, *Cluster* 1):$\sqrt{(379—590)^2+(1896-2999,42)^2+(377-577)^2}$=1141,08

(Record 5,*Cluster* 2):$\sqrt{(379—71,66)^2+(1896-252,33)^2+(377-69,33)^2}$=1700,22

(Record 6, *Cluster* 1):$\sqrt{(210—590)^2+(1106-2999,42)^2+(197-577)^2}$=1968,21

(Record 6,*Cluster* 2):$\sqrt{(210—71,66)^2+(1106-252,33)^2+(197-69,33)^2}$=874,17

(Record 7, *Cluster* 1):$\sqrt{(909—590)^2+(4305-2999,42)^2+(895-577)^2}$=1381,08

(Record 7,*Cluster* 2):√(909—71,66)^2+(4305-252,33)^2+(895-69,33)^2=4219,82

(Record 8, *Cluster* 1):√(901—590)^2+(3126-2999,42)^2+(899-577)^2=465,21

(Record 8,*Cluster* 2):√(901—71,66)^2+(3126-252,33)^2+(899-69,33)^2=3103,88

(Record 9, *Cluster* 1):√(169—590)^2+(450-2999,42)^2+(164-577)^2=2616,75

(Record 9, *Cluster* 2):√(169—71,66)^2+(450-252,33)^2+(164-69,33)^2=239,80

(Record 10, *Cluster* 1):√(853—590)^2+(2560-2999,42)^2+(844-577)^2=577,54

(Record10,*Cluster*2):√(853—71,66)^2+(2560-252,33)^2+(844-69,33)^2=2556,54

Tabel 4.5 Tabel *Centroid* Terdekat Tiap Record (Perulangan 2)

Record	*Cluster*
1	1
2	1
3	2
4	2
5	1
6	2
7	1
8	1
9	2
10	1

Langkah 13 : Untuk tiap *cluster* j (1 ≤ j ≤ k), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.

Cluster 1 :

Record	A1	A2	A3
1	577	6110	542
2	301	1893	285
5	379	1896	377
7	909	4305	895

8	901	3126	899
10	853	2560	844
Rataan	653,33	3315	640,33

Cluster 2 :

Record	**A1**	**A2**	**A3**
3	35	244	33
4	11	63	11
6	210	1106	197
9	169	450	164
Rataan	106,25	465,75	101,25

Langkah 14 : Hitung jarak tiap baris data di (1 ≤ i ≤ n) dengan semua *cluster* (1 ≤ j ≤ k) sebagai d(di,cj)

(Record 1, *Cluster* 1):$\sqrt{(577-653,33)^2+(6110-3315)^2+(542-640,33)^2}$=2797,77

(Record 1, *Cluster* 2):$\sqrt{(577-106,25)^2+(6110-465,75)^2+(542-101,25)^2}$=5680,97

(Record 2, *Cluster* 1):$\sqrt{(301-653,33)^2+(1893-3315)^2+(285-640,33)^2}$=1507,47

(Record 2, *Cluster* 2):$\sqrt{(301-106,25)^2+(1893-465,75)^2+(285-101,25)^2}$=1452,14

(Record 3, *Cluster* 1):$\sqrt{(35-653,33)^2+(244-3315)^2+(33-640,33)^2}$=3190,96

(Record 3, *Cluster* 2):$\sqrt{(35-106,25)^2+(244-465,75)^2+(33-101,25)^2}$=242,70

(Record 4, *Cluster* 1):$\sqrt{(11-653,33)^2+(63-3315)^2+(11-640,33)^2}$=3374,04

(Record 4, *Cluster* 2):$\sqrt{(11-106,25)^2+(63-465,75)^2+(11-101,25)^2}$=423,58

(Record 5, *Cluster* 1):$\sqrt{(379-653,33)^2+(1896-3315)^2+(377-640,33)^2}$=1469,06

(Record 5, *Cluster* 2):$\sqrt{(379-106,25)^2+(1896-465,75)^2+(377-101,25)^2}$=1481,90

(Record 6, *Cluster* 1):$\sqrt{(210-653,33)^2+(1106-3315)^2+(197-640,33)^2}$=2296,25

(Record 6, *Cluster* 2):$\sqrt{(210-106,25)^2+(1106-465,75)^2+(197-101,25)^2}$=655,63

(Record 7, *Cluster* 1):√(909—653,33)^2+(4305-3315)^2+(895-640,33)^2=1053,71

(Record 7, *Cluster* 2):√(909—106,25)^2+(4305-465,75)^2+(895-101,25)^2=4001,78

(Record 8, *Cluster* 1):√(901—653,33)^2+(3126-3315)^2+(899-640,33)^2=404,92

(Record 8, *Cluster* 2):√(901—106,25)^2+(3126-465,75)^2+(899-101,25)^2=2888,76

(Record 9, *Cluster* 1):√(169—653,33)^2+(450-3315)^2+(164-640,33)^2=2944,43

(Record 9, *Cluster* 2):√(169—106,25)^2+(450-465,75)^2+(164-101,25)^2=90,12

(Record 10, *Cluster* 1):√(853—653,33)^2+(2560-3315)^2+(844-640,33)^2=807,07

(Record 10, *Cluster* 2):√(853—106,25)^2+(2560-465,75)^2+(844-101,25)^2=2344,18

Tabel 4.6 Tabel *Centroid* terdekat tiap record (Perulangan 3)

Record	*Cluster*
1	1
2	2
3	2
4	2
5	1
6	2
7	1
8	1
9	2
10	1

Langkah 15 : Untuk tiap *cluster* j (1 ≤ j ≤ k), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.

Cluster 1 :

Record	A1	A2	A3
1	577	6110	542
5	379	1896	377
7	909	4305	895

8	901	3126	899
10	853	2560	844
Rataan	723,8	3599,4	711,4

Cluster 2 :

Record	**A1**	**A2**	**A3**
2	301	1893	285
3	35	244	33
4	11	63	11
6	210	1106	197
9	169	450	164
Rataan	145,2	751,2	138

Langkah 16 : Hitung jarak tiap baris data di (1 ≤ i ≤ n) dengan semua *cluster* (1 ≤ j ≤ k) sebagai d(di,cj)

(Record 1, *Cluster* 1):√(577—723,8)^2+(6110-3599,4)^2+(542-711,4)^2=2520,58

(Record 1, *Cluster* 2):√(577—145,2)^2+(6110-751,2)^2+(542-138)^2=5391,327

(Record 2, *Cluster* 1):√(301—723,8)^2+(1893-3599,4)^2+(285-711,4)^2=1808,97

(Record 2, *Cluster* 2):√(301—145,2)^2+(1893-751,2)^2+(285-138)^2=1161,71

(Record 3, *Cluster* 1):√(35—723,8)^2+(244-3599,4)^2+(33-711,4)^2=3491,90

(Record 3, *Cluster* 2):√(35—145,2)^2+(244-751,2)^2+(33-138)^2=529,54

(Record 4, *Cluster* 1):√(11—723,8)^2+(63-3599,4)^2+(11-711,4)^2=3674,88

(Record 4, *Cluster* 2):√(11—145,2)^2+(63-751,2)^2+(11-138)^2=712,57

(Record 5, *Cluster* 1):√(379—723,8)^2+(1896-3599,4)^2+(377-711,4)^2=1769,82

(Record 5, *Cluster* 2):√(379—145,2)^2+(1896-751,2)^2+(377-138)^2=1192,62

(Record 6, *Cluster* 1):√(210—723,8)^2+(1106-3599,4)^2+(197-711,4)^2=2597,23

(Record 6, *Cluster* 2):√(210—145,2)^2+(1106-751,2)^2+(197-138)^2=365,4628

(Record 7, *Cluster* 1):√(909—723,8)^2+(4305-3599,4)^2+(895-711,4)^2=752,24

(Record 7, *Cluster* 2):√(909—145,2)^2+(4305-751,2)^2+(895-138)^2=3712,94

(Record 8, *Cluster* 1):√(901—723,8)^2+(3126-3599,4)^2+(899-711,4)^2=539,16

(Record 8, *Cluster* 2):√(901—145,2)^2+(3126-751,2)^2+(899-138)^2=2605,76

(Record 9, *Cluster* 1):√(169—723,8)^2+(450-3599,4)^2+(164-711,4)^2=3244,40

(Record 9, *Cluster* 2):√(169—145,2)^2+(450-751,2)^2+(164-138)^2=303,25

(Record 10, *Cluster* 1):√(853—723,8)^2+(2560-3599,4)^2+(844-711,4)^2=1055,75

(Record 10, *Cluster* 2):√(853—145,2)^2+(2560-751,2)^2+(844-138)^2=2066,68

Tabel 4.7 Tabel *Centroid* terdekat tiap record (Perulangan 4)

Record	*Cluster*
1	1
2	2
3	2
4	2
5	2
6	2
7	1
8	1
9	2
10	1

Langkah 17 : Untuk tiap *cluster* j (1 ≤ j ≤ k), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.

Cluster 1 :

Record	A1	A2	A3
1	577	6110	542
7	909	4305	895

8	901	3126	899
10	853	2560	844
Rataan	810	4025,25	795

Cluster 2 :

Record	**A1**	**A2**	**A3**
2	301	1893	285
3	35	244	33
4	11	63	11
5	379	1896	377
6	210	1106	197
9	169	450	164
Rataan	184,16	942	177,83

Langkah 18 : Hitung jarak tiap baris data di (1 ≤ i ≤ n) dengan semua *cluster* (1<=j<=k) sebagai d(di,cj)

(Record 1, *Cluster* 1):$\sqrt{(577-810)^2+(6110-4015,25)^2+(542-795)^2}$=2112,93

(Record 1, *Cluster* 2):$\sqrt{(577-184,16)^2+(6110-94)^2+(542-177,83)^2}$=5195,68

(Record 2, *Cluster* 1):$\sqrt{(301-810)^2+(1893-4015,25)^2+(285-795)^2}$=2250,70

(Record 2, *Cluster* 2):$\sqrt{(301-184,16)^2+(1893-942)^2+(285-177,83)^2}$=964,12

(Record 3, *Cluster* 1):$\sqrt{(35-810)^2+(244-4015,25)^2+(33-795)^2}$=3934,351

(Record 3, *Cluster* 2):$\sqrt{(35-184,16)^2+(244-9)^2+(33-177,83)^2}$=728,30

(Record 4, *Cluster* 1):$\sqrt{(11-810)^2+(63-4015,25)^2+(11-795)^2}$=4117,33

(Record 4, *Cluster* 2):$\sqrt{(11-184,16)^2+(63-942)^2+(11-177,83)^2}$=911,29

(Record 5, *Cluster* 1):$\sqrt{(379-810)^2+(1896-4015,25)^2+(377-795)^2}$=2212,28

(Record 5, *Cluster* 2):$\sqrt{(379-184,16)^2+(1896-942)^2+(377-177,83)^2}$=993,85

(Record 6, *Cluster* 1):$\sqrt{(210-810)^2+(1106-4015,25)^2+(197-795)^2}$=3039,67

(Record 6, *Cluster* 2):√(210—184,16)^2+(1106-942)^2+(197-177,83)^2=167,12

(Record 7, *Cluster* 1):√(909—810)^2+(4305-4015,25)^2+(895-795)^2=313,14

(Record 7, *Cluster* 2):√(909—184,16)^2+(4305-942)^2+(895-177,83)^2=3514,18

(Record 8, *Cluster* 1):√(901—810)^2+(3126-4015,25)^2+(899-795)^2=909,80

(Record 8, *Cluster* 2):√(901—184,16)^2+(3126-942)^2+(899-177,83)^2=2409,10

(Record 9, *Cluster* 1):√(169—810)^2+(450-4015,25)^2+(164-795)^2=3686,65

(Record 9, *Cluster* 2):√(169—184,16)^2+(450-942)^2+(164-177,83)^2=492,42

(Record 10, *Cluster* 1):√(853—810)^2+(2560-4015,25)^2+(844-795)^2=1466,7

(Record10,*Cluster* 2):√(853—184,16)^2+(2560-942)^2+(844-177,83)^2=1873,24

Tabel 4.8 Tabel *Centroid* terdekat tiap record (Perulangan 5)

Record	*Cluster*
1	1
2	2
3	2
4	2
5	2
6	2
7	1
8	1
9	2
10	1

Pada Tabel 4.8 dapat dilihat bahwa tidak ada record yang berpindah *cluster*. Dengan kata lain sudah konvergen, sehingga perhitungan berhenti.

