



TESIS - TE2599

**KOMPARASI GABUNGAN ALGORITMA  
AVERAGE LINKAGE DAN K-MEANS DENGAN K-  
MEANS CLUSTERING UNTUK ANALISA  
FAKTOR PENGANGGURAN DAN ANGKATAN  
KERJA**

Endik Kuswantoro

NRP : 2214206706

Dosen Pembimbing

Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.

Dr. Ir. Wirawan, DEA

PROGRAM MAGISTER

BIDANG KEAHLIAN TELEMATIKA CIO

JURUSAN TEKNIK ELEKTRO

FAKULTAS TEKNOLOGI INDUSTRI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2016



**TESIS - TE2599**

# Comparison Of Algorithms Average Linkage and K-Means Clustering with K-Means Clustering to Analyze the Unemployment Factor on Work Force

Endik Kuswanto

NRP : 2214206706

SUPERVISORS

Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.

Dr. Ir. Wirawan, DEA

PROGRAM MAGISTER

KONSENTRASI CHIEF INFORMATION OFFICER

JURUSAN TEKNIK ELEKTRO

FAKULTAS TEKNOLOGI INDUSTRI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2016

**Tesis ini disusun untuk memenuhi salah satu syarat memperoleh gelar**

**Magister Teknik (MT)**

**di**

**Institut Teknologi Sepuluh Nopember**

**Oleh :**

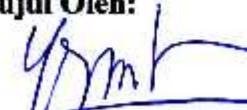
**Endik Kuswantoro**

**Nrp. 2214 206 706**

**Tanggal Ujian : 23 Juni 2016**

**Periode Wisuda : September**

**Disetujui Oleh:**

  
1. Dr. Ir. Yoyon K Suprpto, M.Sc  
NIP. 195409251978031001

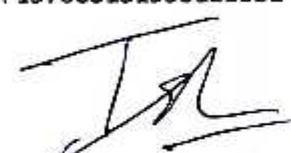
(Pembimbing I)

  
2. Dr. Ir. Wirawan, DEA  
NIP. 196311091989031001

(Pembimbing II)

  
3. Dr. Supeno Mardhi S.N., ST., MT  
NIP. 197003131995121001

(Penguji I)

  
4. Dr. Istas Pratomo, ST., MT  
NIP. 197903252003121001

(Penguji II)

  
Direktur Program Pasca Sarjana,  
  
Prof. Ir. Djuhar Manjari, M.Sc, Ph.d  
NIP. 196012021987011001  
PASCASARJANA

# KOMPARASI GABUNGAN ALGORITMA AVERAGE LINKAGE DAN K-MEANS DENGAN K-MEANS CLUSTERING UNTUK ANALISA FAKTOR PENGANGGURAN DAN ANGKATAN KERJA

Nama : Endik Kuswantoro  
NRP : 2214206706  
Pembimbing : Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.  
Dr. Ir. Wirawan, DEA

## ABSTRAK

Masalah pengangguran selalu jadi masalah bagi setiap daerah, di Jawa Timur tingkat pengangguran terbuka menurut data BPS selalau 2 Tahun terakhir mengalami kenaikan, hal itu disinyalir kebijakan pemerintah yang tidak tepat sasaran, dan faktor pengangguran yang bersifat dinamis setiap waktu. Pada penelitian ini akan menganalisa dengan mengelompokkan faktor pengangguran dengan metode clustering, sehingga dalam analisa faktor pengangguran diperlukan pemilihan metoda yang tepat dan efektif. Algoritma k-means merupakan metode clustering yang mudah diimplementasikan akan tetapi Algoritma K-Means mempunyai kekurangan pada waktu membangkitkan nilai centroid awal yang menyebabkan hasil dari pengelompokan terkadang baik terkadang hasilnya jelek. Pada penelitian ini penentuan centroid awal K-Means dengan menggunakan Average Linkage clustering sebagai salah satu alternatifnya. Dari hasil uji coba perbandingan dari dataset faktor pengangguran nilai Silhoutte index nilai validasi Average Linkage dan K-Means mengalami peningkatan nilai validasi 9,01 % sampai 27,4 % dari nilai K-Means murni.

Kata kunci : k-means, Average Linkage, Penentuan centroid awal, Faktor pengangguran, Silhoutte index

# **COMPARISON ALGORITHM COMBINED AVERAGE LINKAGE AND K-MEANS WITH K-MEANS CLUSTERING FOR FACTORS ANALYSIS OF UNEMPLOYMENT AND WORK FORCE**

Nama : Endik Kuswantoro  
NRP : 2214206706  
Pembimbing : Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.  
Dr. Ir. Wirawan, DEA

## **ABSTRACT**

The problem of unemployment is always a problem for every region, in East Java, the open unemployment rate according to BPS data over the last 2 years has increased, it pointed out the government's policy which is not effective, and the factors of unemployment that is dynamic every time, resulting in a factor analysis of unemployment is required the selection of appropriate methods and effective. K-means algorithm is a clustering method that is easy to implement, but the K-Means algorithm has shortcomings at the time of initial centroid generate a value which causes the result of the grouping is sometimes good sometimes bad result. In this study, the determination of the initial centroid K-Means is clustering using Average Linkage as one alternative. From the test results of a comparative dataset unemployment factor Silhouette index value Average value validation Linkage and K-Means increases in value validation 9.01% to 27.4% of the value of the K-Means pure

Keyword: k-means, Average Linkage, initial centroid, Factor of unemployment, Silhoutte index

## DAFTAR ISI

HALAMAN JUDUL.....	i
PERNYATAAN KEASLIAN TESIS.....	ii
LEMBAR PENGESAHAN .....	iii
ABSTRAK.....	iv
ABSTRACT.....	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	viii
DAFTAR GAMBAR.....	xi
DAFTAR TABEL.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Tujuan Penelitian.....	3
1.3 Manfaat Penelitian.....	3
1.4 Rumusan Masalah.....	3
1.5 Batasan Masalah.....	4
BAB II KAJIAN PUSTAKA DAN LANDASAN TEORI.....	5
2.1. Angkatan Kerja Dan Tenaga Kerja.....	5
2.2 Karakteristik Pengangguran.....	10
2.3 Data Mining (Penambangan Data) .....	12
2.3.1 Proses Tahapan Data Mining.....	13
2.4 Metode Clustering.....	15
2.4.1 Pengukuran Jarak Kedekatan.....	16
2.4.2 K-Means Clustering.....	16
2.4.3 Hierarchical Clustering.....	19

2.5	Evaluasi Validasi Clustering.....	23
2.5.1	Metode Validasi Silhouette Index.....	24
2.6	Penelitian Terdahulu.....	26
<b>BAB III METODOLOGI PENELITIAN.....</b>		<b>27</b>
3.1	Preprosesing Data.....	28
3.1.1	Pemilihan data.....	28
3.1.2	Normalisasi Data.....	30
3.1.3	Pembersihan data (Data Cleaning) .....	30
3.2	Pengklasteran Data.....	31
3.2.1	Algoritma Average Linkage dan K-means Clustering.....	32
3.2.2	Perhitungan manual dari data set sample.....	34
3.3	Evaluasi dan Validasi Cluster.....	47
3.3.1	Penentuan Jumlah cluster yang optimal (Ideal) .....	47
3.4	Interpretasi Data.....	48
3.5	Kesimpulan.....	48
<b>BAB IV HASIL DAN PEMBAHASAN.....</b>		<b>49</b>
4.1	Pemilihan data.....	49
4.1.1	Fitur utama untuk Karakteristik Tingkat Pengangguran.....	51
4.1.2	Fitur Untuk Karakteristik Pengangguran Berdasarkan Wilayah.....	52
4.1.3	Fitur Pengangguran Berdasarkan Pendidikan Terakhir.....	52
4.1.4	Fitur Pengangguran Berdasarkan Umur.....	53
4.1.5	Fitur pengangguran Berdasarkan jenis kelamin.....	53
4.1.6	Fitur Pengangguran Berdasarkan Kompetensi Pendidikan.....	53
4.2	Normalisasi Dan Tranformasi Data.....	55
4.3	Pembersihan Data / Data Cleaning.....	56
4.4	Proses Clustering Data.....	57

4.4.1 Implementasi metode Average Linkage dengan Matlab.....	57
4.4.2 Implementasi K-Means clustering dengan menggunakan Matlab.....	59
4.5 Evaluasi Dan Validasi Average Linkage dan K-Means clustering.....	60
4.5.1 Uji Coba Validasi Cluster Tingkat Pengangguran.....	61
4.5.2 Uji Coba Validasi Cluster Pengangguran dengan wilayah.....	65
4.5.3 Uji Coba Validasi Cluster Pengangguran dengan Pendidikan Terakhir .....	68
4.5.4 Uji Validasi Cluster Pengangguran Dengan Jenis Kelamin.....	71
4.5.5 Pengklasteran Karakteristik Pengangguran Dengan Umur.....	75
4.5.6 Pengklasteran Tingkat Pengangguran Dengan Kompetensi Keahlian .....	77
BAB V KESIMPULAN DAN SARAN.....	81
SARAN.....	82
DAFTAR PUSTAKA.....	83

## DAFTAR TABEL

Tabel 3. 1 Daftar Nama Cluster Pada Masing – Masing Data .....	34
Tabel 3. 2 Matrik Perhitungan Jarak Euclidean .....	38
Tabel 3. 3 Matrik gabungan cluster C4 dan C8654.....	39
Tabel 3. 4 Matrik Gabungan C0/C3 .....	40
Tabel 3. 5 Matrik Gabungan C0/C4/C1 dengan C14514 .....	41
Tabel 3. 6 Matrik Gabungan C0/C3/C1/C2 dengan C8654/C4.....	41
Tabel 3. 7 Matrik Gabungan C0/C3/C1/C2/C8654/C4 .....	41
Tabel 3. 8 Anggota cluster yang dibentuk.....	43
Tabel 3. 9 Centroid awal masing – masing cluster.....	43
Tabel 3. 10 Hasil perhitungan jarak cluster dengan centroidnya .....	45
Tabel 3. 11 Anggota cluster baru yang dibentuk.....	46
Tabel 3. 12 Hasil centroid baru masing – masing cluster Iterasi 1.....	46
Tabel 3. 13 Hasil akhir proses clustering .....	46
Tabel 4. 1 Tabel Pemilihan Fitur untuk Karakteristik Tingkat Pengangguran.....	51
Tabel 4. 2 Tabel Fitur Untuk Karakteristik Pengangguran Berdasarkan Wilayah	52
Tabel 4. 3 Tabel Pemilihan Fitur Pengangguran Berdasarkan Pendidikan .....	52
Tabel 4. 4 Tabel Pemilihan Fitur Untuk Karakteristik Pengangguran Berdasarkan Umur.....	53
Tabel 4. 5 Pemilihan Fitur Untuk Karakteristik Penganggur Berdasarkan jenis kelamin .....	53
Tabel 4. 6 Tabel Fitur Untuk Karakteristik Penganggur Berdasarkan Kompetensi Pendidikan .....	54
Tabel 4. 7 Proses Pembersihan Data .....	56
Tabel 4. 8 Hasil Uji Validasi Silhoutte Index dataset karakteristik Pengangguran .....	62
Tabel 4. 9 Perbandingan waktu Komputasi dengan dataset karakteristik Pengangguran .....	63
Tabel 4. 10 Tabel Analisa Hasil Pengelompokan Karakteristik Pengangguran ....	64

Tabel 4. 11 Hasil Uji Validasi Silhoutte Index dataset Pengangguran dengan wilayah .....	65
Tabel 4. 12 Nilai Standar Deviasi Silhoutte Index data set wilayah .....	66
Tabel 4. 13 Perbandingan waktu komputasi data set wilayah .....	66
Tabel 4. 14 Tabel Analisa Pengelompokan Pengangguran berdasarkan wilayah .	68
Tabel 4. 15 Hasil Uji Validasi Silhoutte Index Pengangguran dengan pendidikan terakhir .....	68
Tabel 4. 16 Nilai Standar Deviasi Silhoutte Index Data Set Pendidikan Terakhir	69
Tabel 4.17 Perbandingan waktu komputasi data set pendidikan terakhir .....	69
Tabel 4. 18 Tabel Analisa Hasil Pengelompokan Pengangguran berdasarkan Pendidikan.....	71
Tabel 4. 19 Hasil Uji Validasi Silhoutte Index Pengangguran dengan Jenis Kelamin.....	72
Tabel 4. 20 Nilai Standar Deviasi Silhoutte Index Data Set Jenis Kelamin .....	72
Tabel 4. 21 Perbandingan waktu komputasi dengan data set jenis kelamin .....	72
Tabel 4. 22 Tabel Analisa Pengelompokan Pengangguran berdasarkan Jenis Kelamin.....	74
Tabel 4. 23 Hasil Uji Validasi Silhoutte Index dataset Pengangguran dengan umur .....	75
Tabel 4. 24 Nilai Standar Deviasi dari Silhoutte Index dengan data set umur .....	75
Tabel 4. 25 Perbandingan waktu komputasi dengan data set umur .....	76
Tabel 4. 26 Tabel Analisa Hasil Pengelompokan Pengguran berdasarkan umur .	76
Tabel 4. 27 Hasil Uji Validasi Silhoutte Index dataset karakteristik Pengangguran dengan Kompetensi Keahlian .....	77
Tabel 4. 28 Nilai Standar Deviasi dari Silhoutte Index Data Set Kompetensi Keahlian .....	78
Tabel 4. 29 Tabel Analisa Hasil Pengelompokan Pengguran berdasarkan umur .	79

## DAFTAR GAMBAR

Gambar 1. 1 Tingkat Pengangguran Terbuka Propinsi Jawa Timur.....	1
Gambar 2. 1 Diagram Penduduk dan Tenaga Kerja.....	6
Gambar 2. 2 Bagan Bidang Ilmu Data Mining.....	13
Gambar 2. 3 Bagan tahapan Data Mining .....	14
Gambar 2. 4 Dendrogram penggerombolan berhierarki dengan prosedur .....	20
Gambar 2.5 Metode Penggabungan Single linkage.....	21
Gambar 2. 6 Metode Penggabungan Complete linkage .....	22
Gambar 2. 7 Metode Penggabungan Average linkage .....	22
Gambar 2. 8 Dendrogram Hierarchical Clustering .....	23
Gambar 3. 1 Diagram Penelitian.....	27
Gambar 3. 2 Flowchart algoritma Average Linkage K-means clustering.....	32
Gambar 3. 3 Dendrogram Average Linkage.....	42
Gambar 3. 4 Flowchart Mencari Jumlah Cluster Yang Optimal.....	47
Gambar 4. 1 Print screen dataset dari Data Sakernas sebelum dilakukan Normalisasi dan Pembedaan data.....	55
Gambar 4. 2 Dataset dalam bentuk format teks *.txt .....	57
Gambar 4. 3 Perintah load dataset pada Matlab .....	58
Gambar 4. 4 Perintah Average Linkage Clustering dengan menggunakan Matlab .....	58
Gambar 4. 5 Grafik Dendrogram average linkage.....	59
Gambar 4. 6 Implementasi K-means dan penentuan centroid awal K-Means .....	60
Gambar 4. 7 Implementasi Uji Validasi dengan metode Silhoutte Index dengan menggunakan Matlab .....	61
Gambar 4. 8 Hasil Uji Coba validasi Silhoutte index attribute karakteristik .....	63
Gambar 4. 9 Diagram Anggota Cluster Karakteristik Pengangguran .....	64
Gambar 4. 10 Diagram Jumlah Anggota Cluster Berdasarkan Wilayah.....	67
Gambar 4. 11 Hasil Uji Coba Silhoutte index Attribut Karakteristik Pengangguran dengan wilayah.....	67
Gambar 4. 12 Attribut Karakteristik Pengangguran dengan pendidikan Terakhir	70

Gambar 4. 13 Diagram Jumlah Anggota pada Karakteristik pengangguran berdasarkan pendidikan terakhir.....	70
Gambar 4. 14 Hasil Uji Coba Silhouette index karakteristik pengangguran dengan Jenis Kelamin .....	73
Gambar 4. 15 Diagram Jumlah Anggota pada Karakteristik pengangguran berdasarkan Jenis Kelamin .....	73
Gambar 4. 16 Hasil Uji Coba Silhouette index karakteristik pengangguran berdasarkan Umur.....	76
Gambar 4. 17 Hasil Uji Coba Silhouette index karakteristik pengangguran berdasarkan Kompetensi Keahlian .....	79

## KATA PENGANTAR

Alhamdulillah ,Segala puji hanya milik Allah SWT. Shalawat dan salam selalu tercurah kepada Rasulullah SAW. Berkat limpahan dan rahmat-Nya penyusun mampu menyelesaikan tesis dengan judul “ KOMPARASI GABUNGAN ALGORITMA AVERAGE LINKAGE DAN K-MEANS DENGAN ALGORITMA K-MEANS CLUSTERING UNTUK ANALISA FAKTOR PENGANGGURAN DAN ANGKATAN KERJA”.

Tesis ini disusun untuk memenuhi salah satu persyaratan memperoleh gelar Magister Teknik (M.T.) dalam bidang keahlian Telematika (CIO) pada program S-2 Jurusan Teknik Elektro di Institut Teknologi Sepuluh Nopember Surabaya. Dalam penyusunan tesis ini, tidak sedikit hambatan yang penulis hadapi. Namun penulis menyadari bahwa kelancaran penulisan tesis ini tidak lain berkat bantuan, dorongan, dan bimbingan dari semua pihak.