4.3.2 Perhitungan Manual Pengujian Menggunakan *Silhouette Coefficient*

Analisa metode *Silhouette* ini dilakukan dengan melihat besar nilai *s* dari hasil perhitungan dengan menggunakan bantuan software Excel. Hasil perhitungan

nilai *silhouette coefficient* dapat bervariasi antara -1 hingga 1. Jika s_i = 1 berarti objek *i* sudah berada dalam *cluster* yang tepat. Jika nilai s_i = 0 maka objek *i* berada di antara dua *cluster* sehingga objek tersebut tidak jelas harus dimasukan ke dalam *cluster* A atau *cluster* B. Akan tetapi, jika s_i = -1 artinya struktur *cluster* yang dihasilkan *overlapping*, sehingga objek *i* lebih tepat dimasukan ke dalam *cluster* yang lain. Ketika rata-rata hasil dari penjumlahan nilai *s* jumlahnya paling besar diantara *cluster* yang lainnya, maka artinya hasil *cluster* yang dihasilkan merupakan *cluster* yang terbaik karena semakin sedikit nilai *s* yang nilainya dibawah 0.

Perhitungan manual pengujian Improved K-Means dengan menggunakan *silhouette coefficient* dapat ditunjukkan pada Tabel 4.9.

Tabel 4.9 Tabel *Dataset* Ter*cluster* (*k*=2)

Record	A1	A2	A3	*Cluster*
1	577	6110	542	1
2	301	1893	285	2
3	35	244	33	2
4	11	63	11	2
5	379	1896	377	2
6	210	1106	197	2
7	909	4305	895	1
8	901	3126	899	1
9	169	450	164	2
10	853	2560	844	1

Langkah 1 : Menghitung *α*(i), dimana *α*(i) adalah jarak rata-rata antara record satu dengan record lain dalam satu *cluster.*

Langkah 2 : Menghitung *b*(i), dimana *b*(i) adalah jarak rata-rata antara record yang berada pada *cluster* yang lain.

Langkah 3 : Menghitung nilai *silhouette coefficient* nya

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Dari langkah 1 sampai dengan 3, maka hasilnya dapat ditunjukkan pada Tabel 4.10.

Tabel 4.10 Tabel Hasil Perhitungan Manual Pengujian

Record	A1	A2	A3	*Cluster*	a_i	b_i	S_i
1	577	6110	542	1	2821,7	5196,8	0,45
2	301	1893	285	2	1186,6	2329,2	0,49
3	35	244	33	2	953,9	3968,3	0,75
4	11	63	11	2	1097,2	4149,1	0,73
5	379	1896	377	2	1209,3	2279,2	0,46
6	210	1106	197	2	851,3	3085,2	0,72
7	909	4305	895	1	1598,1	3514,6	0,54
8	901	3126	899	1	1590,8	2415,5	0,341
9	169	450	164	2	862,6	3717,5	0,76
10	853	2560	844	1	1963,6	1892,0	-0,03
Rata-Rata							0,52

Contoh perhitungan manualisasi untuk pengujian (i =1) :

Langkah 1 : Menghitung α(i), dimana α(i) adalah jarak rata-rata antara record satu dengan record lain dalam satu *cluster*.

Record (1,7) = $\sqrt{(577-909)^2+(6110-4305)^2+(542-895)^2}$ = 1868,91

Record (1,8) = $\sqrt{(577-901)^2+(6110-3126)^2+(542-899)^2}$ = 3022,69

Record (1,10) = $\sqrt{(577-853)^2+(6110-2560)^2+(542-844)^2}$ = 3573,70

α(1) =(1868,91+3022,69+3573,70)/3 =2821,7

Langkah 2 : Menghitung *b*(i), dimana *b*(i) adalah jarak rata-rata antara record yang berada pada *cluster* yang lain.

Record (1,2) =$\sqrt{(577-301)^2+(6110-1893)^2+(542-285)^2}$ = 4233,83

Record (1,3) =$\sqrt{(577-35)^2+(6110-244)^2+(542-33)^2}$ = 5912,93

Record (1,4) =$\sqrt{(577-11)^2+(6110-63)^2+(542-11)^2}$ = 6096,6

Record (1,5) =$\sqrt{(577-379)^2+(6110-1896)^2+(542-377)^2}$ = 4221,87

Record (1,6) =$\sqrt{(577-210)^2+(6110-1106)^2+(542-197)^2}$ = 5029,28

Record (1,9) =$\sqrt{(577-169)^2+(6110-450)^2+(542-164)^2}$ = 5687,26

b(1) = (4233,83+5912,93+6096,6+4221,87+5029,28+5687,26)/6 = 5196,96

Langkah 3 : Menghitung nilai *silhouette coefficient* nya

$$s(1) = \frac{b(1)-a(1)}{max(a(1),b(1))}$$

$$= \frac{5196,96-2821,70}{5196,96}$$

$$= 0,45$$

BAB 5 IMPLEMENTASI

Pada bab implementasi dibahas mengenai implementasi sistem yang dibuat baik perangkat keras maupun perangkat lunak, batasan-batasan implementasi, dan implementasi program berupa proses *clustering* menggunakan Improved K-Means berdasarkan analisis kebutuhan dan proses perancangan.

5.1 Implementasi Sistem

Perangkat lunak sistem *clustering* jurnal internasional dikembangkan dalam lingkungan implementasi yang terdiri dari perangkat keras dan perangkat lunak.

5.1.1 Implementasi Perangkat Keras

Implementasi perangkat keras yang dipakai dalam proses pembuatan sistem dijelaskan pada Tabel 5.1.

Tabel 5.1 Implementasi Perangkat Keras Komputer

Ultrabook Dell Inspiron 14R – 5420	
Nama *Hardware*	**Spesifikasi**
Processor	Intel® Core™ i3-3317U (2M Cache, up to 2.4 GHz)
Memory (RAM)	4GB DDR3
Harddisk	500GB SATA (5400RPM)
Motherboard	Dell Inc.
Graphic Card	NVIDIA® GeForce® GT 630M with 1GB GDDR5 VRAM
Monitor	14.0” HD+ (900p) Truelife Infinity Display
Webcam	Skype-Certified Hi-Def Webcam

5.1.2 Implementasi Perangkat Lunak

Implementasi perangkat lunak yang dipakai dalam proses pembuatan sistem dijelaskan pada Tabel 5.2.

Tabel 5.2 Implementasi Perangkat Lunak Komputer

Ultrabook Dell Inspiron 14R – 5420	
Nama *Software*	**Spesifikasi**
Sistem operasi	Microsoft Windows 7 Home Basic 64-bit
Bahasa pemrograman	JAVA
Tool Pemrograman	Netbeans

5.2 Batasan Implementasi

Beberapa batasan dalam mengimplementasikan perangkat lunak *clustering* jurnal internasional adalah sebagai berikut:

1. Perangkat Lunak sistem *clustering* jurnal internasional dirancang dan dijalankan dengan menggunakan *Java Apllication*.
2. File yang digunakan memiliki format *.java* dan *.xls*.

5.3 Implementasi Program

Terdapat beberapa proses dalam program *clustering* jurnal internasional dalam penelitian ini. Adapun proses-prosesnya adalah sebagai berikut:

5.3.1 Implementasi Improved K-Means

5.3.1.1 Implementasi Pembentukan *Centroid* Awal

***Sourcecode* 5.1** Implementasi Pembentukan *Centroid* Awal

```
For (int i = 0; i < (data.size() - 1); i++) {
 for (int j = i + 1; j < data.size(); j++) {
  int x[] = (int[]) data.get(i);
  int y[] = (int[]) data.get(j);

   double dist = 0;
  for (int k = 0; k < Data[0].length; k++) {
   dist = dist + Math.pow((x[k] - y[k]), 2);
  }

  dist = Math.sqrt(dist);

  if (min < 0) {
   min = dist;
   index1 = x;
   index2 = y;
  } else if (dist < min) {
    min = dist;
    index1 = x;
    index2 = y;
   }
  }
 }

 data.remove(index1);
 data.remove(index2);

 centroid[m].add(index1);
```

```
29    centroid[m].add(index2);
30
31    while (centroid[m].size() < (int)(Threshold * (Data.length)
32   /nCluster)) {
33     min = -1;
34     for (int i = 0; i < (data.size()); i++) {
35      for (int j = 0; j < centroid[m].size(); j++) {
36       int x[] = (int[]) data.get(i);
37       int y[] = (int[]) centroid[m].get(j);
38
39       double dist = 0;
40
41       for (int k = 0; k < Data[0].length; k++) {
42        dist = dist + Math.pow((x[k] - y[k]), 2);
43       }
44
45       dist = Math.sqrt(dist);
46
47       if (min < 0) {
48        min = dist;
49        index1 = x;
50       } else if (dist < min) {
51        min = dist;
52        index1 = x;
53       }
54      }
55     }
56
57     data.remove(index1);
58     centroid[m].add(index1);
59    }
60   }
61   for(int l=0;l<Data[0].length;l++){
62     InitialCentroid[i][l] = InitialCentroid[i][l]/
63   centroid[i].size();
64    }
65   }
66   return InitialCentroid;
```

Penjelasan *Sourcecode* 5.1 Implementasi pembentukan *centroid* awal adalah:

- Baris 1 adalah perulangan sebanyak *centroid* inputan.
- Baris 2 adalah membentuk *centroid* baru.

- Baris 3 – 4 adalah untuk menyimpan indeks dari 2 data terdekat.
- Baris 7 – 23 adalah proses mennghitung jarak data dengan *euclidean distance.*
- Baris 25 – 26 adalah proses menghapus dua data terdekat yang diperoleh.
- Baris 28 – 29 adalah proses menambahkan data yang dihapus ke *centroid* baru.
- Baris 31 adalah perulangan selama jumlah data di *centroid* kurang dari *threshold*.
- Baris 34 – 55 adalah proses mencari data terdekat dari *centroid* baru.
- Baris 39 – 45 adalah proses menghitung jarak data dengan *euclidean distance.*
- Baris 57 adalah proses menghapus data terdekat.
- Baris 58 adalah proses menambahkan ke *centroid* baru.
- Proses diulangi dari baris 1 untuk membuat *centroid* berikutnya.
- Baris 61 – 65 adalah proses menghitung rata-rata dari *centroid* yang sudah dibentuk.
- Baris 66 adalah output berupa rata-rata dari setiap *centroid* sebagai *centroid* awal.

5.3.1.2 Implementasi Penempatan Data Pada *Cluster*

***Sourcecode* 5.2** Implementasi Penempatan Data pada *Cluster*

```
1  private void KMeansCore(int Data[][], int nCluster, double
2  InitialCentroid[][]) {
3    double NearestDistance[] = new double[data.size()];
4     for (int i = 0; i < (data.size()); i++) {
5     int x[] = (int[]) data.get(i);
6
7    NearestDistance[i] = -1;
8    for (int j = 0; j < InitialCentroid.length; j++) {
9    double dist = 0;
10   for (int k = 0; k < Data[0].length; k++) {
11   dist = dist + Math.pow(x[k] - InitialCentroid[j][k],2);
12   }
13   dist = Math.sqrt(dist);
14
15       if (NearestDistance[i] < 0) {
16         NearestDistance[i] = dist;
17         FinalCluster[i] = j;
18       } else if (dist < NearestDistance[i]) {
19         NearestDistance[i] = dist;
20         FinalCluster[i] = j;
```

```
21          }
22         }
23        }
24
25  double[][] centroid = new double[nCluster][Data[0].length];
26  for (int i = 0; i < centroid.length; i++) {
27   int counter = 0;
28   for (int j = 0; j < data.size(); j++) {
29    int x[] = (int[]) data.get(j);
30    if (FinalCluster[j] == i) {
31    for (int k = 0; k < centroid[0].length; k++) {
32    centroid[i][k] = centroid[i][k] + x[k];
33    }
34    counter++;
35    }
36    }
37   for (int k = 0; k < centroid[0].length; k++) {
38   centroid[i][k] = centroid[i][k] / counter;
39   }
40  }
41  int newCluster[] = new int[FinalCluster.length];
42  for(int i=0;i<newCluster.length;i++){
43   newCluster[i] = FinalCluster[i];
44  }
45
46  double newCentroid[][] = new double[centroid.length]
47  [centroid[0].length];
48  for(int i=0;i<centroid.length;i++){
49    for(int j=0;j<centroid[0].length;j++){
50     newCentroid[i][j] = centroid[i][j];
51    }
52    }
53
54    AllCluster.add(newCluster);
55    AllCentroid.add(newCentroid);
56
57    boolean konvergen;
58    do {
59      konvergen = true;
60      for (int i = 0; i < (data.size()); i++) {
61        int x[] = (int[]) data.get(i); //current data
62        double dist = 0;
63        for (int k = 0; k < Data[0].length; k++) {
64         dist = dist + Math.pow(x[k] - centroid
```

```
[FinalCluster[i]][k], 2);
    }
   dist = Math.sqrt(dist);

   if (dist <= NearestDistance[i]) {
     NearestDistance[i] = dist;
   } else if (dist > NearestDistance[i]) {
   NearestDistance[i] = dist;
   for (int j = 0; j < centroid.length; j++) {
     dist = 0;
   for (int k = 0; k < Data[0].length; k++) {
     dist = dist + Math.pow(x[k] - centroid[j][k], 2);
    }
   dist = Math.sqrt(dist);

   if (dist <= NearestDistance[i]) {
     NearestDistance[i] = dist;
     if (FinalCluster[i] != j) {
       FinalCluster[i] = j;
       konvergen = false;
     }
   }
  }
 }
 }

  centroid = new double[nCluster][Data[0].length];
  for (int i = 0; i < centroid.length; i++) { //i centroid
   int counter = 0;
   for (int j = 0; j < data.size(); j++) {//j data
    int x[] = (int[]) data.get(j);
    if (FinalCluster[j] == i) {
     for (int k = 0; k < centroid[0].length; k++) {//k fitur
     centroid[i][k] = centroid[i][k] + x[k];
   }
    counter++;
   }
  }
  for (int k = 0; k < centroid[0].length; k++) {
   centroid[i][k] = centroid[i][k] / counter;
  }
 }

  newCluster = new int[FinalCluster.length];
```

```
109   for (int i = 0; i < newCluster.length; i++) {
110    newCluster[i] = FinalCluster[i];
111   }
112   newCentroid=newdouble[centroid.length][centroid[0].length];
113    for (int i = 0; i < centroid.length; i++) {
114    for (int j = 0; j < centroid[0].length; j++) {
115     newCentroid[i][j] = centroid[i][j];
116    }
117   }
118  AllCluster.add(newCluster);
119  AllCentroid.add(newCentroid);
120  } while (konvergen != true);
```

Penjelasan *Sourcecode* 5.2 Implementasi penempatan data pada *cluster* adalah:

- Baris 1-23 adalah proses mengklasifikasikan data ke *centroid* terdekat, nilai *centroid* dimasukkan pada variabel Final*Cluster*.
- Baris 26-40 adalah menghitung nilai *centroid* baru dari rata-rata setiap data pada *cluster* tersebut.
- Baris 58-84 adalah menghitung jarak data dengan *cluster* akhir terdekat. Jika jarak sama atau kurang dari *cluster* akhir yang terdekat, data tetap berada dalam *cluster.* Jika tidak maka data akan berpidah *cluster*.
- Baris 91-106 adalah update *centroid*.
- Baris 120 adalah ketika kriteria konvergen ditemui, yaitu ketika tidak ada data yang berpindah *cluster.*

5.3.2 Implementasi K-Means

***Sourcecode* 5.3** Implementasi K-Means

```
1   public KMeans(int Data[][], int nCluster) {
2    this.Data = Data;
3    this.nCluster = nCluster;
4    this.InitialCentroid = generateCentroid(Data, nCluster);
5    KMeansCore(Data, nCluster, InitialCentroid);
6   }
7   private double[][] generateCentroid(int Data[][], int
8   nCluster) {
9    Random rand = new Random();
10   double[][] centroid = new double[nCluster][Data[0].length];
11
12   ArrayList data = new ArrayList();
13   data.addAll(Arrays.asList(Data));
```

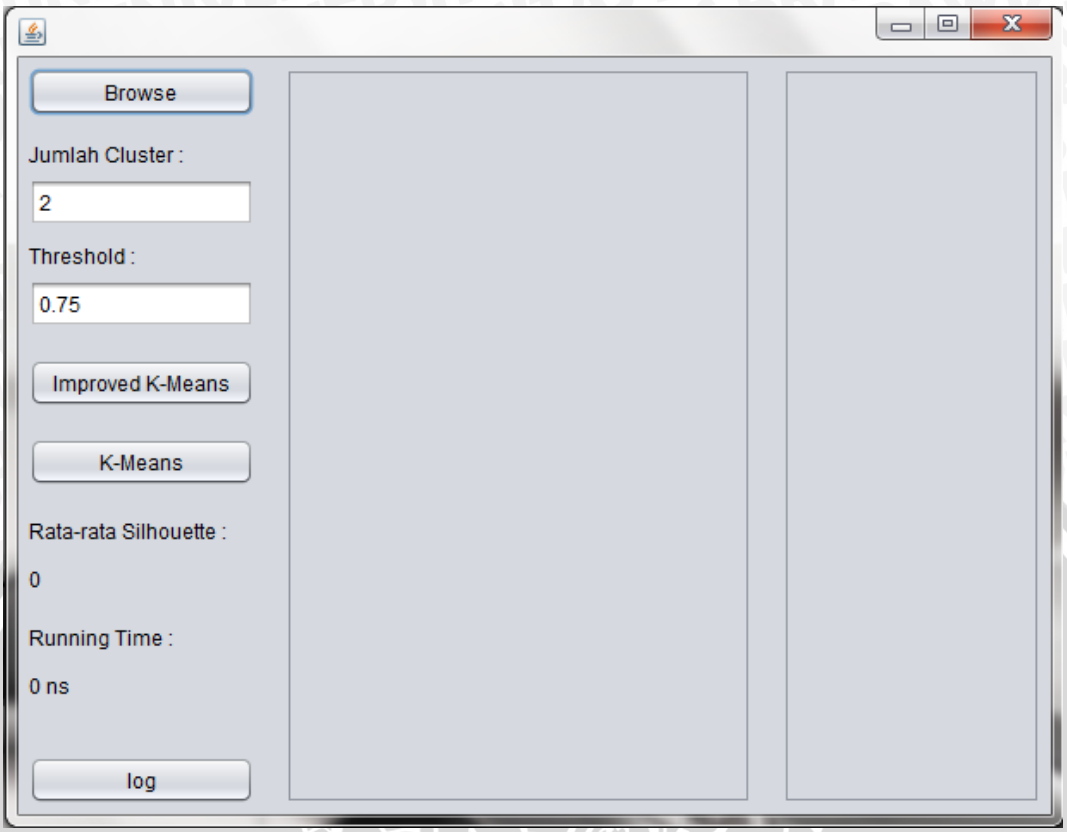
```
14
15    for (int i = 0; i < nCluster; i++) {
16     int n = rand.nextInt(data.size());
17     int x[] = (int[])data.get(n);
18     for(int j=0;j<Data[0].length;j++){
19      centroid[i][j] = (double)x[j];
20     }
21     data.remove(n);
22    }
23    return centroid;
24   }
25
26   for (int i = 0; i < (data.size()); i++) {
27    int x[] = (int[]) data.get(i);
28    NearestDistance[i] = -1;
29
30    for (int j = 0; j < InitialCentroid.length; j++) {
31     double dist = 0;
32     for (int k = 0; k < Data[0].length; k++) {
33      dist = dist + Math.pow(x[k] - InitialCentroid[j][k], 2);
34     }
35     dist = Math.sqrt(dist);
36
37     if (NearestDistance[i] < 0) {
38      NearestDistance[i] = dist;
39      FinalCluster[i] = j;
40     } else if (dist < NearestDistance[i]) {
41      NearestDistance[i] = dist;
42      FinalCluster[i] = j;
43     }
44    }
45   }
46
47   for (int i = 0; i < centroid.length; i++) {
48    int counter = 0;
49    for (int j = 0; j < data.size(); j++) {
50     int x[] = (int[]) data.get(j);
51     if (FinalCluster[j] == i) {
52     for (int k = 0; k < centroid[0].length; k++) {
53
54     centroid[i][k] = centroid[i][k] + x[k];
55     }
56     counter++;
57    }
```

```
58  }
59  for (int k = 0; k < centroid[0].length; k++) {
60   centroid[i][k] = centroid[i][k] / counter;
61   }
62  }
63
64  Centroid = new double[nCluster][Data[0].length];
65   for (int i = 0; i < centroid.length; i++) {
66    int counter = 0;
67    for (int j = 0; j < data.size(); j++) {
68     int x[] = (int[]) data.get(j);
69     if (FinalCluster[j] == i) {
70      for (int k = 0; k < centroid[0].length; k++) {
71
72       centroid[i][k] = centroid[i][k] + x[k];
73
74      }
75      counter++;
76     }
77    }
78    for (int k = 0; k < centroid[0].length; k++) {
79     centroid[i][k] = centroid[i][k] / counter;
80    }
81   }
```