Untuk itu kami tidak lupa menyampaikan banyak terima kasih kepada semua pihak yang membantu kami dalam menyelesaikan tesis ini, yaitu kepada :

1. Bapak Dr. Ir. Yoyon K Suprpto, M.Sc. dan Dr. Ir. Wirawan, DEA atas bimbingan, arahan dan waktu yang telah diluangkan kepada penulis untuk berdiskusi selama menjadi dosen pembimbing dan perkuliahan, sehingga tesis ini dapat terselesaikan dengan baik.
2. Bapak Dr. Supeno Mardi Susiki Nugroho, ST., MT dan Bapak Dr. Istas Pratomo, ST., MT selaku penguji dalam penulisan tesis ini yang senantiasa memberikan saran perbaikan dan masukan demi penyelesaian tesis ini.
3. Dr. Ir. Djoko Purwanto, M.Eng. selaku Ketua Jurusan Teknik Elektro, yang memberikan ruang kepada penulis untuk mengembangkan diri dan berkarya.
4. Prof. Ir. Djauhar Manfaat, M.Sc, Ph.D. selaku Direktur Pasca Sarjana, yang memberikan kesempatan kepada penulis untuk mengembangkan diri dan berkarya di Institut Teknologi Sepuluh Nopember.
5. Semua Dosen-dosen Telematika-CIO ITS yang telah memberikan ilmu yang bermanfaat bagi penulis, mudah-mudahan dianggap sebagai amal jariyah oleh Allah SWT.

6. Kementrian Komunikasi Dan Informatika yang telah memberikan beasiswa pendidikan S-2 Telematika CIO, semoga apa yang kami dapatkan bermanfaat dan barokah.
7. Badan Pusat Statistik Propinsi Jawa Timur yang telah memberikan Raw data dan arahan yang berhubungan dengan tesis penulis.
8. Ayahanda Suyitno dan Ibunda Hj.Kustinah, Bapak mertua Murlan dan Ibu mertua Sutiyanah atas semua dukungan, bantuan dan doa yang tidak pernah putus selama penulis menempuh pendidikan S-2.
9. Kakak tercinta Susi Indarwati dan Rochmad yang telah memberikan dukungan secara moril maupun materiil selama penulis berada di Surabaya untuk menempuh pendidikan S-2.
10. Istriku tercinta Beti astuti atas dukungan dan doanya yang telah sabar menunggu selama penulis menempuh pendidikan S-2 di Surabaya dan anakku Al Ghazali Shidqi Prabaswara yang selalu merindukan ayahnya untuk segera pulang yang memberikan semangat kepada penulis untuk segera menyelesaikan pendidikan S-2, mudah – mudahan menjadi anak yang sholeh.
11. Rekan-rekan S2 Telematika-CIO dan S2 Telematika seangkatan maupun beda angkatan yang selalu memberikan keceriaan, motivasi maupun bantuan lainnya. Terima kasih, kalian adalah teman yang terbaik.
12. Rekan-rekan SMKN 1 Trenggalek atas semua dukungan dan doanya, sehingga penulis dapat menyelesaikan studi S-2.
13. Kepada semua pihak yang telah membantu, mendoakan, memberikan motivasi dan dorongan yang tidak dapat saya sebutkan satu persatu.

Penulis menyadari bahwa penulisan tesis ini masih sangat jauh dari sempurna. Oleh sebab itu, penulis sangat mengharapkan kritik dan saran agar tesis ini menjadi lebih baik. Akhir kata, penulis berharap tesis ini memberikan manfaat bagi kita semua terutama untuk pengembangan ilmu pengetahuan.

Surabaya, Juni 2016

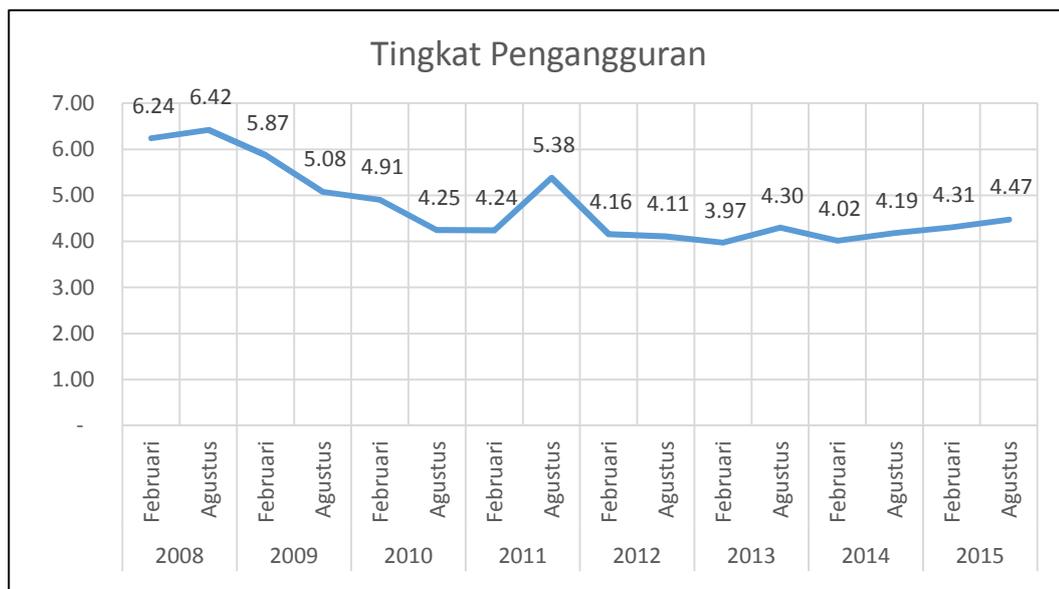
Endik Kuswantoro

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Jumlah angkatan kerja di Indonesia terus meningkat seiring dengan pertambahan jumlah penduduk. Semakin besar jumlah penduduk maka angkatan kerja jadi semakin besar. Hal itu dapat menjadi beban tersendiri bagi perekonomian. Karena jika meningkatnya angkatan kerja yang tidak diimbangi dengan bertambahnya lapangan kerja akan menyebabkan masalah pengangguran. Orang-orang yang menganggur ini secara otomatis tidak akan memperoleh penghasilan. Akibatnya untuk memenuhi kebutuhan pun mereka tidak bisa. Kondisi tersebut dapat menyebabkan kesejahteraannya menurun. Oleh karena itu permasalahan pengangguran juga tidak terlepas dari bagian jumlah angkatan kerja, semakin besar jumlah angkatan kerja maka semakin besar pula jumlah penganggurannya. Semakin bertambahnya jumlah penduduk yang tidak diiringi dengan jumlah lapangan pekerjaan yang memadai menyebabkan jumlah pengangguran yang tinggi, pada propinsi Jawa Timur tingkat pengangguran mulai meningkat pada tahun 2014 bulan februari ,ditunjukkan pada Gambar 1.1



Gambar 1. 1 Tingkat Pengangguran Terbuka Propinsi Jawa Timur

Angka pengangguran yang pada awal mengalami penurunan pada tahun 2013 pada bulan Agustus, mengalami kenaikan kembali dari tahun 2014 sampai tahun 2015, hal itu disinyalir karena beberapa faktor yang mempengaruhi seperti lapangan kerja yang ada tidak mampu menampung (mempekerjakan) tenaga kerja yang ada, lebih-lebih tenaga kerja yang tidak terampil atau berpendidikan. Masalah ini akan menyebabkan semakin meningkatnya tingkat pengangguran sehingga jumlah penduduk miskin juga semakin besar dan memiliki efek-efek negatif yang lain pula. Masalah ketenagakerjaan di Indonesia sekarang ini sudah mencapai kondisi yang cukup memprihatinkan ditandai dengan jumlah penganggur dan setengah penganggur yang besar, pendapatan yang relatif rendah dan kurang merata. Sebaliknya pengangguran dan setengah pengangguran yang tinggi merupakan pemborosan pemborosan sumber daya dan potensi yang ada, menjadi beban keluarga dan masyarakat, sumber utama kemiskinan, dapat mendorong peningkatan keresahan sosial dan kriminal, dan dapat menghambat pembangunan dalam jangka panjang. Permasalahan dalam ketenagakerjaan juga disebabkan karena kekurangan kebutuhan kompetensi dan keahlian yang dibutuhkan oleh pengguna tenaga kerja, hal itu disebabkan oleh tidak meratanya kebutuhan kompetensi tenaga kerja di masing – masing daerah, karena tidak ada pemetaan oleh pemerintah untuk memetakan kebutuhan kompetensi keahlian tenaga kerja yang dibutuhkan oleh pengguna tenaga kerja. Beragam kebijakan pemerintah yang terkait dalam pengangguran selalu mengalami kendala dan sering tidak tepat sasaran karena faktor – faktor penyebab pengangguran di setiap wilayah yang mempunyai karakteristik yang berbeda – beda dan bersifat dinamis yang selalu berubah ubah setiap tahunnya, sehingga perlu untuk dilakukan analisa pada setiap wilayah tertentu. Dalam analisa faktor pengangguran diperlukan pemilihan metoda clustering yang baik dan efektif, metode K-means merupakan metode yang cukup mudah untuk di implementasikan dan mudah untuk diadaptasi akan tetapi memiliki kelemahan yaitu pada saat membangkitkan centroid awal k-means akan menghasilkan cluster yang terkadang baik, terkadang jelek, sehingga memerlukan pengembangan dengan metode lainnya Dalam penelitian ini bertujuan untuk memetakan tingkat pengangguran berdasarkan faktor - faktor pengangguran sehingga dapat memberikan alternatif dalam pengambilan kebijakan oleh

pemerintah, yang terkait dalam mengatasi masalah pengangguran khususnya di Propinsi Jawa Timur.

Dalam penelitian sebelumnya (Febti Eka Pratiwi, Ismaini Zain, 2014) yang berkaitan dengan permasalahan ketenagakerjaan dan pengangguran yaitu menentukan tingkat pengangguran terbuka dan faktor penyebabnya dengan menggunakan metode klasifikasi dengan hasil akurasi 78,9 % sehingga diperlukan penelitian menggunakan metode yang berbeda, faktor – faktor pengangguran terbuka yang mempengaruhi yaitu wilayah, jenis kelamin, pendidikan terakhir, dan usia.

### **1.2 Tujuan Penelitian**

Tujuan dari penelitian ini adalah untuk menentukan jumlah cluster yang ideal dari pengelompokan gabungan dari Average linkage dan K-Means clustering, kemudian membandingkan nilai validasi dari gabungan Average Linkage dan K-Means clustering dengan K-means clustering, dan menganalisa faktor –faktor pengangguran dari hasil pengelompokan dan jumlah cluster yang ideal dari gabungan Average Linkage dan K-Means clustering.

### **1.3 Manfaat Penelitian**

Adapun manfaat yang diperoleh dari hasil penelitian ini adalah menghasilkan kebijakan yang tepat dalam pengambilan keputusan oleh pemerintah Propinsi Jawa Timur yang terkait masalah pengangguran, selain itu memberikan pilihan alternatif bagi peneliti dalam menggunakan algoritma clustering.

### **1.4 Rumusan Masalah**

Rumusan masalah yang mendasari penelitian ini yaitu beragam kebijakan pemerintah yang terkait dalam pengangguran selalu mengalami kendala dan sering tidak tepat sasaran karena faktor – faktor penyebab pengangguran di setiap wilayah yang mempunyai karakteristik yang berbeda – beda dan bersifat dinamis yang selalu berubah ubah setiap tahunnya, sehingga perlu untuk dilakukan analisa pada setiap wilayah tertentu. Dalam analisa faktor pengangguran diperlukan pemilihan metoda clustering yang baik dan efektif, metode K-means merupakan metode yang

cukup mudah untuk di implementasikan dan mudah untuk diadaptasi akan tetapi memiliki kelemahan yaitu pada saat membangkitkan centroid awal k-means akan menghasilkan cluster yang terkadang baik, terkadang jelek, sehingga memerlukan pengembangan dengan metode lainnya.

### **1.5 Batasan Masalah**

Batasan masalah yang ditentukan pada penelitian ini adalah penggunaan sumber data atau data set diambil dari Data SAKERNAS (Survey Angkatan Kerja Nasional) Tahun 2014 yang dilaksanakan secara Triwulanan pada bulan february yang berjumlah 13.266 sample rumah tangga. Fitur yang dipakai dalam penelitian ini sejumlah 11 fitur yang terdiri 6 fitur untuk menentukan jenis pengangguran dan 5 fitur tambahan yang akan dikombinasi Berdasar wilayah di Pedesaan atau Perkotaan, pendidikan terakhir, umur, jenis kelamin, dan kompetensi keahlian. Kemudian melakukan pengklasteran pada faktor pengangguran berdasarkan fitur – fitur yang sudah ditetapkan. Hasil dari clustering dilakukan uji validasi untuk mengukur perbandingan antara metode Average Linkage dan K-Means dengan K-Means clustering, dengan menggunakan nilai validasi Silhoutte index.

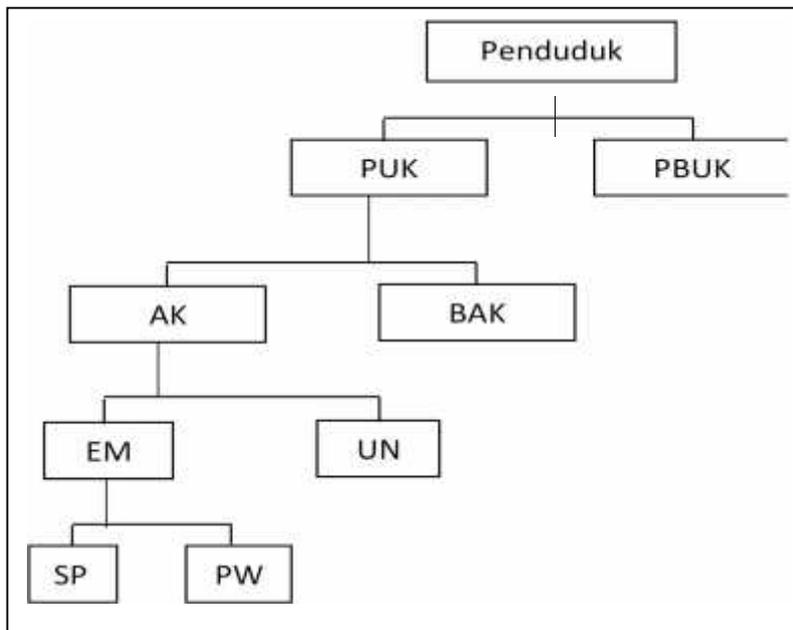
## **BAB II**

### **KAJIAN PUSTAKA DAN LANDASAN TEORI**

Bab ini membahas tentang teori penunjang dalam penelitian yang membahas tentang teori data mining untuk pengelompokan faktor pengangguran dan angkatan kerja, dengan menggunakan gabungan algoritma Average Linkage dan K-Means. Untuk itu perlu diketahui tentang teori angkatan kerja dan ketenagakerjaan yaitu :

#### **2.1. Angkatan Kerja Dan Tenaga Kerja**

Dalam menentukan konsep Angkatan Kerja dan ketenagakerjaan penulis menggunakan konsep dari (BPS,2014) tentang laporan eksekutif Angkatan kerja, sedangkan BPS sendiri menggunakan konsep yang digunakan oleh ILO (International Labor Organization) sebagaimana disajikan pada Gambar 1. Untuk penentuan batas minimum dari usia kerja, pada dasarnya ILO tidak memberikan batasan usia tertentu karena setiap negara memiliki karakteristik dan pertumbuhan ekonomi yang berbeda satu sama lain sehingga memiliki batasan yang berbeda-beda pula untuk menyatakan batas minimum pada usia berapa penduduknya dikatakan telah memasuki usia kerja. Walaupun Undang-Undang Republik Indonesia Nomor 13 Tahun 2003 Tentang Ketenagakerjaan menyatakan bahwa batas usia kerja adalah 18 tahun (karena di bawah usia 18 tahun masih dikategorikan sebagai anak sebagaimana disebutkan pada Pasal 1), namun BPS menggunakan batasan usia kerja 15 tahun agar relevan dengan data-data yang disajikan oleh ILO dan World Bank sehingga data yang dihasilkan bisa diukur keterbandingannya dengan negara lain. Pada Gambar 1. Bagan Ketenagakerjaan menerangkan konsep dari BPS tentang Ketenagakerjaan bahwa penduduk yang termasuk ke dalam usia kerja adalah penduduk berumur 15 tahun dan lebih.



Gambar 2. 1 Diagram Penduduk dan Tenaga Kerja

Keterangan Gambar 1 :

PUK : Penduduk Usia Kerja

PBUK : Penduduk Bukan Usia Kerja

AK : Angkatan Kerja

BAK : Bukan Angkatan Kerja

Em : Employ (Pekerja)

Un : Unemploy (Bukan Pekerja)

SP : Setengah Penganggur

PW : Pekerja Paruh Waktu

Beberapa konsep dari gambar 1. Bagan ketenaga kerjaan diatas yaitu:

1. **Penduduk** dibagi menjadi 2 bagian yaitu :
  - a. Penduduk Angkatan kerja ,Penduduk yang termasuk angkatan kerja adalah penduduk usia kerja (15 tahun dan lebih) yang bekerja, atau punya pekerjaan namun sementara tidak bekerja, dan pengangguran.
  - b. Penduduk yang termasuk bukan angkatan kerja adalah penduduk usia kerja (15 tahun dan lebih) yang masih sekolah, mengurus rumah tangga atau melaksanakan kegiatan lainnya selain kegiatan pribadi.
2. **Bekerja** adalah kegiatan ekonomi yang dilakukan oleh seseorang dengan

maksud memperoleh atau membantu memperoleh pendapatan atau keuntungan, paling sedikit 1 jam (tidak terputus) dalam seminggu yang lalu. Kegiatan termasuk pula kegiatan pekerja tak dibayar yang membantu dalam suatu usaha/kegiatan ekonomi.

- ✓ **Punya pekerjaan tetapi sementara tidak bekerja** adalah keadaan dari seseorang yang mempunyai pekerjaan tetapi selama seminggu yang lalu sementara tidak bekerja karena berbagai sebab, seperti: sakit, cuti, menunggu panen, mogok kerja, dan sebagainya.