Penjelasan *Sourcecode* 5.3 Implementasi K-Means adalah:

- Baris 1-6 adalah konstruktor untuk K-Means standar. Nilai *centroid* awal di generate random melalui method generate*Centroid*().
- Baris 7-24 adalah method generate *centroid*. Nomer index di generate, lalu diambil data dari index yang terpilih. Pada data.remove(n), hapus data dari index yang terpilih agar data tersebut tidak terpilih sebagai *centroid* berikutnya. Setiap *centroid* asti beda data. Hasilnya berupa output dari method generate *centroid*.
- Baris 26-45 adalah proses mengklasifikasikan data ke *centroid* terdekat, nilai *centroid*nya dimasukkan ke variabel Final*Cluster*.
- Baris 47-62 adalah menghitung nilai *centroid* baru dari rata- rata setiap data pada *cluster* tersebut.
- Baris 64-81 adalah update *centroid*.

5.4 Implementasi Interface



Gambar 5.1. Tampilan Awal

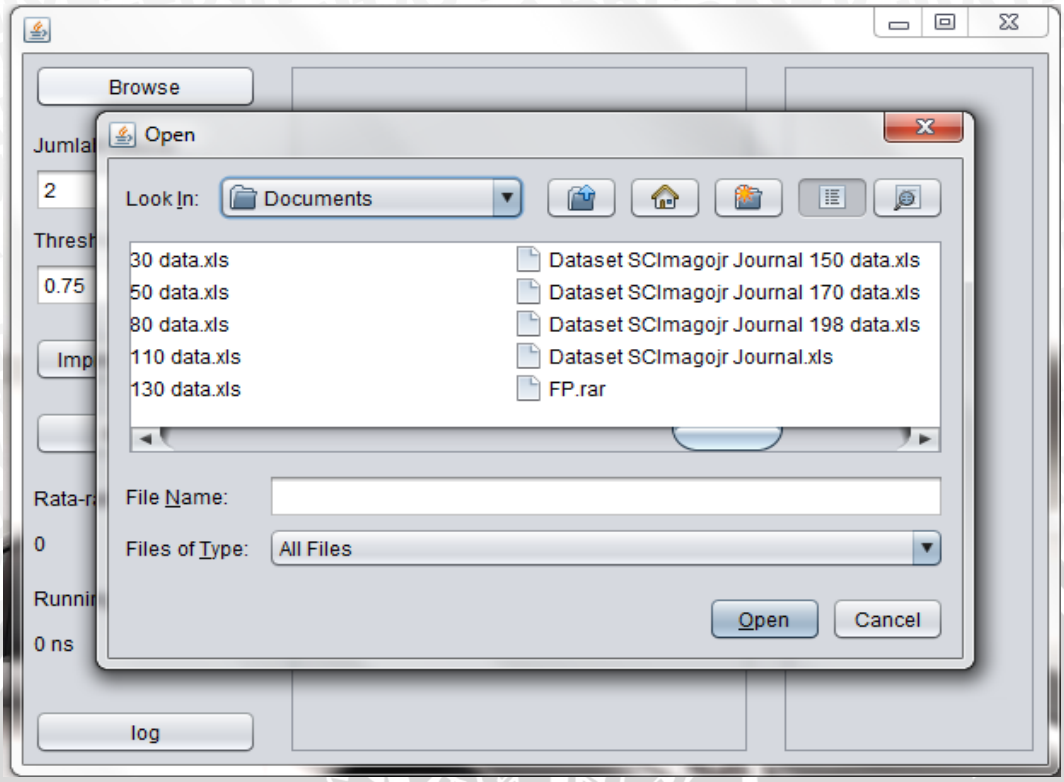
Gambar 5.1 menampilkan informasi umum tentang program *cluster*ing portal jurnal internasional menggunakan Improved K-Means dan K-Means Standar. Halaman antarmuka akan tampil pertama kali pada saat program dijalankan.

Gambar 5.2 berisi tampilan *browse* data. User dapat memilih data tersimpan yang akan di *cluster*. Data yang digunakan untuk *clustering* memiliki format .xls.

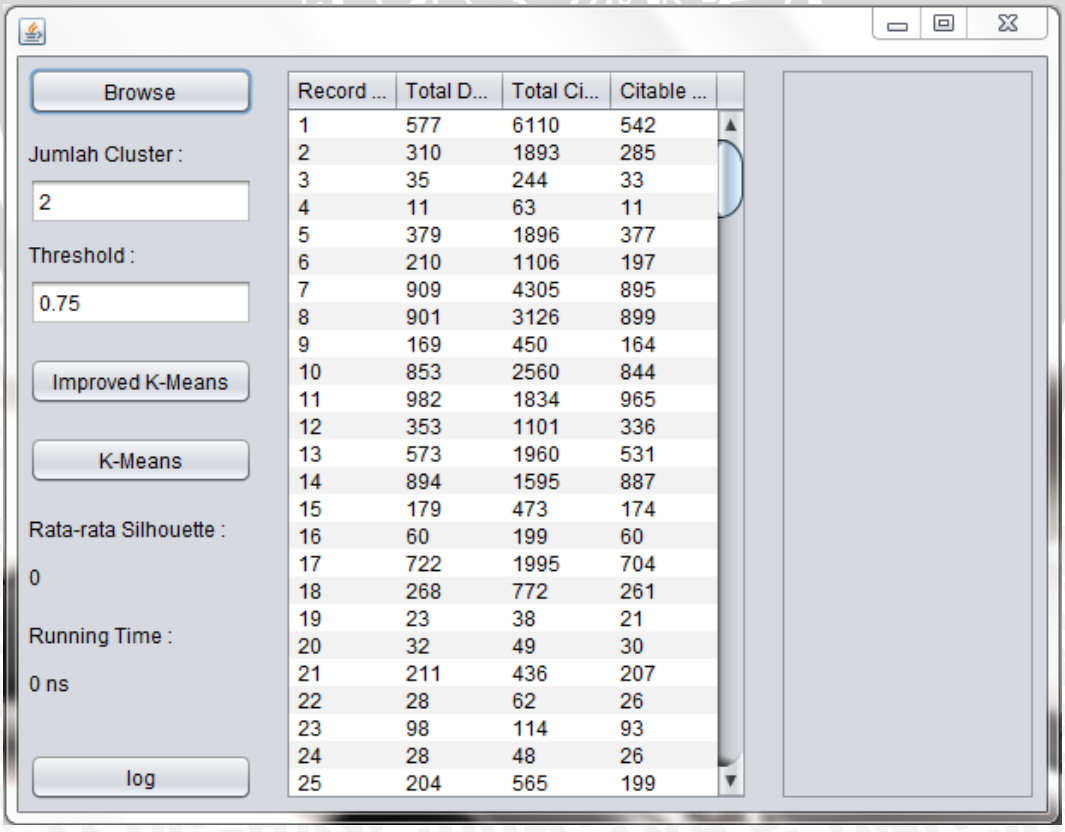
Gambar 5.3 menampilkan data sebelum di *cluster* dengan Improved K-Means. Terdapat 4 kolom yang terdiri dari Record, Total Docs (3 years), Total Cites (3 years), Citable Docs (3 years).

Gambar 5.4 menampilkan data yg setelah di *cluster* dengan Improved K-Means. Pada sisi kanan dari *dataset* dapat dilihat hasil dari *clustering* dimana terdapat 3 kolom yaitu Record, *Cluster*, dan Silhouette.

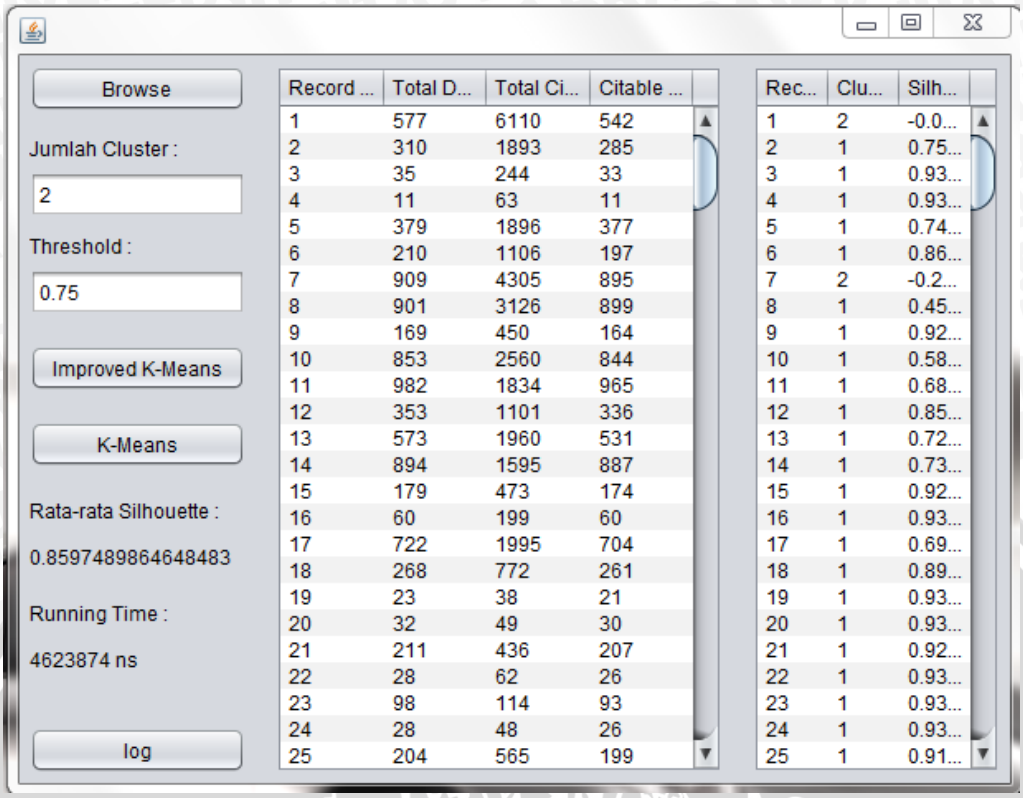
Gambar 5.5 menampilkan log dari proses Improved K-Means. User dapat melihat banyanya jumlah iterasi dan perubahan *centroid* pada setiap iterasinya.



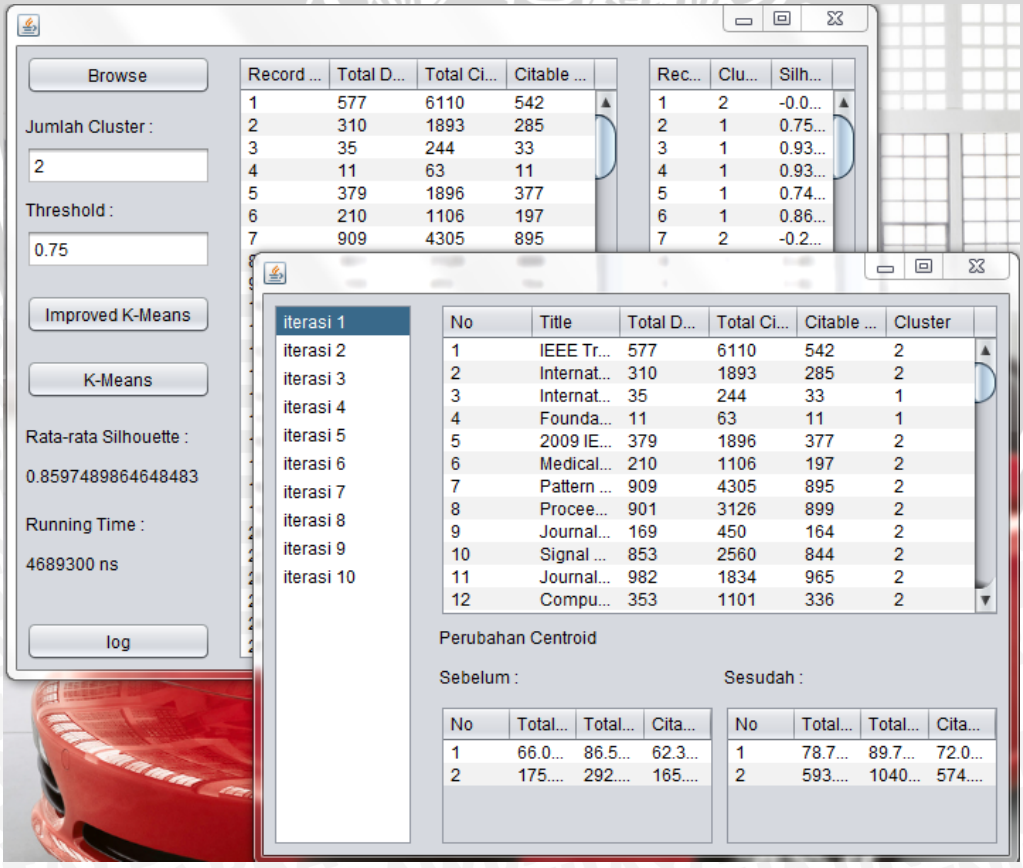
Gambar 5.2. Tampilan Browse Data



Gambar 5.3. Tampilan Sebelum Proses Improved K-Means



Gambar 5.4. Tampilan Setelah Proses Improved K-Means



Gambar 5.5. Tampilan Log

BAB 6 PENGUJIAN

Bab ini membahas mengenai proses pengujian dari implementasi metode Improved K-Means pada portal jurnal internasional. Proses pengujian akan dilakukan dengan 3 skenario, yaitu pengujian 198 data dengan menggunakan percobaan 2 sampai 25 *cluster*, pengujian menggunakan 10, 30, 50, 80, 110, 150, 170 dan 198 data dan pengujian ketiga yaitu percobaan 2 sampai 25 *cluster* pada metode Improved K-Means kemudian dibandingkan dengan hasil *silhouette coefficient* pada pengujian K-Means Standar.

6.1 Pengujian *Cluster* Improved K-Means

Pengujian *cluster* ini dilakukan untuk mengetahui berapa jumlah *cluster* terbaik yang menghasilkan nilai s*ilhouette coefficient* tertinggi untuk digunakan pada *clustering* portal jurnal internasional. Nilai kualitas *cluster* memiliki rentang dari -1 sampai 1.

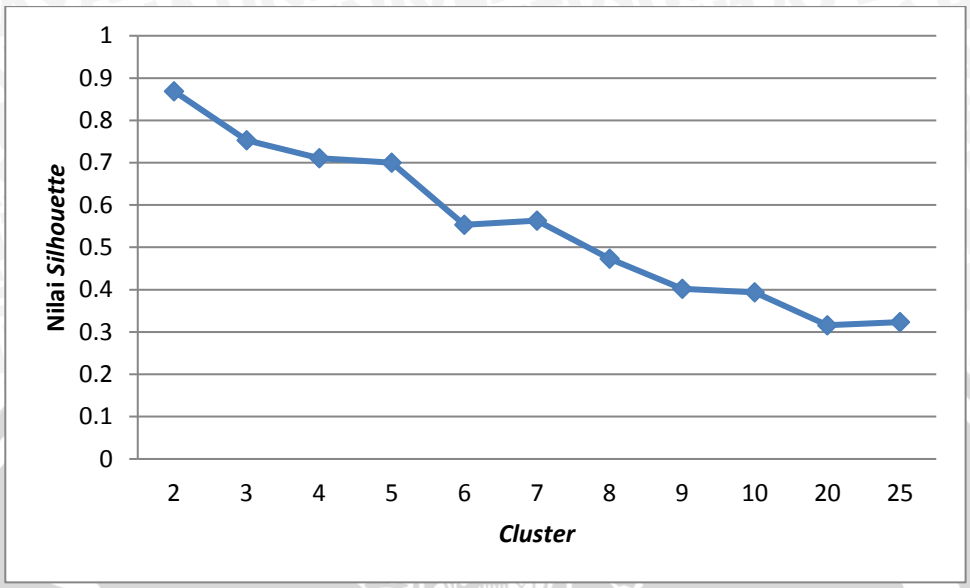
Pengujian jumlah *cluster* dilakukan dengan menggunakan 198 data portal jurnal internasional dengan 3 fitur. Jumlah *cluster* yang digunakan pada pengujian ini adalah 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, dan 25. Adapun hasil pengujian jumlah *cluster* ditunjukkan pada Tabel 6.1.

Tabel 6.1 Pengujian Jumlah *Cluster* Improved K-Means

Cluster	**Nilai *Silhouette***
2	0,859
3	0,753
4	0,711
5	0,7
6	0,553
7	0,563
8	0,473
9	0,402
10	0,394
20	0,316
25	0,323

Pada Tabel 6.1 dapat dilihat bahwa hasil pengujian yang dilakukan dengan metode *silhouette coefficient* menghasilkan nilai tertinggi pada penggunaan 2 *cluster*. Semakin tinggi jumlah *cluster* semakin rendah nilai *silhouette* yang di

dapat. Adapun hasil dari pengujian dalam bentuk statistik ditunjukkan pada Gambar 6.1.



Gambar 6.1 Statistik Nilai *Silhouette Cluster* 2 sampai 25

Berdasarkan grafik hasil pengujian yang ditunjukkan Gambar 6.1 dapat dilihat bahwa pengujian dengan menggunakan jumlah *cluster* sebanyak 2 sampai dengan 25 *cluster* menghasilkan nilai *silhouette* paling tinggi pada jumlah *cluster* 2, hal tersebut dipengaruhi oleh jarak kerapatan data pada satu *cluster* memiliki jarak yang cukup baik dan jarak kerapatan data *cluster* satu dengan *cluster* lainnya cukup besar. Penurunan nilai *silhouette* pada penggunaan *cluster* yang besar dipengaruhi oleh jumlah *cluster* yang digunakan dan *centroid* yang terbentuk sehingga memberikan pengaruh pada kerapatan data pada *cluster* dan jarak antar *cluster* satu dengan *cluster* yang lainnya. Pada hasil pengujian tersebut menunjukkan bahwa penggunaan 2 *cluster* pada penelitian ini merupakan jumlah yang tepat karena memiliki kualitas nilai *silhouette* yang paling baik.

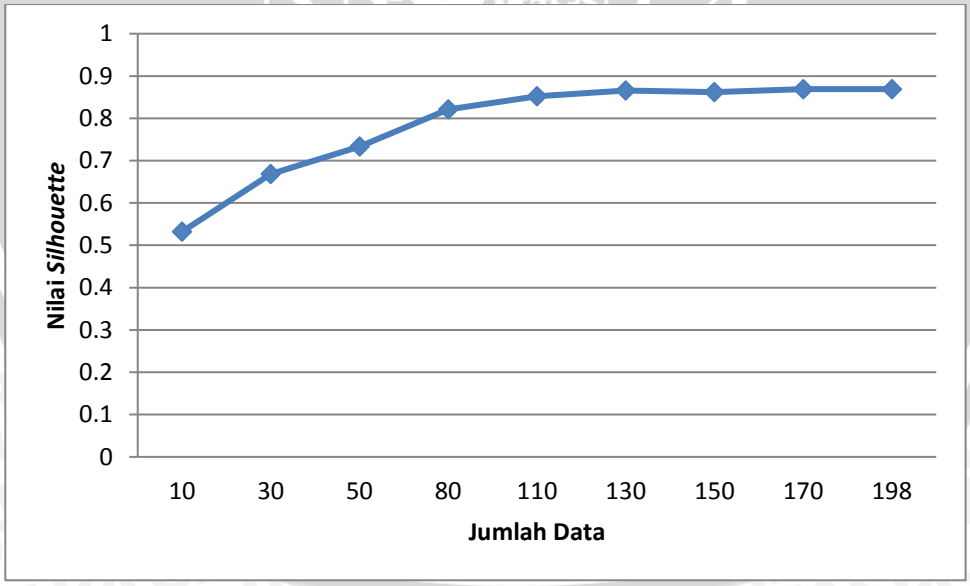
6.2 Pengujian Jumlah Data Improved K-Means

Pengujian jumlah data ini dilakukan untuk mengetahui seberapa besar efek dari banyak data terhadap hasil dari pengujian *silhouette coefficient.* Pengujian jumlah data dilakukan dengan menggunakan 2 *cluster* karena pada pengujian sebelumnya memiliki nilai s*ilhouette coefficient* yang paling bagus. Jumlah data yang akan di uji pada pengujian ini adalah 10, 30, 50, 80, 110, 130, 150, 170, dan 198 data. Adapun hasil pengujian jumlah data ditunjukkan pada Tabel 6.2.

Tabel 6.2 Pengujian Jumlah Data Improved K-Means

Jumlah Data	Nilai *Silhouette*
10	0,532
30	0,668
50	0,733
80	0,821
110	0,852
130	0,866
150	0,862
170	0,869
198	0,869

Pada Tabel 6.2 dapat dilihat bahwa hasil pengujian jumlah data yang dilakukan dengan metode *silhouette coefficient* menghasilkan nilai tertinggi yaitu pada penggunaan 198 data dengan 2 *cluster.* Semakin sedikit jumlah data yang diuji maka semakin rendah nilai *silhouette* yang di dapat. Adapun hasil dari pengujian dalam bentuk statistik ditunjukkan pada Gambar 6.2.