Contoh:

- a) Pekerja tetap, pegawai pemerintah/swasta yang sedang tidak bekerja karena cuti, sakit, mogok kerja, mangkir, mesin/peralatan perusahaan mengalami kerusakan, dan sebagainya.
  - b) Petani yang mengusahakan tanah pertanian dan sedang tidak bekerja karena alasan sakit atau menunggu pekerjaan berikutnya (menunggu panen atau musim hujan untuk menggarap sawah).
  - c) Pekerja profesional (mempunyai keahlian tertentu/khusus) yang sedang tidak bekerja karena sakit, menunggu pekerjaan berikutnya /pesanan dan sebagainya, seperti: dalang, tukang cukur, tukang pijat, dukun, penyanyi, komersial, dan sebagainya.
3. **Sekolah** adalah kegiatan seseorang untuk bersekolah di sekolah formal, mulai dari pendidikan dasar sampai dengan pendidikan tinggi selama seminggu yang lalu sebelum pencacahan. Tidak termasuk yang sedang libur sekolah.
  4. **Mengurus rumah tangga** adalah kegiatan seseorang yang mengurus rumah tangga tanpa mendapatkan upah, misalnya: ibu-ibu rumah tangga dan anaknya yang membantu mengurus rumah tangga. Sebaliknya pembantu rumah tangga yang mendapatkan upah walaupun pekerjaannya mengurus rumah tangga dianggap bekerja.
  5. **Kegiatan lainnya** adalah kegiatan seseorang selain diatas, yakni mereka yang sudah pensiun, orang-orang yang cacat jasmani (buta, bisu, dan sebagainya) yang tidak melakukan sesuatu pekerjaan seminggu yang lalu.
  6. **Pendidikan tertinggi** yang ditamatkan adalah tingkat pendidikan yang dicapai

seseorang setelah mengikuti pelajaran pada kelas tertinggi suatu tingkatan sekolah dengan mendapatkan tanda tamat (ijazah).

7. **Jumlah jam kerja seluruh pekerjaan** adalah lamanya waktu dalam jam yang digunakan untuk bekerja dari seluruh pekerjaan, tidak termasuk jam kerja istirahat resmi dan jam kerja yang digunakan untuk hal-hal di luar pekerjaan selama seminggu yang lalu. Bagi pedagang keliling, jumlah jam kerja dihitung mulai berangkat dari rumah sampai tiba kembali dirumah dikurangi waktu yang tidak merupakan jam kerja, seperti mampir ke rumah famili/kawan, dan sebagainya.
8. **Lapangan usaha** adalah bidang kegiatan dari pekerjaan/ usaha/ perusahaan/ kantor tempat seseorang bekerja. Lapangan pekerjaan pada publikasi ini didasarkan pada Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) 2009.
9. **Jenis pekerjaan/jabatan** adalah macam pekerjaan yang dilakukan oleh seseorang atau ditugaskan kepada seseorang yang sedang bekerja atau yang sementara tidak bekerja. Jenis pekerjaan pada publikasi ini, didasarkan atas Klasifikasi Baku Jenis Pekerjaan Indonesia (KBJI) 2002 yang mengacu kepada ISCO 88
10. **Upah/gaji bersih** adalah imbalan yang diterima selama sebulan oleh buruh/karyawan baik berupa uang atau barang yang dibayarkan perusahaan/kantor/majikan. Imbalan dalam bentuk barang dinilai dengan harga setempat. Upah/gaji bersih yang dimaksud tersebut adalah setelah dikurangi dengan potongan-potongan iuran wajib, pajak penghasilan, dan sebagainya.
11. **Status pekerjaan** adalah jenis kedudukan seseorang dalam melakukan pekerjaan disuatu unit usaha/kegiatan. Mulai tahun 2001 status pekerjaan dibedakan menjadi 7 (tujuh) kategori, yaitu:
  1. Berusaha sendiri, adalah bekerja atau berusaha dengan menanggung resiko secara ekonomis, yaitu dengan tidak kembalinya ongkos produksi yang telah dikeluarkan dalam rangka usahanya tersebut, serta tidak menggunakan pekerja dibayar maupun pekerja tak dibayar, termasuk yang sifat pekerjaannya memerlukan teknologi atau keahlian khusus.
  2. Berusaha dibantu buruh tidak tetap/buruh tak dibayar, adalah bekerja atau berusaha atas resiko sendiri, dan menggunakan buruh/pekerja tak dibayar

dan atau buruh/pekerja tidak tetap.

3. Berusaha dibantu buruh tetap/buruh dibayar, adalah berusaha atas resiko sendiri dan mempekerjakan paling sedikit satu orang buruh/pekerja tetap yang dibayar.
4. Buruh/karyawan/pegawai, adalah seseorang yang bekerja pada orang lain atau instansi/kantor/perusahaan secara tetap dengan menerima upah/gaji baik berupa uang maupun barang. Buruh yang tidak mempunyai majikan tetap, tidak digolongkan sebagai buruh/karyawan, tetapi sebagai pekerja bebas. Seseorang dianggap memiliki majikan tetap jika memiliki 1 (satu) majikan (orang/rumah tangga) yang sama dalam sebulan terakhir. Khusus pada sektor bangunan, batasannya adalah tiga bulan. Apabila majikannya adalah instansi/lembaga, boleh lebih dari satu.
5. **Pekerja bebas di pertanian**, adalah seseorang yang bekerja pada orang lain/majikan/institusi yang tidak tetap (lebih dari 1 majikan dalam sebulan terakhir) di usaha pertanian baik berupa usaha rumah tangga maupun bukan usaha rumah tangga atas dasar balas jasa dengan menerima upah atau imbalan baik berupa uang maupun barang, dan baik dengan system pembayaran harian maupun borongan. Usaha pertanian meliputi: pertanian tanaman pangan, perkebunan, kehutanan, peternakan, perikanan dan pemburuan, termasuk juga jasa pertanian. Majikan adalah orang atau pihak yang memberikan pekerjaan dengan pembayaran yang disepakati.
6. **Pekerja bebas di non pertanian** adalah seseorang yang bekerja pada orang lain/majikan/institusi yang tidak tetap (lebih dari 1 majikan dalam sebulan terakhir), di usaha non pertanian dengan menerima upah atau imbalan baik berupa uang maupun barang dan baik dengan sistem pembayaran harian maupun borongan. Usaha non pertanian meliputi: usaha di sektor pertambangan, industri, listrik, gas dan air, sektor konstruksi bangunan, sektor perdagangan, sektor angkutan, perdagangan dan komunikasi, sektor keuangan, asuransi, usaha persewaan bangunan, tanah dan jasa perusahaan, sektor jasa kemasyarakatan, sosial dan perorangan.
7. **Pekerja keluarga/tak dibayar** adalah seseorang yang bekerja membantu orang lain yang berusaha dengan tidak mendapat upah/gaji, baik berupa

uang maupun barang. Pekerja tak dibayar tersebut dapat terdiri dari:

- ✓ Anggota rumah tangga dari orang yang dibantunya, seperti istri/anak yang membantu suaminya/ayahnya bekerja di sawah dan tidak dibayar.
- ✓ Bukan anggota rumah tangga tetapi keluarga dari orang yang dibantunya, seperti famili yang membantu melayani penjualan di warung dan tidak dibayar.
- ✓ Bukan anggota rumah tangga dan bukan keluarga dari orang yang dibantunya, seperti orang yang membantu menganyam topi pada industri rumah tangga tetangganya dan tidak dibayar.

## 2.2 Karakteristik Pengangguran

Dari konsep BPS yang mengacu pada Gambar 1. Kita bisa menentukan karakteristik dari pengangguran dari Bagan ketenagakerjaan diatas, berikut merupakan karakteristik dari pengangguran sebagai berikut :

### 1. **Pengangur terbuka**, terdiri atas:

- a) Mereka yang tak punya pekerjaan dan mencari pekerjaan.
  - b) Mereka yang tak punya pekerjaan dan mempersiapkan usaha.
  - c) Mereka yang tak punya pekerjaan dan tidak mencari pekerjaan, karena merasa tidak mungkin mendapatkan pekerjaan.
  - d) Mereka yang sudah punya pekerjaan, tetapi belum mulai bekerja. (lihat pada “An ILO Manual on Concepts and Methods”)
- Mencari pekerjaan adalah kegiatan seseorang yang pada saat survey orang tersebut sedang mencari pekerjaan, seperti:
- 1) Yang belum pernah bekerja dan sedang berusaha mendapatkan pekerjaan.
  - 2) Yang sudah bekerja, karena sesuatu hal berhenti atau diberhentikan dan sedang berusaha untuk mendapatkan pekerjaan.
  - 3) Yang bekerja atau mempunyai pekerjaan, tetapi karena sesuatu hal masih berusaha untuk mendapatkan pekerjaan lain. Usaha mencari pekerjaan ini tidak terbatas pada seminggu sebelum pencacahan, jadi mereka yang sedang berusaha mendapatkan pekerjaan dan permohonannya telah dikirim lebih dari satu minggu yang lalu tetap

dianggap sebagai mencari pekerjaan asalkan seminggu yang lalu masih mengharapkan pekerjaan yang dicari. Mereka yang sedang bekerja dan berusaha untuk mendapatkan pekerjaan yang lain tidak dapat disebut sebagai **penganggur terbuka**.

- 4) Mempersiapkan suatu usaha adalah suatu kegiatan yang dilakukan seseorang dalam rangka mempersiapkan suatu usaha/pekerjaan yang “baru”, yang bertujuan untuk memperoleh penghasilan/keuntungan atas resiko sendiri, baik dengan atau tanpa mempekerjakan buruh/pekerja dibayar maupun tidak dibayar. Mempersiapkan yang dimaksud adalah apabila “tindakannya nyata”, seperti: mengumpulkan modal atau perlengkapan/alat, mencari lokasi/tempat, mengurus surat izin usaha, dan sebagainya telah/sedang dilakukan. Mempersiapkan usaha tidak termasuk yang baru merencanakan, berniat, dan baru mengikuti kursus/pelatihan dalam rangka membuka usaha. Mempersiapkan suatu usaha yang nantinya cenderung pada pekerjaan sebagai berusaha sendiri (own account worker) atau sebagai berusaha dibantu buruh tidak tetap/buruh tidak dibayar atau sebagai berusaha dibantu buruh tetap/buruh dibayar.

Penjelasan:

Kegiatan mempersiapkan suatu usaha/pekerjaan tidak terbatas dalam jangka waktu seminggu yang lalu saja, tetapi bisa dilakukan beberapa waktu yang lalu asalkan seminggu yang lalu masih berusaha untuk mempersiapkan suatu kegiatan usaha.

2. **Pekerja Tidak Penuh** adalah mereka yang bekerja di bawah jam kerja normal (kurang dari 35 jam seminggu). Pekerja Tidak Penuh terdiri dari:
  - a) **Setengah Penganggur** adalah mereka yang bekerja dibawah jam kerja normal (kurang dari 35 jam seminggu), dan masih mencari pekerjaan atau masih bersedia menerima pekerjaan (dahulu disebut setengah pengangguran terpaksa).
  - b) **Pekerja paruh waktu** adalah mereka yang bekerja dibawah jam kerja normal (kurang dari 35 jam seminggu), tetapi tidak mencari pekerjaan atau

tidak bersedia menerima pekerjaan lain (dahulu disebut setengah pengangguran sukarela).

1.

2.

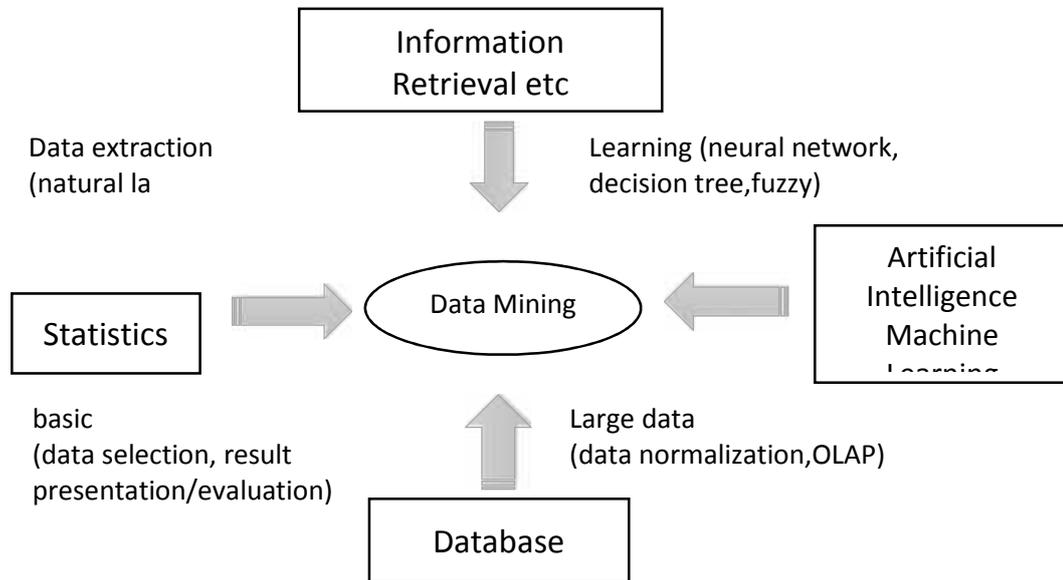
2.1.

### 2.3 Data Mining (Penambangan Data)

Dalam proses pengolahan data pada data Sakernas yang dijadikan dataset yang menjadi bahan penelitian dalam hal ini penulis menggunakan teori Data Mining, Data mining merupakan pemilihan atau penggalian pengetahuan dari jumlah data yang banyak (Han dan Kamber, 2001). Data mining merupakan penemuan pengetahuan atau cara untuk menemukan pola yang tersembunyi pada data. Data mining adalah proses untuk analisis data dari perspektif yang berbeda dan diringkas menjadi informasi yang bermanfaat (Segall et al., 2008). Data mining adalah menganalisis secara otomatis dari data yang berjumlah besar atau kompleks untuk menemukan suatu pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya (Pramudiono, 2006). Data mining adalah analisis meninjau sekumpulan data untuk menemukan suatu hubungan yang tidak diduga dan meringkas data secara berbeda dengan sebelumnya, yang bermanfaat dan dipahami oleh pemilik data (Larose, 2006). Maka dari itu, data mining adalah proses untuk analisis data dalam jumlah besar sehingga membentuk suatu pola yang menjadi informasi berguna. Dari berbagai definisi yang telah disampaikan, berikut merupakan beberapa hal penting yang terkait dengan data mining:

- ✓ Data mining adalah suatu proses otomatis yang dilakukan terhadap data yang telah ada.
- ✓ Data yang akan diproses merupakan data yang berjumlah sangat besar.
- ✓ Tujuan dari data mining adalah untuk mendapatkan hubungan atau pola yang kemungkinan memberikan indikasi bermanfaat. Data mining merupakan suatu bidang ilmu yang telah lama ada. Kesulitan untuk mendefinisikan data mining salah satunya karena data mining mewarisi banyak aspek dan teknik dari berbagai bidang ilmu yang sudah mapan terlebih dahulu. Dari Gambar 2 memperlihatkan

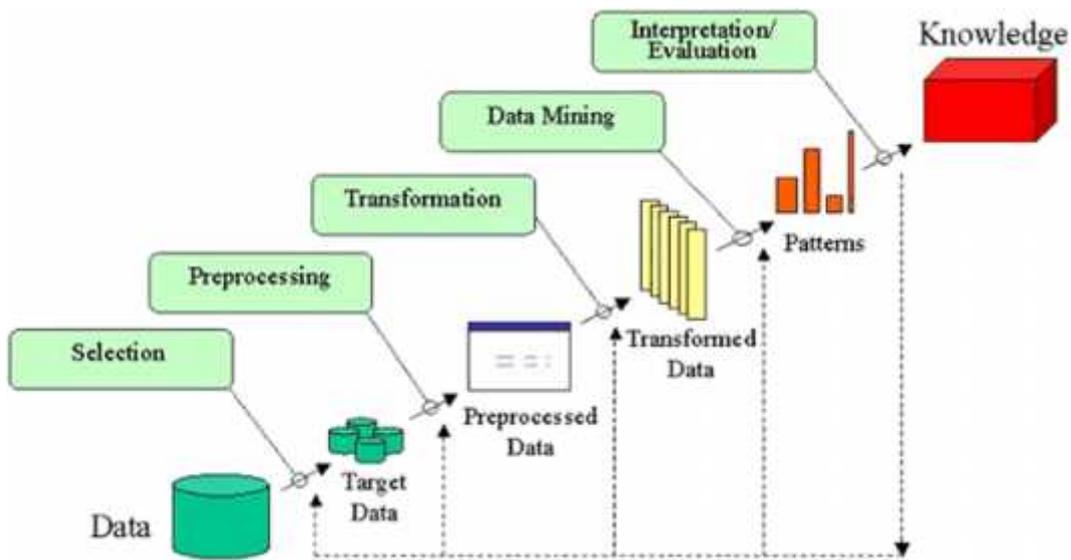
bahwa data mining memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligence), machine learning, statistic, database, serta information retrieval (Pramudiono, 2006).



Gambar 2. 2 Bagan Bidang Ilmu Data Mining

### 2.3.1 Proses Tahapan Data Mining

Dalam melakukan penelitian pengolahan data menggunakan konsep data mining maka harus mengetahui bagaimana proses data mining tersebut, untuk penjelasan proses penggalian informasi yang tersembunyi dalam suatu basis data yang besar atau disebut istilah data mining dan Knowledge Discovery in Database (KDD) sering digunakan bergantian. Pada implementasinya, data mining adalah bagian dari proses KDD. Sebagai komponen dalam proses KDD, data mining berkaitan dengan ekstraksi dan penghitungan pola-pola data yang ditelaah pada basis data. KDD melingkupi keseluruhan proses pencarian pola atau informasi pada basis data yang akan dimulai dengan pemilihan dan persiapan data hingga representasi pola yang akan ditentukan kedalam bentuk yang lebih mudah dimengerti oleh pihak berkepentingan. Secara garis besar proses KDD terdiri atas beberapa tahap (Fayyad, 1996). Berikut merupakan representasi tahapan dalam penelitian menggunakan konsep data mining pada Gambar 3 :



Gambar 2. 3 Bagan tahapan Data Mining

Keterangan dari Gambar 3 :

1. Data Seleksi

Pemilihan (seleksi) data dilakukan dari suatu kumpulan data operasional, sebelum tahap penggalian informasi dalam KDD dimulai proses ini perlu dilakukan. Data hasil seleksi disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pembersihan Data

Proses cleaning perlu dilakukan pada data yang menjadi fokus KDD sebelum proses data mining dapat dilakukan. Proses cleaning melingkupi antara lain membuang data yang memiliki duplikasi, data yang tidak konsisten diperiksa, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (typo), juga dilakukan proses enrichment, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformasi data

Proses transformasi pada data yang telah dipilih adalah Coding, sehingga sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

#### 4. Data Mining

Data mining merupakan proses untuk mencari suatu pola atau informasi yang menarik dalam data yang terpilih dengan teknik atau metode tertentu. Data mining memiliki teknik, metode, atau algoritma dalam sangat bervariasi. Pemilihan metode dan algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

#### 5. Interpretasi Data

Interpretation merupakan proses untuk menampilkan pola informasi yang dihasilkan dari proses data mining oleh pihak yang berkepentingan. Tahap ini meliputi pemeriksaan terhadap pola atau informasi yang ditemukan agar tidak bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

### 2.4 Metode Clustering

Dalam konsep penelitian yang menggunakan data mining Algoritma yang digunakan pada penelitian ini ialah menggunakan Algoritma Clustering, karena data set yang akan diolah belum diketahui label kelasnya. Teknik Clustering adalah suatu teknik analisis multivariate untuk menemukan suatu kelompok-kelompok dari sekumpulan objek maupun individu berdasarkan karakteristik yang mempunyai kemiripan antar anggota kelompok (cluster) yang tinggi dan mempunyai karakteristik antar kelompok (cluster) berbeda yang rendah. Tujuan objektif secara khusus dari teknik ini adalah untuk mengklasifikasikan sampel entitas (berupa individu/objek atau yang lainnya) yang berdasarkan kemiripan antara entitas menjadi sejumlah kelompok tertentu. Untuk cluster banyaknya suatu kelompok tidak ditentukan terlebih dahulu. Karena kelompok (cluster) pada data pada diidentifikasi menggunakan teknik ini. Hasil cluster (kelompok) yang akan terbentuk diharapkan memiliki tingkat homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi pula. Hal pokok dalam analisis cluster adalah ukuran kemiripan antar objek. Metode pengukuran kemiripan pada cluster yang biasa digunakan ada tiga yaitu kemiripan antar objek biasa diukur menggunakan jarak (distance), kemiripan antar variabel (jika variabel yang diclusterkan) diukur menggunakan ukuran korelasi, dan asosiasi antar variabel. Metode utama pada analisis cluster adalah metode hierarki (Average Linkage clustering) dan metode

tidak berhierarki/partisi (K-means clustering). Metode hierarki (Average Linkage clustering) adalah metode yang melibatkan pembangunan berhierarki struktur pohon yang menggunakan prosedur penggabungan (agglomerative) atau pembagian (divisive). Metode tidak berhierarki disebut juga dengan metode penyekatan (partitioning / K-means method) dimana jumlah cluster yang ingin dibentuk sudah ditentukan sebelumnya

#### 2.4.1 Pengukuran Jarak Kedekatan

Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan dengan persamaan (2.1):

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \dots\dots\dots(2.1)$$

dimana :

$d(x_i, x_j)$  = jarak dari  $x_i$  dan  $x_j$

#### 2.4.2 K-Means Clustering

Selain menggunakan konsep non partition penelitian ini menggunakan konsep partision clustering dalam hal ini menggunakan Algoritma K-means clustering, Algoritma K-means clustering merupakan salah satu metode data clustering non-hirarki yang mengelompokan data dalam bentuk satu atau lebih cluster/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokan dalam satu cluster/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan cluster/kelompok yang lain sehingga data yang berada dalam satu cluster/kelompok memiliki tingkat variasi yang kecil (Agusta, 2007).

Menurut Santosa (2007), langkah-langkah melakukan clustering dengan metode K-Means adalah sebagai berikut:

- a) Pilih jumlah cluster k.
- b) Inisialisasi k pusat cluster ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat cluster diberi nilai awal dengan angka-angka random,
- c) Alokasikan semua data/ objek ke cluster terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke cluster tertentu ditentukan jarak antara data dengan pusat cluster. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat cluster. Jarak paling antara kedekatan satu data dengan satu cluster tertentu akan menentukan suatu data masuk dalam cluster mana.
- d) Hitung kembali pusat cluster dengan keanggotaan cluster yang sekarang. Pusat cluster adalah rata-rata dari semua data/ objek dalam cluster tertentu. Jika dikehendaki bisa juga menggunakan median dari cluster tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai.
- e) Tugaskan lagi setiap objek memakai pusat cluster yang baru. Jika pusat cluster tidak berubah lagi maka proses clustering selesai. Atau, kembali ke langkah ( c ) sampai pusat cluster tidak berubah lagi.

Beberapa permasalahan yang sering muncul pada saat menggunakan metode K-Means untuk melakukan pengelompokan data adalah:

1. Ditemukannya beberapa model clustering yang berbeda umumnya disebabkan oleh perbedaan proses inisialisasi anggota masing-masing cluster. Proses inisialisasi yang sering digunakan adalah proses inisialisasi secara random. Dalam suatu studi perbandingan, proses inisialisasi secara random mempunyai kecenderungan untuk memberikan hasil yang lebih baik dan independent, walaupun dari segi kecepatan untuk converge lebih lambat.
2. Pemilihan jumlah cluster yang paling tepat merupakan masalah laten dalam metode K-Means. Beberapa pendekatan telah

digunakan dalam menentukan jumlah cluster yang paling tepat untuk suatu dataset yang dianalisa termasuk di antaranya Partition Entropy (PE) dan GAP Statistics. Satu hal yang patut diperhatikan mengenai metode - metode ini adalah pendekatan yang digunakan dalam mengembangkan metode-metode tersebut tidak sama dengan pendekatan yang digunakan oleh K-Means dalam mempartisi data items ke masing-masing cluster.

### 3. Kegagalan untuk converge

Permasalahan kegagalan untuk converge, secara teori memungkinkan untuk terjadi dalam kedua metode K-Means yang dijelaskan di dalam tulisan ini. Kemungkinan ini akan semakin besar terjadi untuk metode Hard K-Means, karena setiap data di dalam dataset dialokasikan secara tegas (hard) untuk menjadi bagian dari suatu cluster tertentu. Perpindahan suatu data ke suatu cluster tertentu dapat mengubah karakteristik model clustering yang dapat menyebabkan data yang telah dipindahkan tersebut lebih sesuai untuk berada di cluster semula sebelum data tersebut dipindahkan. Demikian juga dengan keadaan sebaliknya. Kejadian seperti ini tentu akan mengakibatkan pemodelan tidak akan berhenti dan kegagalan untuk converge akan terjadi. Untuk Fuzzy K-Means, walaupun ada, kemungkinan permasalahan ini untuk terjadi sangatlah kecil, karena setiap data dilengkapi dengan membership function (Fuzzy K-Means) untuk menjadi anggota cluster yang ditemukan

### 4. Pendeteksian outliers

Permasalahan keempat merupakan permasalahan umum yang terjadi hampir di setiap metode yang melakukan pemodelan terhadap data. Khusus untuk metode K-Means hal ini memang menjadi permasalahan yang cukup menentukan. Beberapa hal yang perlu diperhatikan dalam melakukan pendeteksian outliers dalam proses pengelompokan data termasuk bagaimana menentukan apakah suatu data item merupakan outliers dari suatu cluster tertentu dan apakah data dalam jumlah kecil yang membentuk suatu cluster tersendiri dapat dianggap sebagai outliers. Proses ini memerlukan suatu pendekatan khusus yang berbeda dengan proses pendeteksian outliers di dalam suatu dataset yang hanya terdiri dari satu populasi yang homogen

### 5. Bentuk masing-masing cluster

Tidak seperti metode data clustering lainnya termasuk Mixture Modelling, K-Means umumnya tidak mengindahkan bentuk dari masing-masing cluster yang mendasari model yang terbentuk, walaupun secara natural masing-masing cluster umumnya berbentuk bundar. Untuk dataset yang diperkirakan mempunyai bentuk yang tidak biasa, beberapa pendekatan perlu untuk diterapkan.

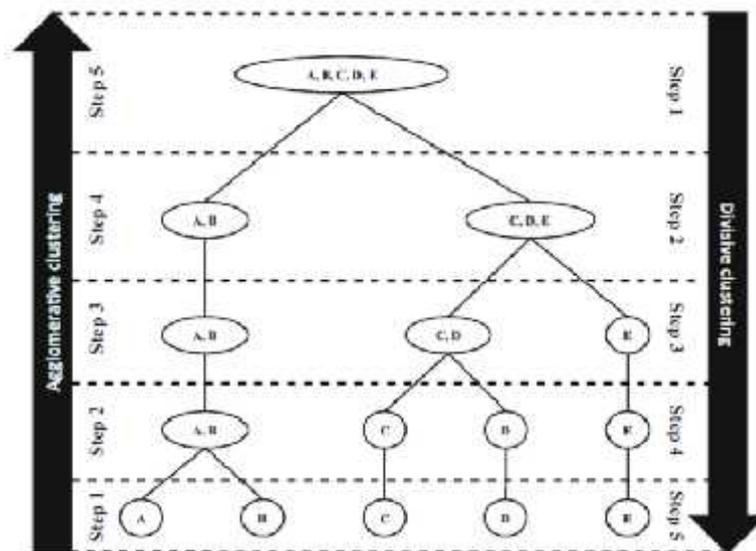
#### 6. Masalah overlapping

Masalah overlapping sebagai permasalahan terakhir sering sekali diabaikan karena umumnya masalah ini sulit terdeteksi. Hal ini terjadi untuk metode Hard K-Means dan Fuzzy K-Means, karena secara teori, metode ini tidak dilengkapi feature untuk mendeteksi apakah di dalam suatu cluster ada cluster lain yang kemungkinan tersembunyi.

### 2.4.3 Hierarchical Clustering

Konsep Clustering yang digunakan dalam penelitian ini menggunakan gabungan dari konsep partition clustering dan non partition clustering. Dalam menggunakan konsep non partition clustering memakai Algoritma Hierarchical clustering, Hierarchical clustering memiliki dua tipe metode yaitu agglomerative dan divisive. Agglomerative clustering atau disebut juga dengan metode "bottomup" karena dianggap setiap objek sebagai cluster tunggal kemudian cluster-cluster tersebut digabungkan sehingga hanya tersisa satu cluster saja. Divisive clustering juga disebut dengan metode "top-down". Pada metode divisive awalnya seluruh objek dianggap menjadi satu kesatuan cluster yang sama, kemudian dilakukan proses pemecahan cluster menjadi dua cluster dan seterusnya hingga setiap objek dianggap satu cluster tunggal. Proses pembentukan cluster pada hierarchical clustering digambarkan melalui diagram dua dimensi yang disebut dendrogram. Gambar 4 merupakan pembentukan cluster baik dengan prosedur agglomerative maupun divisive dalam bentuk dendrogram. Prosedur yang sering digunakan dalam metode clustering berhierarki adalah prosedur agglomerative. Awalnya terdapat  $n$  anggota/observasi yang dianggap  $n$  cluster atau kelompok tunggal dan pada akhirnya menghasilkan satu cluster atau satu kelompok yang berisi  $n$  anggota. Operasi untuk penggabungan dasar yang sering

digunakan adalah diantaranya single linkage, complete linkage, dan average linkage. Ketiga metode tersebut memiliki perbedaan dalam menentukan jarak (similarity) antara objek dengan kelompok objek ataupun kelompok objek dengan kelompok objek. Single linkage merupakan operasi penggabungan yang menggunakan jarak minimum atau tetangga terdekat (nearest neighbor). Sebaliknya, complete linkage menggunakan jarak terjauh (farthest neighbor). Sedangkan average linkage menggunakan rata-rata jarak untuk menentukan jarak antar objek. Metode operasi penggabungan lainnya antara lain Ward's minimal variance, centroid method, median method, dan average linkage wighted.



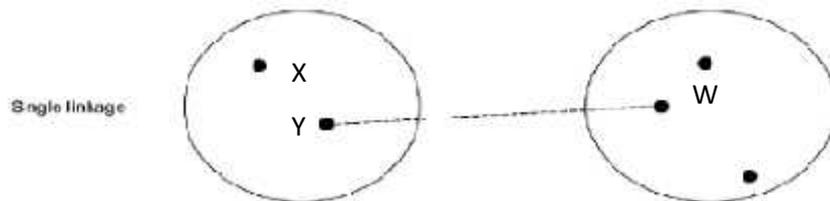
Gambar 2. 4 Dendrogram penggerombolan berhierarki dengan prosedur

Algoritma agglomerative hierarchical clustering dengan banyak objek adalah  $N$  secara umum adalah sebagai berikut (Izenman, 2008):

1. Diawali dengan  $N$  buah cluster yang mana dari setiap cluster merupakan entitas tunggal dengan sebuah matriks jarak (matriks kedekatan) berukuran  $N \times N$  yang dituliskan  $D = [d_{ik}]$ .
2. Menentukan jarak pasangan cluster pada matriks yang jaraknya paling dekat (atau paling mirip) yang mana jarak paling dekat adalah pasangan cluster yang memiliki nilai jarak terkecil. Misalnya jarak paling dekat

terdapat pada pasangan cluster X dan Y, maka jarak tersebut diberi inisial  $d(X,Y)$ .

- Cluster X dan Y digabungkan menjadi suatu cluster baru yaitu cluster XY. Perbaharui ukuran matriks jarak menjadi  $(N - 1) \times (N - 1)$ . Penghitungan jarak antara cluster baru yang dibentuk dengan N-1 cluster yang telah ada, dapat dilakukan dengan berbagai metode penggabungan single linkage, complete linkage, average linkage ataupun yang lainnya. Dari persamaan (2.2) nilai dari jarak ketika cluster XY dan W digabungkan adalah jarak minimum antar cluster X dengan cluster W dan cluster Y dengan cluster W. Dimana pada persamaan (2.2) merupakan bentuk matematis dari metode penggabungan single linkage. Dan persamaan (2.3) adalah bentuk matematis dari metode complete linkage jarak ketika cluster XY dan W digabungkan adalah jarak maksimum antar cluster X dengan cluster W dan cluster Y dengan cluster W. Sedangkan pada metode average linkage, jarak dua cluster merupakan rata-rata jarak dari keduanya, dapat dilihat pada persamaan (2.4).



Gambar 2.5 Metode Penggabungan Single linkage

$$d_{(xy)w} = \min\{d_{(xy)}, d_{(yw)}\}, \dots \dots \dots (2.2)$$

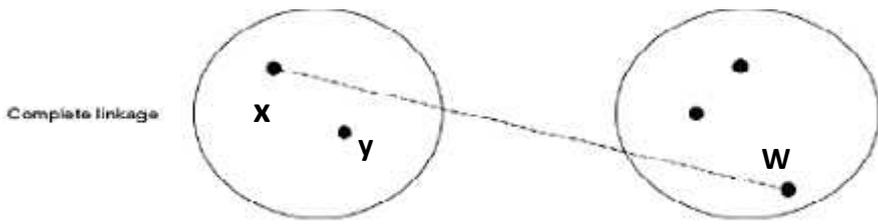
dimana :

$d_{(xy)w}$  = Jarak antar cluster xy dengan cluster w

$d_{(yw)}$  = Jarak antar cluster y dengan cluster w

$d_{(xw)}$  = Jarak antar cluster x dengan cluster w

$\min\{d_{(xw)}, d_{(yw)}\}$  = Jarak minimum antar cluster xw dengan cluster yw



Gambar 2. 6 Metode Penggabungan Complete linkage

$$d_{(xy)w} = \max \{ d_{(xy)}, d_{(yw)} \}, \dots \dots \dots (2.3)$$

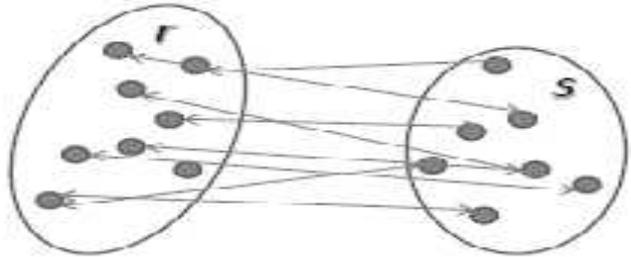
dimana :

$d_{(xy)w}$  = Jarak antar cluster xy dengan cluster w

$d_{(yw)}$  = Jarak antar cluster y dengan cluster w

$d_{(xw)}$  = Jarak antar cluster x dengan cluster w

$\max \{ d_{(xw)}, d_{(yw)} \}$  = Jarak maksimum antar cluster xw dengan cluster yw



Gambar 2. 7 Metode Penggabungan Average linkage

$$L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(X_r, X_s) \dots \dots \dots (2.4)$$

Keterangan :

$L(r,s)$  = Jarak antar cluster r dengan cluster s

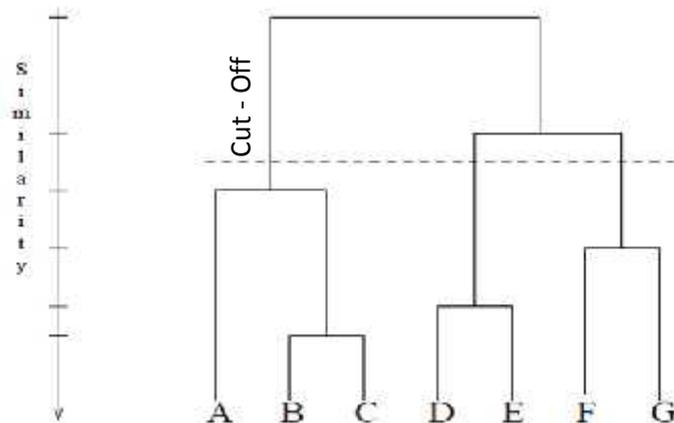
$n_r$  = Jumlah anggota cluster r

$n_s$  = Jumlah anggota cluster s

$D(x_{(ri)}, x_{(sj)})$  = Jarak antar cluster r pada anggota cluster ke i dengan cluster s pada anggota cluster ke j

4. Lakukan kembali langkah 2 dan 3 sebanyak  $N-1$  kali. Kemudian catat setiap identitas cluster yang bergabung dan tingkat jarak untuk setiap penggabungan kemudian gambarkan dalam dendrogram.