Gambar 6.2 Statistik Nilai *Silhouette* Data 10 sampai 198

Berdasarkan grafik hasil pengujian jumlah *cluster* yang ditunjukkan pada Gambar 6.2 dapat dilihat bahwa pengujian dengan menggunakan data sebanyak 198 data dengan 2 *cluster* menghasilkan nilai *silhouette* yang paling tinggi, hal

tersebut dipengaruhi oleh jarak antar data pada *cluster* memiliki kerapatan yang cukup baik dan jarak data antar *cluster* memiliki nilai yang cukup besar. Penurunan nilai *silhouette* pada penggunaan data yang sedikit dipengaruhi oleh jumlah data yang digunakan dan *centroid* yang terbentuk sehingga memberikan pengaruh pada kerapatan data pada *cluster* dan jarak antar *cluster* satu dengan *cluster* lainnya. Pada hasil pengujian tersebut menunjukkan bahwa penggunaan 198 data pada penelitian ini merupakan jumlah yang tepat karena memiliki kualitas nilai *silhouette* yang baik.

6.3 Perbandingan Pengujian Improved K-Means dengan K-Means Standar

Sub bab ini membandingkan hasil dari pengujian yang dilakukan oleh metode Improved K-Means dengan metode K-Means Standar. Adapun poin yang dilakukan perbandingan sub bab ini yaitu pada jumlah *cluster.* Jumlah *cluster* yang digunakan pada pengujian ini adalah 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, dan 25. Pada tiap pengujian *cluster* dilakukan 5 kali *run* program. Adapun hasil pengujian ditunjukkan pada Tabel 6.3 sampai Tabel 6.7 pengujian *cluster*.

Tabel 6.3 Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-1

Cluster	**Nilai *Silhouette***	
	Improved K-Means	**K-Means Standar**
2	0,869	0,869
3	0,753	0,753
4	0,711	0,711
5	0,7	0,721
6	0,553	0,553
7	0,563	0,548
8	0,473	0,513
9	0,402	0,403
10	0,394	0,413
20	0,316	0,298
25	0,323	0,348

Tabel 6.4 Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-2

Cluster	**Nilai *Silhouette***	
	Improved K-Means	**K-Means Standar**
2	0,869	0,869

3	0,753	0,759
4	0,711	0,71
5	0,7	0,721
6	0,553	0,549
7	0,563	0,562
8	0,473	0,473
9	0,402	0,412
10	0,394	0,404
20	0,316	0,349
25	0,323	0,324

Tabel 6.5 Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-3

Cluster	**Nilai *Silhouette***	
	Improved K-Means	**K-Means Standar**
2	0,869	0,869
3	0,753	0,848
4	0,711	0,711
5	0,7	0,648
6	0,553	0,55
7	0,563	0,563
8	0,473	0,431
9	0,402	0,387
10	0,394	0,413
20	0,316	0,344
25	0,323	0,335

Tabel 6.6 Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-4

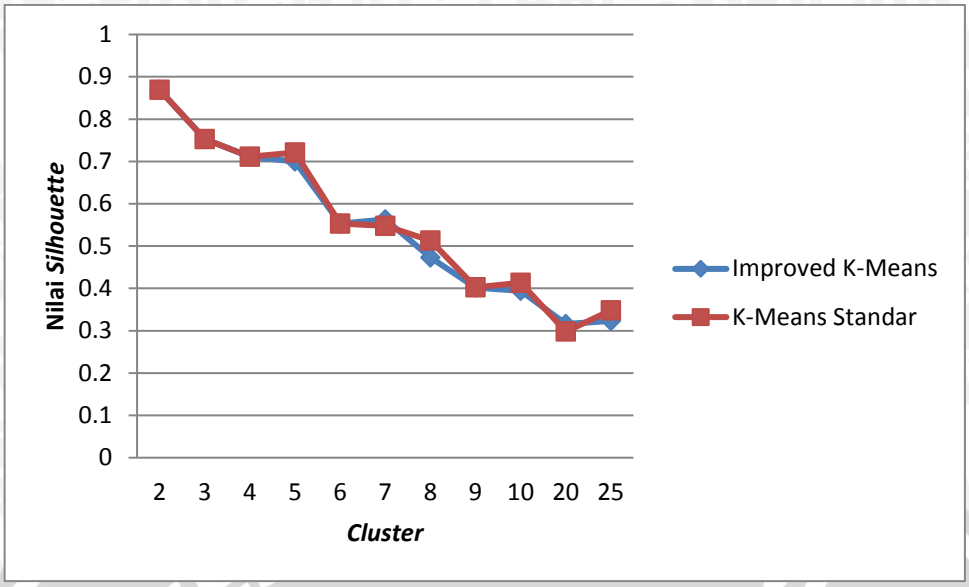
Cluster	**Nilai *Silhouette***	
	Improved K-Means	**K-Means Standar**
2	0,869	0,869
3	0,753	0,848

4	0,711	0,711
5	0,7	0,7
6	0,553	0,553
7	0,563	0,562
8	0,473	0,514
9	0,402	0,482
10	0,394	0,469
20	0,316	0,329
25	0,323	0,328

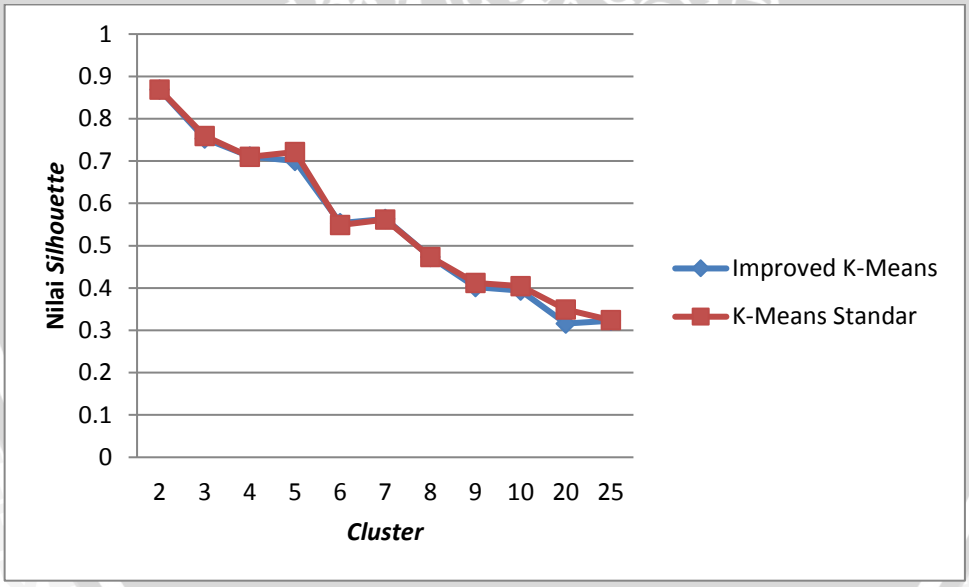
Tabel 6.7 Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-5

Cluster	**Nilai *Silhouette***	
	Improved K-Means	**K-Means Standar**
2	0,869	0,869
3	0,753	0,753
4	0,711	0,763
5	0,7	0,584
6	0,553	0,709
7	0,563	0,563
8	0,473	0,546
9	0,402	0,361
10	0,394	0,415
20	0,316	0,34
25	0,323	0,324

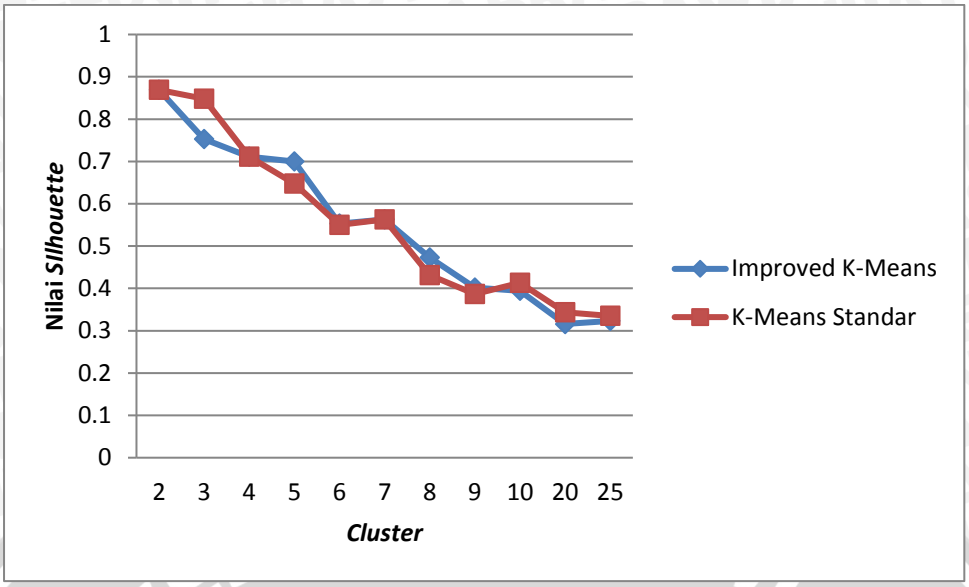
Pada Tabel 6.3 sampai Tabel 6.7 dapat dilihat bahwa hasil pengujian yang dilakukan dengan metode *silhouette coefficient* pada metode K-Means standar menunjukkan bahwa penggunaan 2 *cluster* memiliki nilai *silhouette* paling tinggi dibanding dengan menggunakan jumlah *cluster* yang lain. Adapun hasil dalam bentuk statistik ditunjukkan pada Gambar 6.3 sampai Gambar 6.7.