Dari dendrogram yang dihasilkan dapat diketahui urutan bagaimana objek dikelompokkan ke dalam suatu cluster. Jumlah cluster yang dapat ditentukan dengan memotong dendrogram pada suatu ketinggian tertentu. Pada gambar tersebut terdapat garis yang menunjukkan dimana terdapat Cut – Off pada dendrogram. Dari kita ingin mengelompokkan objek kedalam dua cluster. Berikut Gambar 8. Dendrogram merupakan gambaran dalam menentukan jumlah cluster.



Gambar 2. 8 Dendrogram Hierarchical Clustering

## 2.5 Evaluasi Validasi Clustering

Kemampuan untuk mendeteksi ada atau tidaknya struktur tidak acak pada data merupakan salah satu aspek penting dalam evaluasi validasi klaster. Berikut ini merupakan beberapa aspek penting dalam validasi klaster (Tan et al. 2006) :

1. Menentukan tujuan pengklasteran dari data.
2. Menentukan jumlah klaster yang tepat.
3. Mengevaluasi seberapa baik hasil analisis klaster tanpa diberikan informasi eksternal.
4. Membandingkan hasil analisis klaster terhadap hasil eksternal yang diketahui, misalnya label kelas eksternal

5. Membandingkan dua himpunan klaster untuk menentukan yang lebih baik. Aspek 1, 2 dan 3 tidak memerlukan informasi eksternal, yang merupakan teknik tak terbimbing/terpandu, sedangkan aspek 4 membutuhkan informasi eksternal dilakukan terhadap teknik terbimbing. Aspek 5 dapat dilakukan terhadap teknik terbimbing atau tak terbimbing. Ukuran - ukuran evaluasi dapat digolongkan menjadi tiga jenis (Tan et al. 2006):

1. Teknik tak terbimbing (Unsupervised): mengukur kebaikan dari struktur pengklasteran tanpa informasi eksternal, salah satu contohnya adalah SSE. Besaran teknik tak terbimbing dibagi dua yaitu pertama kesatuan klaster (kekompakan, keketatan) yang menunjukkan seberapa dekat objek dalam satu klaster dan kedua adalah pemisahan klaster (terisolasi) yang menunjukkan seberapa jauh atau kebaikan pemisahan suatu klaster dengan klaster lain.
2. Teknik terbimbing (Supervised): mengukur kecocokan struktur pengklasteran dengan struktur eksternal, salah satu contohnya adalah entropy.
3. Relative: membandingkan pengklasteran yang berbeda. Besaran evaluasi pengklasteran relatif merupakan teknik terbimbing dan tak terbimbing yang digunakan untuk tujuan perbandingan. Pada Sub bab ini akan dilakukan evaluasi terhadap hasil algoritma metode clustering. Pengujian ini dilakukan untuk melihat apakah kombinasi algoritma Average Linkage clustering dan k-means menghasilkan pengelompokan data yang lebih baik jika dibandingkan dengan k-means itu sendiri. Adapun pengujian yang dilakukan adalah dengan metode Cluster Silhoutte Index.

### 2.5.1 Metode Validasi Silhoutte Index

Untuk mengukur validasi cluster dalam penelitian ini sangatlah penting, karena untuk mengetahui kualitas cluter baik atau artinya jarak antar anggota cluster kecil sedangkan jarak antar cluster besar. Dalam pengukuran validasi cluster penelitian ini menggunakan metode Silhoutte index (SI), metode ini dapat memvalidasi cluster yang menggabungkan kohesi dan separasi. Untuk menghitung nilai SI dari sebuah data ke  $- i$ , ada 2 komponen yaitu  $a_i$  dan  $b_i$ ,  $a_i$  adalah rata - rata

jarak data ke –  $i$  terhadap semua data lainnya dalam satu cluster, sedangkan  $b_i$  didapat dengan menghitung rata – rata jarak data ke –  $i$  terhadap semua data dari cluster yang lain tidak dalam satu cluster dengan data ke –  $i$ , kemudian diambil dari nilai yang terkecil. Nilai *Silhouette Index* bervariasi dari -1 dan 1. Nilai *Silhouette Index* mendekati 1 menunjukkan bahwa data tersebut tepat berada pada kluster tersebut. SI bernilai 0 atau mendekati 0 maka posisi data berada pada perbatasan dua kluster. Nilai negatif menandakan jarak rata-rata antar objek jauh.

Berikut ini merupakan ukuran nilai silhouette menurut Kaufman dan Rousseuw.

Nilai *Silhouette Index* (SI):

- $0.7 < SI \leq 1$  strong structure
- $0.5 < SI \leq 0.7$  medium structure
- $0.25 < SI \leq 0.5$  weak structure
- $SI \leq 0.25$  no structure

Formula untuk menghitung  $a_i^j$  pada persamaan (2.5) :

$$a_i^j = \frac{1}{m_j - 1} \sum_{r=1}^{m_j} d(x_i^j, x_r^j), \quad i = 1, 2, \dots, m_j \quad \dots\dots\dots(2.5)$$

Formula untuk menghitung nilai  $b_i^j$  pada persamaan(2.6) :

$$b_i^j = \min_{\substack{n=1, \dots, k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq i}}^{m_n} d(x_i^j, x_r^n) \right\}, \quad i = 1, 2, \dots, m_n \quad \dots\dots\dots(2.6)$$

Untuk mendapatkan Silhouette Index (SI) data ke –  $i$  menggunakan persamaan (2.7) :

$$S_i^j = \frac{b_i^j - a_i^j}{\max \{a_i^j, b_i^j\}} \quad \dots\dots\dots(2.7)$$

Dimana :

- $a_i$  = Rata-rata jarak dari data ke- $i$  terhadap semua data lainnya dalam satu cluster
- $b_i$  = Rata-rata jarak dari data ke- $i$  terhadap semua data dari cluster lain
- $d(x_i^r, x_r^r)$  = Jarak data ke- $i$  dengan data ke- $r$  dalam satu cluster  $j$
- $m_j$  = Jumlah data dalam cluster ke- $j$
- $S_i^j$  = Nilai Silhouette Index pada data ke- $i$

Untuk mendapatkan nilai SI dari dari sebuah *cluster* dapat dilakukan dengan menghitung nilai rata-rata SI dalam *cluster* tersebut dan dapat dilihat pada persamaan (2.8).

$$S_i^j = \frac{1}{m_j} \sum_{i=1}^{m_j} S_i^j \dots\dots\dots(2.8)$$

Untuk mendapatkan nilai SI Global dilakukan dengan menghitung rata-rata dari keseluruhan nilai dapat dilihat pada persamaan (2.9).

$$S = \frac{1}{k} \sum_{j=1}^k S \dots\dots\dots(2.9)$$

Rata-rata SI dari seluruh data dalam suatu *cluster* menunjukkan seberapa dekat kemiripan data dalam suatu *cluster* yang juga menunjukkan seberapa tepat data telah dikelompokkan. Semakin dekat SI kepada 1, maka semakin baik pengelompokkan data Sebaliknya, semakin dekat SI kepada -1, maka semakin buruk pengelompokkan data

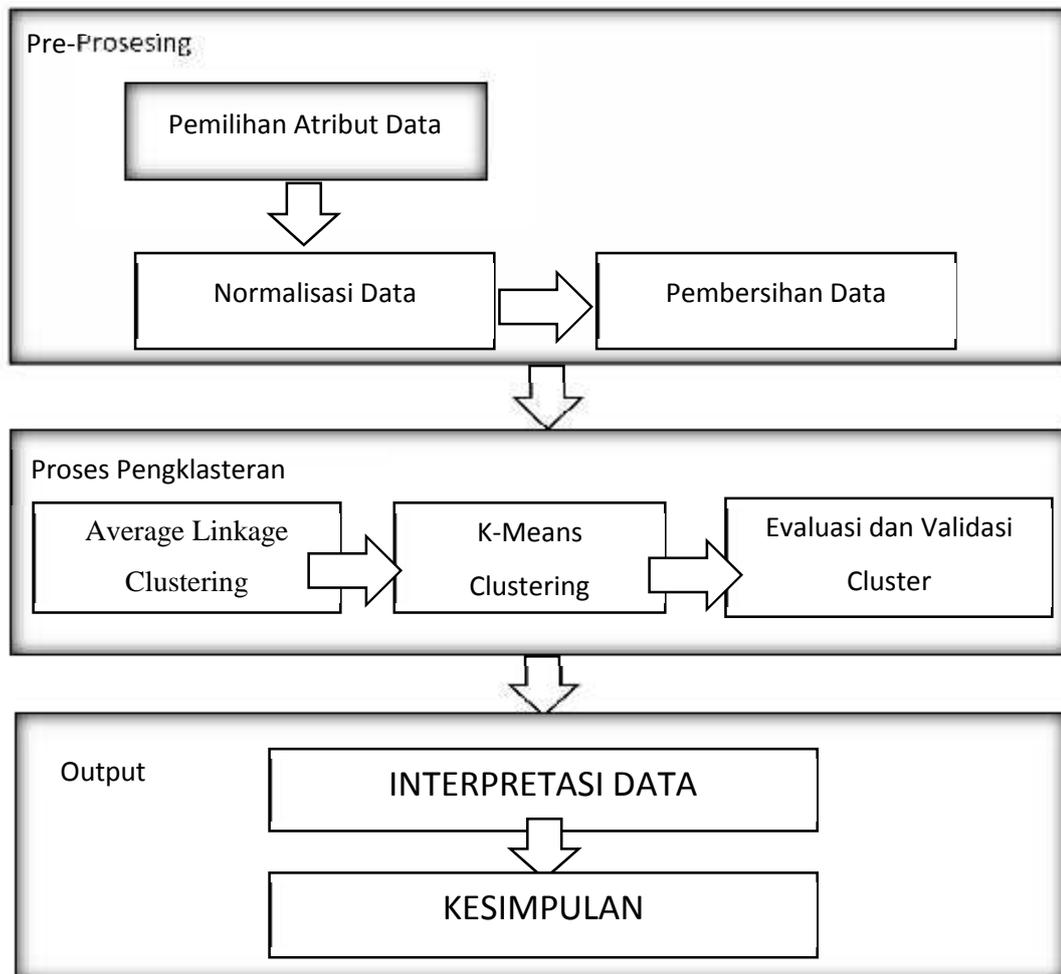
## 2.6 Penelitian Terdahulu

Pada bagian ini akan dijabarkan beberapa penelitian terdahulu. Hierarchical K-Means telah digunakan untuk beberapa penelitian seperti yang dilakukan oleh Widyasari pada penelitian “Analisis Penerapan Metode Single Linkage dan K-Means untuk Pewilayahan Pola Hujan Provinsi Sumatera Selatan Berbasis Arc View”. Penelitian lain dilakukan oleh Tahta Alfina, dkk pada tahun 2012 dengan judul “Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)”. Rendy Handoyo pada tahun 2014 juga melakukan penelitian dengan judul “Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K - Means Pada Pengelompokan Dokumen”. Dan Kohei Arai dan Ali Ridho Barakbah melakukan penelitian dengan judul “Hierarchical K-means: an algorithm for centroids initialization for K-means”.

### BAB III

## METODOLOGI PENELITIAN

Pada bagian ini akan menjelaskan tentang langkah – langkah metodologi penelitian secara sistematis dan terarah yang akan dijadikan acuan dalam kerangka penelitian yang membahas tentang Pengelompokan Angkatan kerja, guna mengukur tingkat pengangguran, faktor – faktor pengangguran dan persebaran jenis pekerjaan utama dengan menggunakan gabungan dari Algoritma Average Linkage clustering dan K-Means clustering. Berikut merupakan diagram penelitian yang disajikan pada gambar 3.1 :



Gambar 3. 1 Diagram Penelitian

### 3.1 Preprocessing Data

Pada tahap ini menjelaskan tentang pemrosesan awal dari pengolahan data, pada tahap pemrosesan awal ini merupakan tahapan yang sangat penting karena kualitas dari pengolahan pada data mining tergantung dari benar tidaknya pada proses awal (Preprocessing). Dalam penelitian biasanya data yang dibutuhkan tidak digunakan semuanya akan tetapi dari data yang didapat akan diambil beberapa atribut data yang memenuhi karakteristik dan tujuan dari penelitian tersebut, dalam hal ini adanya pemilihan fitur atau atribut data. Data yang dipilih kemudian dilakukan pemrosesan dan analisa untuk mencari apakah ada data yang hilang atau salah dalam melakukan pengentrian data, biasanya pada data tertentu selalu ada nilai yang salah atau kosong pada satu atau lebih fitur dari satu atau lebih vektor dalam data keseluruhan.

#### 3.1.1 Pemilihan data

Sumber data yang digunakan dalam penelitian ini berasal dari hasil pengumpulan data Survey Angkatan Kerja Nasional Tahun 2014 yang dilakukan oleh Badan Pusat Statistik Propinsi Jawa Timur pada bulan Februari, survey tersebut pada setiap tahunnya diselenggarakan secara serentak diseluruh kabupaten/kota di Indonesia. Sakernas Februari 2014 ditujukan untuk menghasilkan angka estimasi sampai dengan tingkat kabupaten/kota. Setelah data didapatkan kemudian dilakukan analisa untuk menentukan pemilihan fitur data yang akan diolah, pengolahan data dilakukan dengan menggunakan Algoritma Average Linkage clustering dan K-means clustering total data yang digunakan sejumlah 13.266 data sample. Pada pemilihan fitur tentunya harus sesuai dengan tujuan penelitian yaitu untuk mengukur estimasi tingkat pengangguran pada propinsi Jawa Timur, dari studi literatur mengenai konsep pengangguran maka 12 fitur yang digunakan dari data yang diambil dari poin – poin Survey yang menunjukkan karakteristik pengangguran yaitu :

1. Status Bekerja dalam 1 minggu : Responden dalam waktu kurun 1 minggu melakukan pekerjaan sekurang kurangnya 1 jam berturut turut, untuk menentukan responden penganggur atau tidak.

2. Sedang Cari kerja : Responden mempunyai karakteristik pengangguran jika responden sedang berusaha mencari pekerjaan
3. Jam kerja : Jumlah jam kerja responden melakukan pekerjaan yang terakumulasi dalam rentan waktu satu minggu, jika jam kerja responden kurang dari 35 jam dalam 1 minggu maka bisa dikatakan mempunyai karakteristik pengangguran
4. Status Sementara tidak bekerja : jika responden pada waktu dilakukan survey statusnya sementara tidak bekerja mungkin karena responden sedang mengambil cuti atau sedang sakit maka responden tersebut mempunyai karakteristik pengangguran.
5. Masih mencari usaha baru : Jika responden pada waktu dilakukan survey sedang mempersiapkan usaha baru maka juga bisa dikatakan mempunyai karakteristik pengangguran.
6. Alasan tidak mencari pekerjaan : Alasan utama responden tidak mencari pekerjaan atau mempersiapkan usaha baru ,alasanya apakah sudah tidak mampu bekerja, sudah merasa cukup, putus asa dalam mencari pekerjaan, sekolah ,mengurus rumah tangga atau yang lain. Jika alasan responden sudah tidak mampu bekerja atau sudah putus asa maka juga mempunyai karakteristik pengangguran.
7. Klasifikasi wilayah : wilayah tempat responden yang terletak di pedesaan atau perkotaan.
8. Umur : Umur responden, menurut konsep BPS umur tenaga kerja ialah diatas 15 tahun.
9. Pendidikan terakhir : Pendidikan terakhir yang ditamatkan oleh responden.
10. Jurusan : Jurusan pendidikan terakhir yang diikuti responden untuk melihat persebaran kompetensi responden yang bekerja atau yang menganggur
11. Pernah Mengikuti Pelatihan : apakah responden pernah mengikuti pelatihan kerja atau tidak pernah.
12. Hubungan dengan kepala rumah tangga : hubungan responden dalam keluarga apakah sebagai kepala rumah tangga kalau tidak apakah hubungannya dengan kepala rumah tangga.

### 3.1.2 Normalisasi Data

Pada tahap ini adalah tahapan dimana data pada waktu akan diolah dalam praktiknya biasanya selalu ada nilai yang salah atau kosong pada satu atau lebih fitur dalam data keseluruhan. Nilai yang salah ini seharusnya bernilai angka tetapi bernilai karakter, atau nilai yang disimpan berada di luar jangkauan nilai yang seharusnya dimasukkan. Masalah seperti ini bisa terjadi karena banyak penyebab, seperti input dari user yang dilakukan sembarangan, data yang didapat dari kuesioner yang biasanya juga tidak diisi secara lengkap oleh responden, basis data, dan antar muka aplikasi yang tidak taat integritas data, alat ukur yang sudah tidak standar sehingga memberikan hasil yang salah dan sebagainya. Pada tahap normalisasi data ini juga tahap dimana data set yang diperoleh dari sakernas dari BPS tersebut sudah didapat dalam format \*.sav atau format SPSS, data survey tersebut sudah berupa data Nominal dan data Ordinal, kemudian dilakukan transformasi data dengan dilakukan export ke dalam software Ms.Excell. Data yang bersifat kosong yang diisi dengan karakter “.” diubah menjadi nilai numerik “0” yang bersifat numerik. Kemudian menghapus karakter yang tidak digunakan dalam penelitian ini, dan setelah selesai data siap untuk dilakukan data cleaning untuk dapat diolah.

### 3.1.3 Pembersihan data (Data Cleaning)

Pada tahap ini dilakukan setelah melalui proses normalisasi data yaitu menangani data yang salah sehingga data yang didapat sudah mengalami pembetulan data. Langkah selanjutnya ialah melakukan pembersihan data, proses pembersihan data dalam hal ini ialah menghapus data yang tidak diperlukan dalam penelitian. Data tersebut tidak sesuai dengan konsep dan tujuan dari penelitian ini, dalam hal ini langkah – langkah pembersihan data :

1. Menghapus rentan nilai pada atribut umur pada responden yang berusia dibawah 15 tahun : Dari konsep tenaga kerja menurut Badan pusat Statistik (BPS) penduduk usia kerja ialah penduduk yang berusia diatas 15 tahun jadi penduduk yang dibawah 15 tahun tidak termasuk dalam penelitian ini.