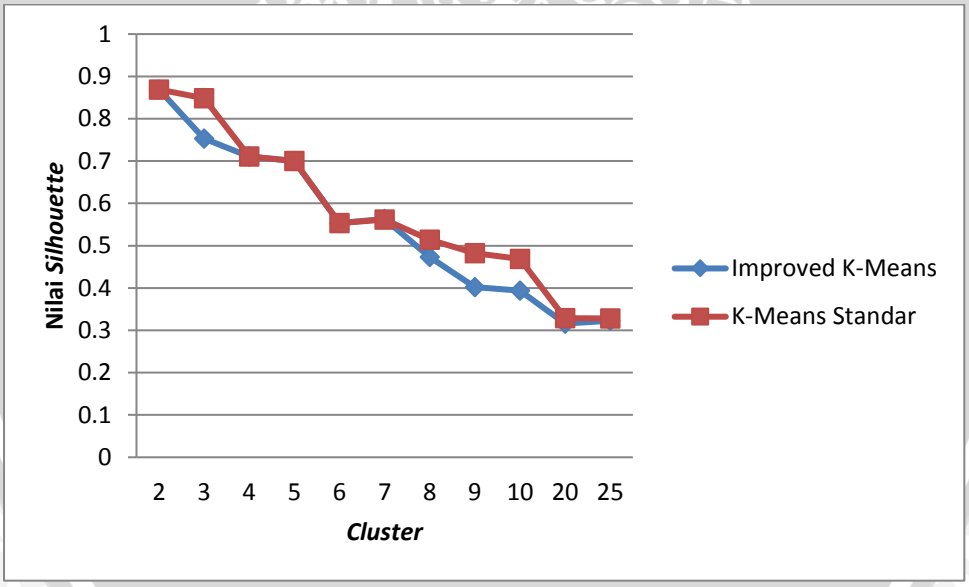
Gambar 6.3 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-1



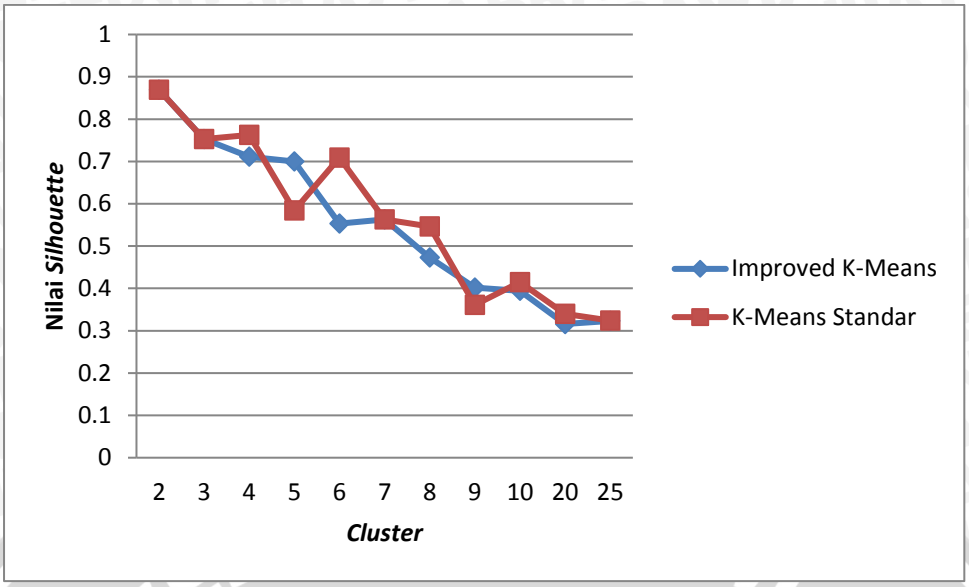
Gambar 6.4 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-2



Gambar 6.5 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-3



Gambar 6.6 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-4



Gambar 6.7 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-5

Berdasarkan grafik hasil perbandingan nilai *silhouette* pengujian *cluster* yang ditunjukkan pada Gambar 6.3 sampai Gambar 6.7 dapat dilihat bahwa pengujian jumlah *cluster* metode Improved K-Means memiliki kecenderungan nilai *silhouette* yang hampir sama terhadap nilai *silhouette* yang dihasilkan pada metode K-Means Standar pada penggunaan 2 sampai 25 *cluster.* Akan tetapi, nilai *silhouette* pada K-Means Standar tidaklah mutlak karena pada K-Means Standar inisialisasi *centroid* awal ditentukan secara acak sehingga memungkinkan memiliki nilai *silhouette* yang lebih baik ataupun lebih buruk.

BAB 7 PENUTUP

7.1 Kesimpulan

Berdasarkan hasil perancangan, implementasi dan pengujian yang telah dilakukan, maka didapatkan kesimpulan antara lain sebagai berikut:

1. Metode Improved K-Means dapat diterapkan pada *clustering* portal jurnal internasional dimana memiliki hasil yang lebih stabil dibandingkan dengan metode K-Means Standar.
2. Metode Improved K-Means dapat di implementasikan pada *clustering* portal jurnal internasional dengan nilai kualitas *cluster* terbaik yang didapatkan adalah 0,869 untuk 198 data dengan penggunaan 2 *cluster.*
3. Penambahan dan pengurangan jumlah *cluster* pada metode Improved K-Means pada *clustering* portal jurnal internasional memiliki pengaruh terhadap nilai dari kualitas *cluster.* Pada penggunaan 2 *cluster* memiliki nilai kualitas yang paling baik daripada penggunaan jumlah *cluster* yang lain.
4. Penambahan dan pengurangan data pada Improved K-Means pada *clustering* portal jurnal internasional memiliki pengaruh terhadap meningkat atau menurunnya kualitas *cluster.* Pada penggunaan 198 data dengan 2 *cluster* memiliki nilai kualitas yang lebih baik daripada penggunaan data yang lebih sedikit.
5. Pada pengujian jumlah *cluster* metode Improved K-Means memiliki kecenderungan nilai *silhouette* yang hampir sama terhadap nilai *silhouette* yang dihasilkan pada metode K-Means Standar pada penggunaan 2 sampai 25 *cluster*. Akan tetapi, hasil pada K-Means Standar tersebut tidaklah mutlak karena pada metode K-Means Standar inisialisasi *centroid* awal ditentukan secara acak.
6. Dari hasil pengujian yang telah dilakukan sebelumnya, penggunaan metode Improved K-Means memiliki kestabilan pada nilai *silhouette* dibandingkan dengan penggunaan metode K-Means Standar. Penggunaan metode K-Means Standar memiliki kecenderungan nilai *silhouette* yang berubah-ubah karena inisialisasi *centroid* dilakukan secara acak.

7.2 Saran

Berdasarkan hasil penelitian, penulis menyarankan agar hasil penelitian tentang Implementasi Algoritma Improved K-Means pada inisialisasi *centroid* awal untuk *clustering* portal jurnal internasional digunakan di berbagai macam

database untuk membuktikan ketangguhan dan kestabilan nilai *silhouette* algoritmanya. Kemudian pada sistem didesain lebih *user friendly* agar mempermudah *user* untuk menggunakan sistem ini.

DAFTAR PUSTAKA

Ahmed, A.H. & Ashour, W., 2011. An Initialization Method for the K-means Algorithm using RNN and Coupling Degree. International Journal of Computer Applications, XXV(1), pp.1-6.

Al-Zoubi, Moh'd Belal, Mohammad al Rawi, 2008. An Efficient Approach for Computing *Silhouette*Coefficients, Journal of Computer Science.

Berkhin, Pavel, 2002. *Survey of Clustering Data Mining Techniques.* Accrue Software, Inc.

Carlisle, A. and Dozier, G., 2001. An Off-The Shelf PSO, Proceedings of the 2001 Worksop on Particle Swarm Optimization.pp. Indianapolis, IN.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E., 2008. *Comparison of SCImago journal rank indicator with journal impact factor. The FASEB Journal, 22*(8), 2623-2628.

Frand, Jason, 2009. Data Mining: What is Data Mining?. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologite/palace/datamining.htm. Diakses pada 3 Mei 2016.

Garfield, E, 2006. The History and Meaning of The Journal Impact Factor. Journal of the American Medical Association (JAMA), (293):90-93, January 2006. (Abridged version Published).

Han, Jiawei; & Kamber, Micheline, 2001. Data Mining Concepts and Techniques Second Edition. San Francisco: Morgan Kauffman.

Giyanto, Herribertus, 2008. Penerapan Algoritma *Cluster*ing K-Means, K-Medoid, Gath Geva. Tesis Tidak terpublikasi. Yogyakarta: Universitas Gajah Mada.

Joshi, K.D & Nalwade, P.S., 2013. Modified K-Means for Better Initial *Cluster* Centres. International Journal of Computer Science and Mobile Computing, II(7), pp.219-23.

Kantardzic, Mehmed, 2003. Data Mining Concepts Models, Methods, and Algorithm. New Jersey: IEEE.

Kusrini, 2009. Algoritma Data Mining, Penerbit Andi, Yogyakarta.

Manadokota, 2013. Apakah Diseminasi Informasi Itu. http://www.manadokota.go.id/berita-1194-apakah--diseminasi--informasi--itu.html. Diakses Mei 2016.

Mathworks, 2014. Mathworks statistics toolbox: http://www.mathworks.com/help/stats/cophenet.html .

M. Lutfi Firdaus, D.Sc., 2012. Teknik Publikasi Karya Ilmiah di Jurnal Nasional dan Internasional. FKIP UNB Press.

Nazeer and Sebastian, 2009. Improving the Accuracy and Efficiency of the K-Means *Cluster*ing Algorithm. Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

Oded Maimon and Lior Rokach, 2010. Data Mining and Knowledge Discovery Handbook Second Edition. Springer Science+Business Media, LLC 2005.2010.

Osmar R. Zaiane, 1999. CMPUT690 *Principles of Knowledge Discovery in Database Chapter I: Introduction to Data Mining.* University of Alberta.

Pustakaristek, 2016. http://pustaka.ristek.go.id/main/about. Akses pada 3 Mei 2016.

Rossner M., Van Epps H., Hill E, 2007. Show me the data.J. Cell Biol.179, 1091-1092.doi: 10.1083/jcb.200711140.

Saracli et al.,2013. Comparison of Hierarchical *Cluster* Analysis Methods by Cophenetic Correlation. Journal of Inequalities and Applications 2013:203

Scimagojr, 2014. About Us. http://scimagojr.com/aboutus.php. Diakses pada 3 Mei 2016.

Scimagojr, 2016. SCImago Journal & Country Rank. http://www.scimagojr.com/countryrank.php. Diakses pada 3 Mei 2016.

Steinbach, M., G. Karypis and Vipin Kumar, 2000. A comparison of document *clustering* techniques, Minnesota: University of Minnesota, Department of Computer Science and Engineering. http://glaros.dtc.umn.edu/gkhome/fetch/papers/doc*cluster*.pdf

Tahta Alfina, Budi Santosa, dan Ali Ridho Barakbah, 2012. Analisa Perbandingan Metode Hierarchical *Cluster*ing, K-Means dan Gabungan Keduanya dalam *Cluster* Data (Studi kasus: Problem Kerja praktek Jurusan Teknik Industri ITS). Jurnal Teknik ITS Vol. 1, (Sept,2012) ISSN:

Turban, Efraim, et al., 2005. Decision Support Systems and Intelligent Systems 7th Ed. New Jersey: Pearson Education.