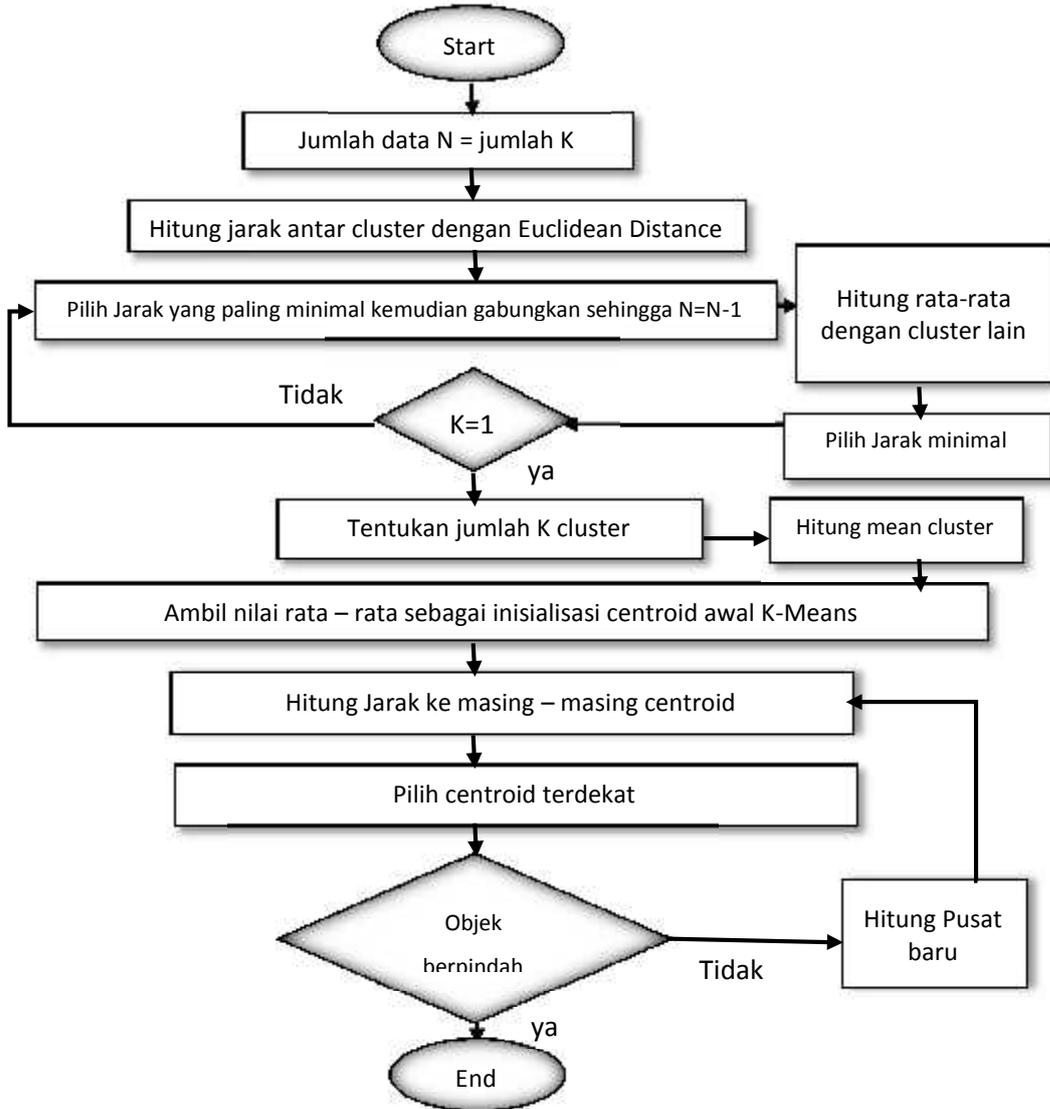
2. Menghapus data responden yang mempunyai status bukan angkatan kerja : yang termasuk bukan angkatan kerja adalah mereka yang dalam kegiatannya selama seminggu yang lalu paling banyak mengurus rumah tangga, pelajar yang masih bersekolah di sekolah formal atau informal atau kegiatan pribadi yang tidak berhubungan dengan kegiatan ekonomi.
3. Menghapus data responden yang bukan angkatan kerja dan yang tidak mempunyai jam kerja yang berhubungan dengan ekonomi keluarga.
4. Menghapus data responden yang bukan angkatan kerja yang tidak mencari kerja dengan alasan sekolah atau mengurus rumah tangga dan tidak mempunyai jam kerja dalam perekonomian.

### 3.2 Pengklasteran Data

Pada tahap ini ialah tahap setelah dilakukan normalisasi dan proses pembersihan data yaitu melakukan proses pengklasteran, dimana pengklasteran ini ialah tahap pengelompokan data angkatan kerja untuk mengelompokkan responden yang berstatus sebagai pengangguran atau bukan, dan mengelompokkan karakteristik pengangguran dengan beberapa faktor yang berhubungan yaitu wilayah, pendidikan terakhir, umur, Jenis Kelamin, Kompetensi keahlian, hubungan status dalam keluarga. Pengklasteran ini dilakukan melalui 2 tahap yaitu melalui Average Linkage clustering dan K-Means clustering tujuannya dilakukannya proses Average Linkage clustering ialah adanya kelemahan algoritma K-means yang bersifat local optima yaitu hasil yang didapat K-means terkadang baik terkadang jelek, hal tersebut terjadi karena tidak ada perhitungan pasti untuk membangkitkan pusat centroid awal yang dilakukan algoritma K-means, karena selama ini untuk menentukan pusat centroid awal dengan cara random sehingga jika centroid yang digunakan tepat maka hasil yang didapatkan bagus sebaliknya jika centroid yang digunakan kurang tepat maka hasil yang didapatkan kurang bagus, bagus atau tidak maksudnya adalah jarak kemiripan antar anggota cluster kecil dan jarak antar cluster yang berbeda tinggi. Maka dengan dilakukan pengklasteran dengan Average Linkage clustering adalah untuk mendapatkan nilai pusat centroid awal yang digunakan untuk algoritma K-means.

### 3.2.1 Algoritma Average Linkage dan K-means Clustering

Pada tahap ini akan menjelaskan proses pengelompokan diawali dengan algoritma Average Linkage clustering kemudian hasil dari pengelompokannya dijadikan acuan untuk penentuan dari centroid awal untuk algoritma K-Means,



Gambar 3. 2 Flowchart algoritma Average Linkage K-means clustering

awal mula dari langkah – langkah pengelompokan pada Average Linkage clustering ialah menentukan jumlah cluster, dalam hal ini pada Average Linkage clustering jumlah cluster merupakan jumlah data kemudian setelah itu menentukan jarak antar cluster dengan menggunakan Euclidean, kemudian hasil pengukuran jarak dimasukkan kedalam matrik pengukuran jarak guna mengetahui nilai jarak

yang paling minimum, kemudian jarak yang paling kecil digabungkan menjadi satu cluster, jika  $N =$  jumlah K cluster maka  $N=N-1$ , kemudian tentukan nilai yang paling minimal, setelah itu apakah jumlah cluster sama dengan satu jika belum maka nilai minimum tadi digabung menjadi satu cluster. Lakukan pengukuran jarak antar cluster kembali untuk menentukan nilai minimumnya kemudian pilih nilai minimumnya dan gabung menjadi satu cluster sehingga jumlah cluster semua tergabung menjadi satu. Setelah jumlah cluster terbentuk menjadi satu kemudian tentukan jumlah cluster yang diinginkan, jika jumlah cluster = 2 maka lakukan cut off pada percabangan pada grafik dendogram menjadi 2 bagian. Persamaan dari algoritma Average Linkage terdapat pada persamaan (2.4). Kemudian setelah pengelompokan melalui Average Linkage clustering didapat menjadi 2 cluster kemudian menghitung rata – rata dari masing – masing cluster untuk menentukan centroid awal yang digunakan untuk algoritma K-means, penentuan centroid awal menggunakan persamaan (3.1).

$$\bar{x}C_m = \sum_{i=1}^N \frac{x_i C_m}{N} \quad \text{-----} \quad (3.1)$$

Dimana :

$\bar{x}C_m$  = centroid awal yang merupakan rata – rata dari data yang berada dalam cluster

$X_i C_m$  = data X ke i pada centroid m

$N$  = Jumlah data

Kemudian setelah didapatkan centroid awal tahap selanjutnya ialah melakukan pengklasteran pada algoritma K-Means, dengan menghitung jarak dengan masing – masing centroid dengan Euclidean distance bertujuan untuk menentukan anggota masing – masing cluster jika jarak anggota cluster lebih dekat dengan centroid cluster 1 maka anggota cluster masuk dalam cluster 1 atau jika anggota cluster dekat dengan centroid cluster 2 maka anggota cluster masuk kedalam cluster 2 , maka persamaan dari pengukuran jarak menggunakan Euclidean ditunjukkan ke dalam dengan persamaan (2.1). Selanjutnya dilakukan secara

berulang menentukan centroid baru kemudian menghitung jarak dengan masing - masing centroid sampai masing - masing anggota cluster tidak berpindah.

### 3.2.2 Perhitungan manual dari data set sample

Dalam hal ini banyak data sama dengan banyak cluster yang digunakan, karena banyak data merupakan jumlah clusternya. Nama parameter yang digunakan akan diubah variabelnya sehingga lebih memudahkan penulis dalam melakukan perhitungan

1. Atribut Bekerja seminggu yang lalu variable X1
2. Atribut Cari kerja menjadi variable X2
3. Atribut Jam Kerja menjadi X3
4. Atribut Sementara tidak bekerja menjadi X4
5. Atribut masih mencari usaha baru menjadi X5
6. Atribut Alasan tidak bekerja menjadi X6

Berikut merupakan masing – masing atribut atau fitur yang disajikan dalam Table

3.1 Tabel Daftar Nama Cluster pada masing – masing data :

Tabel 3. 1 Daftar Nama Cluster Pada Masing – Masing Data

Responden	Cluster ke	X1	X2	X3	X4	X5	X6
1	C0	1	0	2	2	5	2
2	C1	1	0	2	2	3	2
3	C2	1	0	2	2	4	3
4	C3	1	0	2	2	6	3
5	C4	2	0	2	2	7	0
Responden	Cluster ke	X1	X2	X3	X4	X5	X6
...	...	...	...	...	...	...	...
8654	C8654	2	0	1	2	7	0

Setelah selesai memberikan nama cluster pada data yang dimiliki dimulai dengan cluster C0 maka selanjutnya dilakukan penghitungan jarak dengan menggunakan Euclidean jarak C0 dengan cluster yang lainnya :

$$d(C0,C0)=\sqrt{(X1C0 - X1C0)^2 + (X2C0 - X2C0)^2 + \dots \dots (X6C0 - X6C0)^2}$$

$$d(C0,C0)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(5-5)^2+(2-2)^2}$$

$$=0$$

$$d(C0,C1)=\sqrt{(X1C0-X1C1)^2+(X2C0-X2C1)^2+\dots+(X6C0-X6C1)^2}$$

$$d(C0,C1)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(3-5)^2+(2-2)^2}$$

$$=2$$

$$d(C0,C2)=\sqrt{(X1C0-X1C2)^2+(X2C0-X2C2)^2+\dots+(X6C0-X6C2)^2}$$

$$d(C0,C2)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(4-5)^2+(3-2)^2}$$

$$=1,41$$

$$d(C0,C3)=\sqrt{(X1C0-X1C3)^2+(X2C0-X2C3)^2+\dots+(X6C0-X6C3)^2}$$

$$d(C0,C3)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(6-5)^2+(3-2)^2}$$

$$=1,41$$

$$d(C0,C4)=\sqrt{(X1C0-X1C4)^2+(X2C0-X2C4)^2+\dots+(X6C0-X6C4)^2}$$

$$d(C0,C4)=\sqrt{(2-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(7-5)^2+(0-2)^2}$$

$$=3$$

$$d(C0,C8654)=\sqrt{(X1C0-X1C8654)^2+(X2C0-X2C8654)^2+\dots+(X6C0-X6C8654)^2}$$

$$d(C0,C8654)=\sqrt{(2-1)^2+(0-0)^2+(1-2)^2+(2-2)^2+(7-5)^2+(0-2)^2}$$

$$=3,16$$

Menghitung jarak C1 dengan cluster lainnya :

$$d(C1,C0)=\sqrt{(X1C1-X1C0)^2+(X2C1-X2C0)^2+\dots+(X6C1-X6C0)^2}$$

$$d(C1,C0)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(3-5)^2+(2-2)^2}$$

$$=2$$

$$d(C1,C2)=\sqrt{(X1C1-X1C2)^2+(X2C1-X2C2)^2+\dots+(X6C1-X6C2)^2}$$

$$d(C1,C2)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+\dots+(3-4)^2+(2-3)^2}$$

$$=1,41$$

$$d(C1,C3)=\sqrt{(X1C1-X1C3)^2+(X2C1-X2C3)^2+\dots+(X6C1-X6C3)^2}$$

$$d(C1,C3)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+\dots+(3-6)^2+(2-3)^2}$$

$$=3,16$$

$$d(C1,C4)=\sqrt{(X1C1 - X1C4)^2 + (X2C1 - X2C4)^2 + \dots + (X6C1 - X6C4)^2}$$

$$d(C1,C4)=\sqrt{(2-1)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + \dots + (7-3)^2 + (0-2)^2}$$

$$=4,58$$

$$d(C1,C8654)=\sqrt{(X1C1 - X1C8654)^2 + (X2C1 - X2C8654)^2 + \dots + (X6C1 - X6C8654)^2}$$

$$d(C1,C8654)=\sqrt{(2-1)^2 + (0-0)^2 + (1-2)^2 + (2-2)^2 + \dots + (7-3)^2 + (0-2)^2}$$

$$=4,69$$

Menghitung C2 dengan cluster lainnya :

$$d(C2,C0)=\sqrt{(X1C2 - X1C0)^2 + (X2C2 - X2C0)^2 + \dots + (X6C2 - X6C0)^2}$$

$$d(C2,C0)=\sqrt{(1-1)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + (4-5)^2 + (3-2)^2}$$

$$=1,41$$

$$d(C2,C1)=\sqrt{(X1C2 - X1C1)^2 + (X2C2 - X2C1)^2 + \dots + (X6C2 - X1)^2}$$

$$d(C2,C1)=\sqrt{(1-1)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + (4-3)^2 + (3-2)^2}$$

$$=1,41$$

$$d(C2,C3)=\sqrt{(X1C2 - X1C3)^2 + (X2C2 - X2C3)^2 + \dots + (X6C2 - X6C3)^2}$$

$$d(C2,C3)=\sqrt{(1-1)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + (4-6)^2 + (3-3)^2}$$

$$=2$$

$$d(C2,C4)=\sqrt{(X1C2 - X1C4)^2 + (X2C2 - X2C4)^2 + \dots + (X6C2 - X6C4)^2}$$

$$d(C2,C4)=\sqrt{(1-2)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + (4-7)^2 + (3-0)^2}$$

$$=4,36$$

$$d(C2,C8654)=\sqrt{(X1C2 - X1C8654)^2 + (X2C2 - X2C8654)^2 + \dots + (X6C2 - X6C8654)^2}$$

$$d(C2,C8654)=\sqrt{(1-2)^2 + (0-0)^2 + (2-1)^2 + (2-2)^2 + (4-7)^2 + (3-0)^2}$$

$$=4,47$$

Menghitung C3 dengan cluster lainnya :

$$d(C3,C0)=\sqrt{(X1C3 - X1C0)^2 + (X2C3 - X2C0)^2 + \dots + (X6C3 - X6C0)^2}$$

$$d(C3,C0)=\sqrt{(1-2)^2 + (0-0)^2 + (2-2)^2 + (2-2)^2 + (4-7)^2 + (3-0)^2}$$

$$=1,41$$

$$d(C3,C1)=\sqrt{(X1C3 - X1C1)^2 + (X2C3 - X2C1)^2 + \dots + (X6C3 - X6C1)^2}$$

$$d(C3,C1)=\sqrt{(1 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (6 - 3)^2 + (3 - 2)^2}$$

$$=3,16$$

$$d(C3,C2)=\sqrt{(X1C3 - X1C2)^2 + (X2C3 - X2C2)^2 + \dots + (X3C3 - X6C2)^2}$$

$$d(C3,C2)=\sqrt{(1 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (6 - 34 + (3 - 3)^2}$$

$$=2$$

$$d(C3,C4)=\sqrt{(X1C3 - X1C4)^2 + (X2C3 - X2C4)^2 + \dots + (X6C3 - X6C4)^2}$$

$$d(C3,C4)=\sqrt{(1 - 2)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (6 - 37 + (3 - 0)^2}$$

$$=3,31$$

$$d(C3,C8654)=\sqrt{(X1C3 - X1C8654)^2 + (X2C3 - X2C8654)^2 + \dots + (X6C3 - X6C8654)^2}$$

$$d(C3,C8654)=\sqrt{(1 - 2)^2 + (0 - 0)^2 + (2 - 1)^2 + (2 - 2)^2 + (6 - 7)^2 + (3 - 0)^2}$$

$$=3,46$$

Menghitung C4 dengan cluster lainnya :

$$d(C4,C0)=\sqrt{(X1C4 - X1C0)^2 + (X2C4 - X2C0)^2 + \dots + (X6C4 - X6C0)^2}$$

$$d(C4,C0)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (7 - 5)^2 + (0 - 2)^2}$$

$$=3$$

$$d(C4,C1)=\sqrt{(X1C4 - X1C1)^2 + (X2C4 - X2C1)^2 + \dots + (X6C4 - X6C1)^2}$$

$$d(C4,C1)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (7 - 3)^2 + (0 - 2)^2}$$

$$=3$$

$$d(C4,C2)=\sqrt{(X1C4 - X1C2)^2 + (X2C4 - X2C2)^2 + \dots + (X6C4 - X6C2)^2}$$

$$d(C4,C2)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (7 - 4)^2 + (0 - 3)^2}$$

$$=4,36$$

$$d(C4,C3)=\sqrt{(X1C4 - X1C3)^2 + (X2C4 - X2C3)^2 + \dots + (X6C4 - X6C3)^2}$$

$$d(C4,C3)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (2 - 2)^2 + (2 - 2)^2 + (7 - 6)^2 + (0 - 3)^2}$$