LAMPIRAN

Lampiran 1 *Dataset SCImagojr Journal*

No	Title	Total Docs (3 years)	Total Cites (3 years)	Citable Docs (3 years)
1	IEEE Transactions on Pattern Analysis and Machine Intelligence	577	6110	542
2	International Journal of Computer Vision	310	1893	285
3	International Journal of Machine Learning and Cybernetics	35	244	33
4	Foundations and Trends in Computer Graphics and Vision	11	63	11
5	2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009	379	1896	377
6	Medical Image Analysis	210	1106	197
7	Pattern Recognition	909	4305	895
8	Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	901	3126	899
9	Journal of Mathematical Imaging and Vision	169	450	164
10	Signal Processing	853	2560	844
11	Journal of the Optical Society of America A: Optics and Image Science, and Vision	982	1834	965
12	Computer Vision and Image Understanding	353	1101	336
13	IEEE Transactions on Visualization and Computer Graphics	573	1960	531
14	Proceedings of the IEEE International Conference on Computer Vision	894	1595	887
15	Signal Processing: Image Communication	179	473	174
16	Journal of Information Hiding and Multimedia Signal Processing	60	199	60
17	Pattern Recognition Letters	722	1995	704
18	Speech Communication	268	772	261
19	Proceedings - HPG 2011: ACM SIGGRAPH Symposium on High Performance Graphics	23	38	21
20	Proceedings - SCA 2011: ACM SIGGRAPH/ Eurographics Symposium on Computer Animation	32	49	30
21	Journal of Visual Communication and Image Representation	211	436	207

22	IEEE Pacific Visualization Symposium, PacificVis 2009 - Proceedings	28	62	26
23	Computing and Visualization in Science	98	114	93
24	IEEE Pacific Visualization Symposium 2011, PacificVis 2011 - Proceedings	28	48	26
25	Computerized Medical Imaging and Graphics	204	565	199
26	Pattern Analysis and Applications	104	158	102
27	Information Visualization	70	137	66
28	2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011	152	222	151
29	Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding , ASRU 2009	102	94	100
30	Machine Vision and Applications	198	321	183
31	MM'09 - Proceedings of the 2009 ACM Multimedia Conference, with Co-located Workshops and Symposiums	249	370	247
32	Cognitive Computation	109	166	102
33	IEEE Pacific Visualization Symposium 2010, PacificVis 2010 - Proceedings	29	42	27
34	IET Image Processing	148	287	144
35	Intelligent Data Analysis	154	116	136
36	International Journal of Pattern Recognition and Artificial Intelligence	212	238	208
37	Eurographics Workshop on 3D Object Retrieval, EG 3DOR	32	33	28
38	Eye Tracking Research and Applications Symposium (ETRA)	69	133	66
39	International Journal of imaging Systems and Technology	127	185	123
40	International Journal on Document Analysis and Recognition	77	146	72
41	Presence: Teleoperators and Virtual Environments	115	196	108
42	Visual Computer	373	439	315
43	Proceedings - International Conference on Pattern Recognition	1147	927	1144
44	Evolutionary Intelligence	55	71	50
45	IET Computer Vision	88	139	84
46	Journal of Uncertain Systems	28	19	28
47	IEEE International Conference on Intelligent Robots and Systems	788	648	786

48	IEEE-RAS International Conference on Humanoid Robots	118	110	117
49	Journal of Computer and Systems Sciences International	274	100	274
50	International Journal of Biometrics	56	52	56
51	Pattern Recognition and Image Analysis	321	86	317
52	Journal of Advanced Computational Intelligence and Intelligent Informatics	340	149	325
53	Final Program and Proceedings - IS and T/SID Colorr imaging Conference	192	41	187
54	Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation	54	25	53
55	Imaging Science Journal	111	68	106
56	IEICE Transactions on Information and Systems	1055	439	1019
57	International Journal of Speech Technology	62	62	61
58	Moshi Shibie yu Rengong Zhineng/ Pattern Recognition and Artificial Intelligence	395	173	395
59	Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings	7680	2179	7677
60	Forum on Specifications and Design Languages	29	3	27
61	Proceedings of the International Display Workshops	537	49	534
62	ACM International Conference Proceeding Series	6503	412	6138
63	Foundations and Trends in Machine Learning	11	194	11
64	Journal of the ACM	98	524	91
65	Cognitive Psychology	73	447	72
66	ACM Transactions on Intelligent Systems and Technology	75	1100	65
67	IEEE Transactions on Fuzzy Systems	307	2299	302
68	Journal of Memory and Languages	178	673	177
69	International Journal of Robotics Research	307	1559	282
70	Information Sciences	1062	5408	1038
71	Fuzzy Optimization and Decision Making	63	194	63
72	Knowlegde-Based Systems	318	1766	311
73	International Journal of Approximate Reasoning	286	838	263
74	Fuzzy Sets and Systems	612	1894	588
75	Journal of Scheduling	136	256	123

76	IEEE Computational Intelligence Magazine	117	365	86
77	Artificial Intelligence	218	931	213
78	Topics in Cognitive Science	154	495	128
79	Machine Learning	171	639	159
80	Cognitive Science	178	526	171
81	Design Studies	96	314	91
82	Journal of Machine Learning Research	739	2482	724
83	Networks and Spatial Economics	92	143	87
84	Constraints	65	154	59
85	Autonomous Robots	149	575	138
86	Journal of the American Society for Information Science and Technology	618	1934	564
87	Journal of Heuristics	96	230	92
88	Physics of Life Reviews	176	351	68
89	Knowledge and Information Systems	243	631	234
90	Journal of Artificial Intelligence Research	144	624	144
91	Proceedings of the 26th International Conference On Machine Learning, ICML 2009	162	583	160
92	Argument and Computation	17	78	17
93	IEEE Intelligent Systems	231	587	201
94	Expert Systems with Applications	4163	14906	4105
95	IEEE Transactions on Neural Networks	563	2854	540
96	Engineering Applications of Artificial Intelligence	400	1284	392
97	International Journal of Intelligent Systems	187	432	177
98	IEEE Transactions on Autonomous Mental Development	72	250	63
99	Journal of Automated Reasoning	94	189	88
100	Autonomous Agents and Multi - Agent Systems	104	272	96
101	Proceedings of the 2009 ACM SIGPLAN/ SIGOPS International Conference on Virtual Execution Environments, VEE'09	16	99	14
102	Swarm Intelligence	43	109	37
103	Integrated Computer - Aided Engineering	85	240	78
104	AI Communications	64	122	61
105	Journal of Intelligent Manufacturing	241	421	215
106	Neural Networks	399	1286	380
107	Neurocomputing	1128	2987	1087
108	Advanced Engineering Informatics	164	449	152
109	International Journal of Uncertainty, Fuzziness and Knowledge - Based Systems	161	360	157

110	Artificial Intelligence Review	88	306	83
111	Applied Intelligence	163	316	150
112	Cognitive Systems Research	89	199	82
113	Perception	533	612	456
114	Journal of Semantics	43	45	42
115	Theory and Practice of Logic Pregramming	92	149	89
116	Jouenal of Intelligent and Robotic Systems: Theory and Applications	313	590	277
117	Journal of Pragmatics	708	760	639
118	ICML 2010 - Proceedings, 27th International Conference on Machine Learning	161	312	159
119	Synthesis Lectures on Artificial Intelligence and Machine Learning	12	48	11
120	International Journal of Fuzzy Systems	117	204	111
121	Parallel Computing	155	368	145
122	Cybernetics and Systems	122	168	112
123	Artificial Intelligence in Medicine	172	419	163
124	Journal of Intelligent Information Systems	98	163	90
125	Journal of Artificial Intelligence	40	62	37
126	Computational Intelligence	67	129	64
127	Multidimensional Systems and Signal Processing	71	78	63
128	IEEE Transactions on Computational Intelligence and AI in Games	75	357	73
129	VRIPHYS 2010 - 7th Workshop on Virtual Reality Interactions and Physical Simulations	17	13	15
130	Knowledge Engineering Review	62	106	58
131	Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM	85	130	78
132	Proceedings - IEEE International Conference on Robotics and Automation	2435	3427	2429
133	Journal of Parallel and Distributed Computing	332	651	323
134	AI Magazine	126	224	122
135	Neural Processing Letters	109	172	107
136	IJCAI International Joint Conference on Artificial Intelligence	1036	1315	1030
137	i-Perception	56	80	52
138	2009 IEEE Congress on Evolutionary Computation, CEC 2009	449	507	448
139	Intelligent Service Robotics	71	90	67
140	Artificial Life	65	124	63
141	Cognitive Processing	177	210	161

142	Iranian Journal of Fuzzy Systems	104	83	103
143	Kybernetika	208	162	202
144	Frontiers in Neurobotics	21	52	21
145	2009 IEEE/ RSJ International Conference on Intelligent Robots Systems, IROS 2009	950	1183	948
146	Annals of Mathematics and Artificial Intelligence	145	107	127
147	AI and Society	174	132	150
148	Journal of Intelligent and Fuzzy Systems	75	86	69
149	Bulletin of the Polish Academy of Sciences. Technical Sciences	193	268	190
150	Journal of Consciousness Studies	216	177	195
151	Expert Systems	105	90	84
152	International Journal of Artificial Intelligence	77	92	73
153	Artificial Intelligence and Law	42	69	42
154	IEEE/ RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings	992	1059	990
155	6th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009	98	108	96
156	Connection Science	57	55	52
157	International Journal of Machine Consciousness	65	60	60
158	Neural Computing and Applications	343	433	332
159	International Journal of Humanoid Robotics	102	90	96
160	Neural Network World	154	99	146
161	Computational Linguistics	84	219	80
162	Minds and Machines	108	79	99
163	Natural Language Engineering	64	105	61
164	Kybernetes	380	199	380
165	Applied Artificial Intelligence	131	124	126
166	International Journal of Advanced Robotics Systems	143	167	142
167	Kongzhi yu Juece/ Control and Decisionn	1137	716	1137
168	Journal of Computer Science	639	481	639
169	Jiqiren/ Robot	350	207	350
170	Artificial Life and Robotics	352	119	349
171	International Journal on Artificial Intelligence Tools	143	84	133
172	Web Intelligence and Agent Systems	72	52	72
173	WSEAS Transactions on Systems and Control	180	101	180

174	International Journal of Robotics and Automation	116	82	116
175	International Journal of Software Engineering and Knowledge Engineering	145	88	136
176	Microprocessors and Microsystems	144	143	137
177	IEEE International Conference on Fuzzy Systems	825	391	820
178	Studies in Computational Intelligence	3140	1064	2995
179	2009 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2009	81	30	79
180	International Journal of Intelligent Mechatronics and Robotics	20	11	20
181	Springer Tracts in Advanced Robotics	287	93	253
182	Informatica	159	106	150
183	Journal of Intelligent Systems	56	18	54
184	Transactions on Interactive Intelligent Systems	6	11	5
185	International Journal of Cognitive Informatics and Natural Intelligence	81	55	70
186	Journal of Software	640	341	622
187	Journal of Multimedia	192	103	181
188	Intelligent Automation and Soft Computing	253	71	239
189	Frontiers in Artificial Intelligence and Applications	1452	249	1410
190	Sistemi Intelligenti	39	8	35
191	Machine Translation	44	25	41
192	Inteligencia Artificial	43	20	39
193	Understanding Complex Systems	252	27	39
194	Transactions of the Japanese Society for Artificial Intelligence	177	30	177
195	International Journal of Mobile Network Design and Innovation	26	7	26
196	Conference on Design and Architectures for Sifnal and Image Processing DASIP	47	4	45
197	WSEAS Transactions on Signal Processing	72	43	71
198	Cognitive Technologies	43	1	32