$$=3,32$$

$$d(C4,C8654)=\sqrt{(X1C4 - X1C8654)^2 + (X2C4 - X2C8654)^2 + \dots + (X6C8654 - X6C8654)^2}$$

$$d(C4,C8654)=\sqrt{(2 - 2)^2 + (0 - 0)^2 + (2 - 1)^2 + (2 - 2)^2 + (7 - 7)^2 + (0 - 0)^2}$$

$$=1$$

Menghitung C8654 dengan cluter lainnya :

$$d(C8654,C0)=\sqrt{(X1C8654 - X1C0)^2 + (X2C8654 - X2C0)^2 + \dots + (X6C8654 - X6C0)^2}$$

$$d(C8654,C0)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1 - 2)^2 + (2 - 2)^2 + (7 - 5)^2 + (0 - 2)^2}$$

$$=3,16$$

$$d(C8654,C1)=\sqrt{(X1C8654 - X1C1)^2 + (X2C8654 - X2C1)^2 + \dots + (X6C8654 - X6C1)^2}$$

$$d(C8654,C1)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1 - 2)^2 + (2 - 2)^2 + (7 - 3)^2 + (0 - 2)^2}$$

$$=4,69$$

$$d(C8654,C2)=\sqrt{(X1C8654 - X1C2)^2 + (X2C8654 - X2C2)^2 + \dots + (X6C8654 - X6C2)^2}$$

$$d(C8654,C2)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1 - 2)^2 + (2 - 2)^2 + (7 - 4)^2 + (0 - 3)^2}$$

$$=4,47$$

$$d(C8654,C3)=\sqrt{(X1C8654 - X1C3)^2 + (X2C8654 - X2C3)^2 + \dots + (X6C8654 - X6C3)^2}$$

$$d(C8654,C3)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1 - 2)^2 + (2 - 2)^2 + (7 - 6)^2 + (0 - 3)^2}$$

$$=3,46$$

$$d(C8654,C4)=\sqrt{(X1C8654 - X1C4)^2 + (X2C8654 - X2C4)^2 + \dots + (X6C8654 - X6C4)^2}$$

$$d(C8654,C4)=\sqrt{(2 - 2)^2 + (0 - 0)^2 + (1 - 2)^2 + (2 - 2)^2 + (7 - 7)^2 + (0 - 0)^2}$$

$$=1$$

Pada Tabel 3.2 Berikut merupakan matrik perhintungan jarak antar cluster

Tabel 3. 2 Matrik Perhitungan Jarak Euclidean

Cluster	C0	C1	C2	C3	C4	...	C8654
C0	0.00	2.00	1.41	1.41	3.00	...	3.16
C1	2.00	0.00	1.41	3.16	3.00	...	4.69
C2	1.41	1.41	0.00	2.00	4.36	...	4.47
C3	1.41	3.16	2.00	0.00	3.32	...	3.46
C4	3.00	4.58	4.36	3.32	0.00	...	1.00
...	...	...	....	...	...	....	.....
C8654	3.16	4.69	4.47	3.46	1.00	.....	0.00

Setelah Table 3.2 hasil penghitungan jarak didapatkan, langkah berikutnya ialah menentukan jarak yang paling dekat untuk digabungkan menjadi sebuah cluster baru. Dari penggabungan ini maka banyak data menjadi  $N=N-1$ .

Lalu hitung jarak rata – rata cluster yang baru dengan cluster lainnya. Perhitungan jarak cluster baru ini menggunakan Average Linkage. Dari Table 3.2 ditemukan jarak terdekat adalah C4/C8654 dengan jarak 1, keduanya digabung menjadi cluster tunggal yaitu C4/C8654 Kemudian hitung dari jarak objek dengan objek lainnya. Pada clusterisasi Average linkage, aturanya adalah jarak antara objek yang digabung dengan objek lainnya dengan menghitung jarak rata – rata nya kemudian dicari jarak terpendek dari suatu anggota pada cluster, kemudian objek tersebut akan digabungkan. Setelah penggabungan C4 dan C8654 maka akan terbentuk cluster baru seperti pada Tabel 3.3 Matrik gabungan cluster C4 dan C8654, kemudian langkah berikutnya ialah menghitung jarak C4/C8654 terhadap cluster lainnya dengan perhitungan rata – rata.

$$D(C4/C8654)C0 = D(C4,C0) + D(C8654,C0) / 2 = 3,08$$

$$D(C4/C8654)C1 = D(C4,C1) + D(C8654,C1) / 2 = 4,64$$

$$D(C4/C8654)C2 = D(C4,C2) + D(C8654,C2) / 2 = 4,41$$

$$D(C4/C8654)C3 = D(C4,C3) + D(C8654,C3) / 2 = 2,84$$

Setelah melakukan perhitungan jarak cluster maka hasil perhitungan akan membentuk suatu matrik seperti Tabel 3.3 :

Tabel 3. 3 Matrik gabungan cluster C4 dan C8654

Cluster	C0	C1	C2	C3	...	C8654,C4
C0	0	2	1.41	1.41		3.08
C1	2	0	1.41	3.16		4.64
C2	1.41	1.41	0	2		4.42
C3	1.41	3.16	2	0		2.85
...	...	...	....	...	....	....
C8654,C4	3.08	4.63	4.415517449	2.85	.....	0

Setelah di dapatkan tabel matrik seperti pada tabel 3.3 diatas selanjutnya cari sepasang cluster yang jaraknya paling dekat, dan gabungkan sehingga didapat sebuah cluster baru. Dari penggabungan ini maka banyak data akan menjadi  $n = n - 1$ . Lalu hitung jarak antara cluster yang baru dengan masing-masing cluster yang lainnya. Penghitungan jarak cluster baru ini menggunakan Average-Linkage pada persamaan (2.7). Dari tabel matrik pada tabel 3.3 ditemukan pasangan jarak terdekat adalah C0 dan C3, dengan jarak = 0.5. Keduanya digabung menjadi Cluster tunggal yaitu C0/C3. Level dari cluster baru adalah  $L(C0/C3) = 1,41$ . Kemudian dihitung jarak dari objek gabungan ini ke semua objek lainnya. Pada Clusterisasi Average linkage, aturannya adalah jarak antara objek gabungan ke objek lainnya sama dengan jarak terpendek dari nilai rata – rata suatu anggota pada Cluster yang lainnya di luar objek

$$D_{\min}(C0,C3)=1,4$$

$$D(C0,C3)C1= D(C0,C1)+D(C3,C1)/2=2,58$$

$$D(C0,C3)C2= D(C0,C2)+D(C3,C2)/2=1,7$$

$$D((C0,C3),(C4/C8654))= D(C0,C4/C14514)+D(C3,C4/C8654)/2=2,96$$

Setelah melakukan perhitungan jarak cluster maka hasil perhitungan akan membentuk suatu matrik seperti Tabel 3.4 :

Tabel 3. 4 Matrik Gabungan C0/C3

Cluster	C0,C3	C1	C2	...	C8654,C4
C0,C3	0	2.58	1.7	...	2.96
C1	2.58	0	1.41	...	4.64
C2	1.7	1.41	0	...	4.42
...	...	...	...	0	...
C8654,C4	2.96	4.6	4.42	...	0

Pada Tabel 3.4 jarak terdekat pada cluster C1/C2 dengan jarak = 0.5, kemudian gabungkan sehingga didapat sebuah cluster baru. Dari penggabungan ini maka banyak data akan menjadi  $n = n - 1$ . Lalu hitung jarak antara cluster yang baru dengan masing-masing cluster yang lainnya sesuai dengan aturan pada

algoritma Average Linkage. Setelah melakukan perhitungan jarak cluster maka hasil perhitungan akan membentuk suatu matrik seperti Tabel 3.5

Tabel 3. 5 Matrik Gabungan C0/C4/C1 dengan C14514

Cluster	C0C3	C1C2	...	C8654C4
C0C3	0	2.14	...	2.96
C1C2	2.14	0	...	4.53
...	...	...	...	...
C8654C4	2.96	4.53	...	0

Pada Tabel 3.5 jarak terdekat pada cluster C1/C2 dengan C0/C3 dengan jarak = 2.14, kemudian gabungkan sehingga didapat sebuah cluster baru. Dari penggabungan ini maka banyak data akan menjadi  $n = n - 1$ . Lalu hitung jarak antara cluster yang baru dengan masing-masing cluster yang lainnya. Setelah melakukan perhitungan jarak cluster maka hasil perhitungan akan membentuk suatu matrik seperti Tabel 3.6

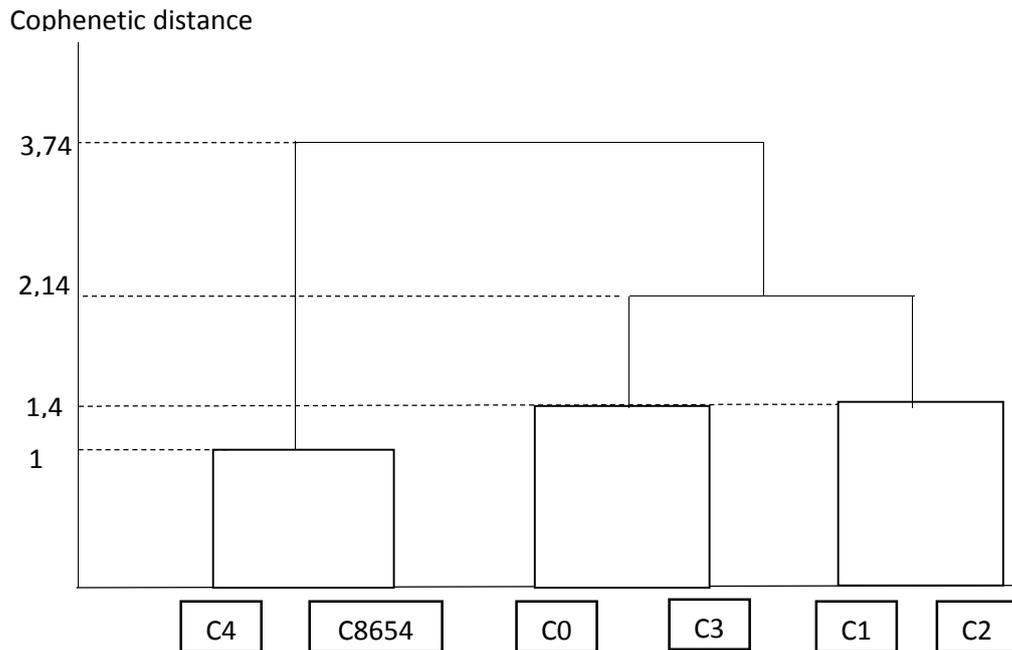
Tabel 3. 6 Matrik Gabungan C0/C3/C1/C2 dengan C8654/C4

Cluster	C0C3C1C2	....	C8654C4
C0C3C1C2	0	...	3.74
...	...	...	...
C8654C4	3.74		0

Tabel 3. 7 Matrik Gabungan C0/C3/C1/C2/C8654/C4

Cluster	C0/C3/C1/C2/C8654/C4
C0/C3/C1/C2/C8654/C4	0

Pada Tabel 3.6 diperoleh pengukuran jarak terakhir sebesar 3,74 kemudian gabung semua cluster menjadi 1 cluster seperti pada Tabel 3.7. Setelah semua cluster tergabung menjadi 1 cluster besar maka proses pada algoritma Average Linkage telah selesai. Hasil dari proses ini ditampilkan dalam suatu Dendogram.



Gambar 3. 3 Dendogram Average Linkage

Setelah proses hierarchical clustering selesai, maka dilanjutkan dengan proses k-means. Proses k-means sendiri diawali dengan penentuan banyak cluster yang ingin di bentuk kemudian dilanjutkan dengan penentuan pusat awal cluster (centroid awal) yang dilakukan secara random. Namun karena menggunakan gabungan antara algoritma hierarchical clustering dan k-means, maka penentuan pusat cluster dilakukan dengan mencari means (rata-rata) dari data yang berada pada sebuah cluster hasil dari perhitungan Average Linkage clustering yang sebelumnya telah dilakukan. Pada kasus ini penulis ingin membentuk 2 cluster. Karena ingin membentuk 2 cluster maka anggota tiap cluster baru adalah adalah :

Cluster A : C4 dan C8654

Cluster B : C0, C1, C2, C3

Pengambilan pusat cluster (centroid) dilakukan dengan menghitung mean (rata - rata) pada masing-masing cluster dengan membagi jumlah data yang didapatkan untuk setiap clusternya. Adapun tujuan dari penghitungan ini adalah agar setiap cluster memiliki anggota data pada iterasi pertama. Untuk mencari mean (rata-rata) dengan menggunakan rumus persamaan (3.1)

Tabel 3. 8 Anggota cluster yang dibentuk

Cluster Baru	Cluster Lama	X1	X2	X3	X4	X5	X6
CB	C0	1	0	2	2	5	2
CB	C1	1	0	2	2	3	2
CB	C2	1	0	2	2	4	3
CB	C3	1	0	2	2	6	3
CA	C4	2	0	2	2	7	0
...	...	...	...	...	...	...	...
CA	C8654	2	0	1	2	7	0

Dari hasil anggota cluster yang dibentuk pada Tabel 3.8, kemudian dihitung nilai rata – rata dari masing – masing cluster untuk menentukan nilai centroid awal yang terdapat pada Tabel 3.9.

Tabel 3. 9 Centroid awal masing – masing cluster

Cluster	X1	X2	X3	X4	X5	X6
CA	2	0	1.5	2	7	0
CB	1	0	2	2	4.5	2.5

Setelah hasil centroid awal dari setiap cluster didapat, selanjutnya dilakukan penghitungan jarak anggota cluster ke setiap centroidnya dengan menggunakan rumus euclidiance distance yakni persamaan (2.1). Hasil dari penghitungan jarak anggota cluster dengan centroidnya akan berpengaruh pada penempatan setiap data ke cluster yang telah ditentukan. Berikut ini merupakan perhitungan anggota cluster pada tiap masing – masing centroid awal :

Jarak cluster CA dengan centroidnya:

$$d(CA,C0)=\sqrt{(X1C - X1C0)^2 + (X2C - X2C0)^2 + \dots \dots + (X6C - X6C0)^2}$$

$$d(CA,C0)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1,5 - 2)^2 + (2 - 2)^2 + (7 - 5)^2 + (0 - 2)^2}$$

$$=3,04$$

$$d(CA,C1)=\sqrt{(X1CA - X1C1)^2 + (X2C - X2C1)^2 + \dots \dots + (X6C - X6C1)^2}$$

$$d(CA,C1)=\sqrt{(2 - 1)^2 + (0 - 0)^2 + (1,5 - 2)^2 + (2 - 2)^2 + (7 - 3)^2 + (0 - 2)^2}$$

$$=4,6$$

$$d(CA,C2)=\sqrt{(X1C - X1C2)^2 + (X2C - X2C2)^2 + \dots \dots + (X6C - X6C2)^2}$$

$$d(CA,C2)=\sqrt{(2-1)^2+(0-0)^2+(1,5-2)^2+(2-2)^2+(7-4)^2+(0-3)^2}$$

$$=4,39$$

$$d(CA,C3)=\sqrt{(X1C - X1C3)^2+(X2C - X2C3)^2+\dots+(X6C - X6C3)^2}$$

$$d(CA,C3)=\sqrt{(2-1)^2+(0-0)^2+(1,5-2)^2+(2-2)^2+(7-6)^2+(0-3)^2}$$

$$=3,35$$

$$d(CA,C4)=\sqrt{(X1C - X1C4)^2+(X2C - X2C4)^2+\dots+(X6C - X6C4)^2}$$

$$d(CA,C4)=\sqrt{(2-2)^2+(0-0)^2+(1,5-2)^2+(2-2)^2+(7-7)^2+(0-0)^2}$$

$$=0,5$$

$$d(CA,C8654)=\sqrt{(X1C - X1C8654)^2+(X2C - X2C8654)^2+\dots+(X6C - X6C8654)^2}$$

$$d(CA,C8654)=\sqrt{(2-2)^2+(0-0)^2+(1,5-1)^2+(2-2)^2+(7-7)^2+(0-0)^2}$$

$$=0,5$$

Jarak cluster CB dengan centroidnya:

$$d(CB,C0)=\sqrt{(X1C - X1C0)^2+(X2C - X2C0)^2+\dots+(X6C - X6C0)^2}$$

$$d(CB,C0)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(4,5-5)^2+(2,5-2)^2}$$

$$=0,7$$

$$d(CB,C1)=\sqrt{(X1C - X1C1)^2+(X2C - X2C1)^2+\dots+(X6C - X6C1)^2}$$

$$d(CB,C1)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(4,5-3)^2+(2,5-2)^2}$$

$$=1,58$$

$$d(CB,C2)=\sqrt{(X1C - X1C2)^2+(X2C - X2C2)^2+\dots+(X6C - X6C2)^2}$$

$$d(CB,C2)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(4,5-4)^2+(2,5-3)^2}$$

$$=0,7$$

$$d(CB,C3)=\sqrt{(X1C - X1C3)^2+(X2C - X2C3)^2+\dots+(X6C - X6C3)^2}$$

$$d(CB,C3)=\sqrt{(1-1)^2+(0-0)^2+(2-2)^2+(2-2)^2+(4,5-6)^2+(2,5-3)^2}$$

$$=1,58$$

$$d(CB,C4)=\sqrt{(X1C - X1C4)^2+(X2C - X2C4)^2+\dots+(X6C - X6C4)^2}$$

$$d(CB,C4)=\sqrt{(1-2)^2+(0-0)^2+(1,5-2)^2+(2-2)^2+(4,5-7)^2+(2,5-0)^2}$$

$$=3,67$$

$$d(CB,C8654)=\sqrt{(X1C - X1C8654)^2+(X2C - X2C8654)^2+\dots+(X6C - X6C8654)^2}$$

$$d(CB,C8654)=\sqrt{(1-2)^2+(0-0)^2+(2-1)^2+(2-2)^2+(4,5-7)^2+(2,5-0)^2}$$

$$=3,8$$

Dari hasil perhitungan jarak Euclidean distance cluster dengan centroidnya ditampilkan dalam Tabel 3.10.

Tabel 3. 10 Hasil perhitungan jarak cluster dengan centroidnya

Cluster	X1	X2	X3	X4	X5	X6	dCA	dCB
C0	1	0	2	2	5	2	3,41	0,7
C1	1	0	2	2	3	2	4,6	1,58
C2	1	0	2	2	4	3	4,39	0,7
C3	1	0	2	2	6	3	3,54	1,58
C4	2	0	2	2	7	0	0,5	3,67
...	...	...	...	...	...	...	...	...
C8654	2	0	1	2	7	0	0,5	3,8

Setelah melakukan penghitungan jarak centroid dengan masing-masing clusternya, selanjutnya adalah dilakukan pengelompokkan jarak terkecil disetiap masing-masing cluster. Pada tabel 3.10 dapat dilihat bahwa jarak paling dekat pada C0 adalah 0.7, maka dari itu cluster C0 merupakan anggota dari cluster CB. Demikian juga C1 jarak paling dekat adalah 1,58, sehingga cluster C1 merupakan anggota dari cluster CB. C2 jarak paling dekat adalah 0,7, maka C2 merupakan anggota cluster dari CB. C3 jarak terdekate adalah 1,58, maka C3 merupakan anggota cluster dari CB. C4 jarak terdekate adalah 0,5, maka C4 merupakan anggota cluster dari CA. C8654 jarak terdekate adalah 0,5, maka C8654 merupakan anggota cluster dari CA. Untuk lebih jelasnya anggota data awal untuk masing-masing cluster baru akan disajikan pada Tabel 3.11.

Hasil dari tabel 3.11 adalah hasil sementara, maka akan terjadi pengulangan (iterasi) pada proses k-means hingga tidak ada perpindahan pada setiap anggota cluster. Maka dari itu proses akan diulang kembali dari penentuan centroid baru dan menghitung kembali jarak antara cluster dan centroidnya. Pengulangan pertama dilakukan dengan menggunakan data yang merupakan tabel yang anggota dari cluster baru telah terbentuk.

Tabel 3. 11 Anggota cluster baru yang dibentuk

Cluster Baru	Cluster Lama	X1	X2	X3	X4	X5	X6
CB	C0	1	0	2	2	5	2
CB	C1	1	0	2	2	3	2
CB	C2	1	0	2	2	4	3
CB	C3	1	0	2	2	6	3
CA	C4	2	0	2	2	7	0
...	...	...	...	...	...	...	...
CA	C8654	2	0	1	2	7	0

Tabel 3. 12 Hasil centroid baru masing – masing cluster Iterasi 1

Cluster	X1	X2	X3	X4	X5	X6
CA	1	0	2	2	4.5	2.5
CB	2	0	1.5	2	7	0

Setelah hasil centroid baru terbentuk pada masing – masing cluster, tahap selanjutnya melakukan perulangan (iterasi) untuk menghitung jarak pada cluster dengan masing – masing centroidnya sampai anggota cluster tidak berpindah dari centroidnya. Sehingga akan menghasilkan hasil akhir clustering yang ditunjukkan pada Tabel 3.14 merupakan hasil akhir dari gabungan algoritma Average Linkage dan K-Means Clustering.

Tabel 3. 13 Hasil akhir proses clustering

Cluster Akhir	Cluster Lama	X1	X2	X3	X4	X5	X6
CB	C0	1	0	2	2	5	2
CB	C1	1	0	2	2	3	2
CB	C2	1	0	2	2	4	3
CB	C3	1	0	2	2	6	3
CA	C4	2	0	2	2	7	0
...	...	...	...	...	...	...	...
CA	C8654	2	0	1	2	7	0

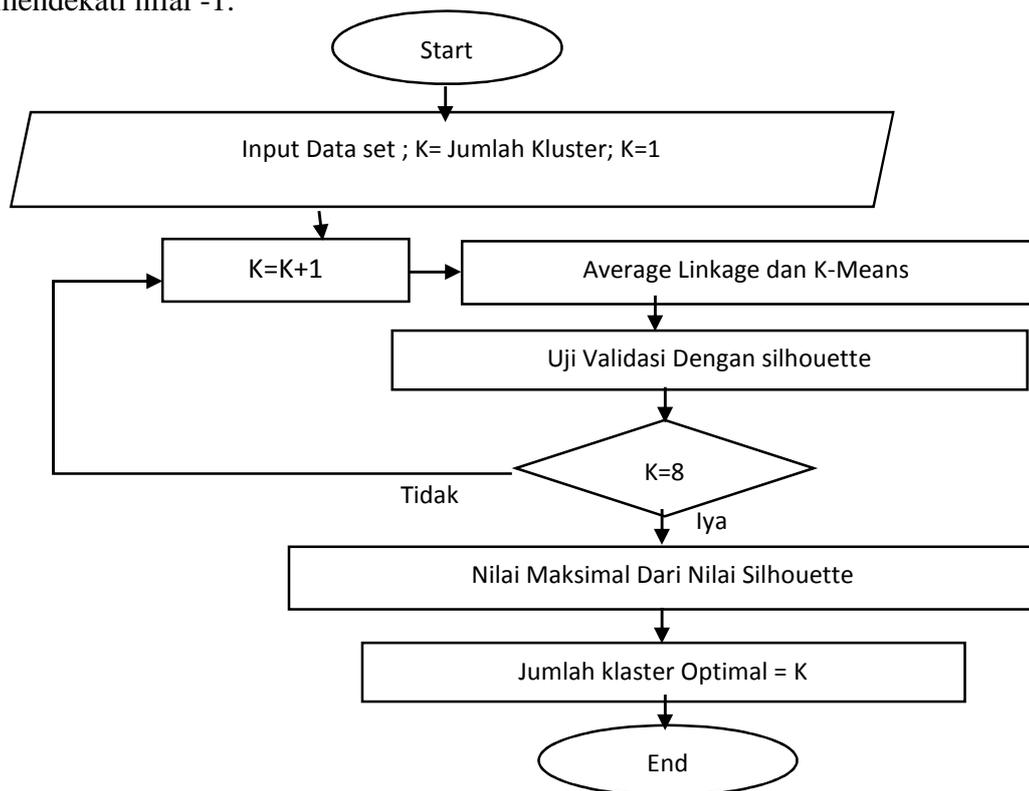
### 3.3 Evaluasi dan Validasi Cluster

proses evaluasi dilakukan setelah dilakukan proses pengklasteran atau pada saat didalam proses clustering itu sendiri, proses ini penting karena evaluasi cluster ini untuk menguji data set yang digunakan untuk melihat kualitas dan

kekuatan cluster, seberapa baik suatu objek ditempatkan dalam suatu cluster. Jika suatu cluster mempunyai kualitas baik maka tingkat homogenitas cluster tinggi, dalam penentuan pola dan analisa cluster akan semakin mudah. Berikut merupakan tahap evaluasi dalam penelitian ini :

### 3.3.1 Penentuan Jumlah cluster yang tepat (Ideal)

Dalam hal ini ialah penentuan jumlah cluster yang ideal, pada proses dilakukan dengan menggunakan pendekatan metode validasi cluster dengan Sillhoutte index, Menurut (Edy Irwansyah dan Muhammad Faizal, 2015) tentang Advanced Clustering bahwa Sillhouete index dihitung sebagai derajat kepercayaan dalam proses clustering dalam suatu pengamatan dengan cluster yang dikatakan terbentuk baik bila nilai index bernilai 1 dan kondisi sebaliknya jika nilai index mendekati nilai -1.



Gambar 3. 4 Flowchart Mencari Jumlah Cluster Yang Optimal

Pengamatan untuk menentukan cluster yang ideal ialah dengan cara melakukan klasterisasi untuk menghasilkan mulai 2 buah cluster sampai 8 cluster

setelah itu masing – masing hasil cluster diukur dengan silhouette index untuk mengetahui nilai kerapatan datanya, nilai yang mendekati angka 1 adalah nilai yang paling bagus. Pada gambar 3.2 merupakan flowchart langkah – langkah untuk mendapatkan jumlah cluster yang optimal, didapat dari nilai validasi Silhoutte yang paling besar melalui uji coba pengelompokan clustering untuk menghasilkan 2 cluster sampai menghasilkan 8 cluster.

### 3.4 Interpretasi Data

Pada bagian ini merupakan proses intepretasi data yaitu setelah dilakukan pengklasteran dan optimalisasi jumlah cluter maka didapat hasil pengelompokan yang optimal langkah berikutnya ialah mendeskripsikan pola hasil dari pengelompokan tersebut, ada berapa jumlah responden pada sample data yang mempunyai karakteristik pengangguran dari masing – masing jenis pengangguran, dan bagaimanakah karakteristik pengangguran yang tersebar di wilayah pedesaan dan perkotaan. Dan dari masing – masing cluster karakterik jenis pekerjaan apa yang paling banyak. Faktor – faktor apa yang paling berpengaruh pada responden yang mempunyai karakteristik pengangguran.

### 3.5 Kesimpulan

Pada tahap ini merupakan proses akhir yaitu menyimpulkan dari hasil yang sudah didapat, kemudian manfaat apa yang didapatkan dari hasil penelitian ini ialah untuk menentukan berbagai kebijakan untuk sebagai bahan pertimbangan.

## **BAB IV**

### **HASIL DAN PEMBAHASAN**

Pada bab ini akan dijelaskan tahap implementasi sistem dan analisa hasil cluster serta pengujian validasi cluster untuk mendapatkan jumlah cluster yang optimal serta membandingkan algoritma gabungan Average Linkage dan K-Means dengan K-Means murni. Pada tahap – tahap percobaan uji validasi dan analisa data yang dilakukan dengan menggunakan metode – metode yang telah disebutkan pada bab 3 akan dijelaskan hasilnya pada Bab IV ini. Pada bab IV ini juga akan menjelaskan komponen pendukung yang dipakai untuk mempresentasikan data dengan menggunakan gabungan dari algoritma Average Linkage clustering dan K-means clustering.

#### **4.1 Pemilihan data**

Penyiapan data dilakukan dengan terlebih dahulu dengan pemilihan data. Data yang dijadikan penelitian adalah data dari Survey Angkatan Kerja Nasional tahun 2014 pada bulan februari yang dilakukan oleh Badan Pusat Statistik Propinsi Jawa Timur, jumlah data input yang akan diolah sebanyak 13.266 data sample. Pemilihan data pada tahun 2014 menggambarkan keadaan yang terjadi pada tahun tersebut ,maka harus mempertimbangkan dasar – dasar pemilihan data dan harus diamati secara seksama untuk mencapai tujuan dari penelitian itu, maka pengambilan dari data sakernas Tahun 2014 pada bulan februari berdasarkan beberapa fakta yang dijadikan alasan untuk penelitian ,yaitu :

1. Peneliti yang lebih berdomisili di Jawa Timur sehingga lebih mengenal secara Geografis dari pada propinsi lain sehingga dijadikan tempat penelitian.
2. Menurut data dari BPS (Badan Pusat Statistik) Jumlah dari Tingkat Partisipasi Angkatan Kerja Propinsi Jawa Timur mulai tahun 2011 hingga tahun 2015, mengalami peningkatan dengan jumlah terbesar pada bulan Feberuari tahun 2014.

3. Menurut data dari BPS (Badan Pusat Statistik) penduduk yang berusia 15 tahun keatas menurut jam kerja utama ,yang mempunyai jam kerja paling banyak mulai tahun 2011 sampai tahun 2015 ialah pada tahun 2014 bulan februari sebanyak 5.324 jam kerja.
4. Pada Tahun 2014 jumlah jam kerja sebanyak 0,- jam atau penganggur terbanyak pada bulan februari

Sehingga dari berbagai dasar fakta tersebut maka diharapkan penelitian dapat menghasilkan data – data yang lebih relevan dan lebih baik dalam menunjang tujuan dari penelitian ini. Dari pemilihan fitur pada pada Bab 3 yang digunakan dalam penelitian atau dalam pengelompokannya nanti tidak digunakan secara bersamaan akan tetapi akan dikombinasikan disesuaikan dengan tujuan dari penelitian ini. Dalam penelitian ini fitur yang digunakan dibagi menjadi 2 jenis fitur, yaitu :

1. Fitur utama untuk karakteristik tingkat pengangguran

Pada fitur utama ini berjumlah 6 atribut data yang bertujuan penentu dalam proses pengelompokan data untuk menentukan jenis – jenis pengangguran, fitur utama ini memperhatikan konsep pengangguran menurut (BPS, 2014) tentang laporan Eksekutif Angkatan Kerja di Propinsi Jawa Timur , fitur atau atribut tersebut terdiri dari atribut Status bekerja selama seminggu yang lalu, sedang mencari pekerjaan, jam kerja selama seminggu, status sementara tidak bekerja,mencari usaha baru, dan alasan tidak mencari pekerjaan.

2. Fitur tambahan atau fitur kombinasi

fitur tambahan ini berjumlah 5 fitur atribut merupakan faktor - faktor yang berhubungan dengan tingkat pengangguran, yang merupakan atribut atau fitur tambahan adalah atribut wilayah, pendidikan terakhir, umur, jenis kelamin, dan kompetensi keahlian

Dari kedua jenis fitur tersebut akan dikombinasikan untuk menganalisa karakteristik pengangguran berdasarkan faktor – faktornya, sehingga dapat

memberikan gambaran yang jelas tentang faktor – faktor yang paling berpengaruh dalam tingkat pengangguran.

#### 4.1.1 Fitur utama untuk Karakteristik Tingkat Pengangguran

Dalam pemilihan fitur untuk pengelompokan harus memperhatikan dengan seksama konsep dasar dari teori pengangguran berdasarkan hasil publikasi dari BPS (Badan Pusat Statistik). Konsep pengangguran menurut (BPS, 2014) adalah Penduduk 15 tahun ke atas yang tidak bekerja dan melakukan kegiatan sedang mencari pekerjaan, dan atau sedang mempersiapkan usaha, atau merasa tidak mungkin mendapatkan pekerjaan, atau sudah mendapat pekerjaan tetapi belum mulai bekerja selama seminggu yang lalu. Maka atribut yang dapat digunakan pada Table 4.1 Pemilihan Atribut untuk karakteristik tingkat Pengangguran

Tabel 4. 1 Tabel Pemilihan Fitur untuk Karakteristik Tingkat Pengangguran

No	Atribut	Tipe Data	Klasifikasi Data
1.	Sudah Bekerja dalam 1 minggu	Nominal	1=Iya 2=Tidak
2.	Sedang Mencari Kerja	Nominal	1=Iya 2=Tidak
3.	Jam Kerja dalam 1 minggu	Nominal	1 jika = 0 Jam 2 jika 1 - 35 Jam 3 jika > 35 Jam
4.	Status Sementara Tidak bekerja	Nominal	1=Iya 2=Tidak
5.	Mencari Usaha Baru	Nominal	1=Iya, 2=Tidak
6.	Alasan Tidak Mencari Kerja	Nominal	1=Putus Asa:Merasa tdk mungkin mendapat pekerjaan 2=Sudah diterima bekerja, tapi belum mulai bekerja 3=Sedang bersekolah 4=Mengurus rumah tangga 5=Sudah mempunyai pekerjaan/usaha 6=Sudah merasa cukup 7=Tidak mampu melakukan pekerjaan 8=Lainnya

#### 4.1.2 Fitur Untuk Karakteristik Pengangguran Berdasarkan Wilayah

Tabel 4. 2 Tabel Fitur Untuk Karakteristik Pengangguran Berdasarkan Wilayah

No	Atribut	Tipe Data	Klasifikasi Data
1.	Klasifikasi Wilayah	Nominal	1=Perkotaan 2=Pedesaan

Pada Tabel 4.2 Tabel Pemilihan Fitur Untuk Karakteristik Pengangguran berdasarkan wilayah Perkotaan dan Pedesaan diatas merupakan pemilihan fitur atau atribut data yang bertujuan untuk menghasilkan pengelompokan data yang menggambarkan karakteristik pengangguran di perkotaan dan pedesaan, sehingga dapat diketahui pada daerah manakah jumlah prosentase pengangguran terjadi pada klasifikasi wilayah tersebut.

#### 4.1.3 Fitur Pengangguran Berdasarkan Pendidikan Terakhir

Pada Tabel 4.3 Tabel Pemilihan Fitur Untuk Karakteristik Pengangguran Berdasarkan Pendidikan merupakan pemilihan fitur atau atribut data yang bertujuan untuk menghasilkan pengelompokan data yang menggambarkan karakteristik pengangguran berdasarkan pendidikan terakhirnya, sehingga dapat diketahui pendidikan dari penganggur mempunyai pendidikan tinggi atau berpendidikan rendah.

Tabel 4. 3 Tabel Pemilihan Fitur Pengangguran Berdasarkan Pendidikan

No	Atribut	Tipe Data	Klasifikasi Data
1.	Pendidikan Terakhir	Ordinal	1=Tdk/blm pernah sekolah 2.=Tdk/blm tamat SD 3=SD/Ibtidaiyah 4=Paket A 5=SMP/Tsanawiyah 6=SMP Kejuruan 7=Paket B 8=SMA/Aliyah 9=SMK 10= Paket C 11=Diploma I/II 12=Diploma III 13=Diploma IV/Universitas 14=S2/S3

#### 4.1.4 Fitur Pengangguran Berdasarkan Umur

Pada Tabel 4.4 Tabel Pemilihan Fitur Untuk Karakteristik Pengangguran Berdasarkan Umur merupakan pemilihan fitur atau atribut data yang bertujuan untuk mengelompokkan karakteristik pengangguran yang berdasarkan umur, sehingga dapat diketahui kisaran berapakah umur penganggur dan yang bukan penganggur

Tabel 4. 4 Tabel Pemilihan Fitur Untuk Karakteristik Pengangguran Berdasarkan Umur

No	Atribut	Tipe Data	Klasifikasi Data
1.	Umur	Nominal	0 =<15 Tahun 1 =15 – 24 Tahun 2 = 24 – 44 Tahun 3 = 45 – 64 Tahun 4 = > 64 Tahun

#### 4.1.5 Fitur pengangguran Berdasarkan jenis kelamin

Pada Tabel 4.5 Pemilihan Fitur Untuk Karakteristik Penganggur Berdasarkan jenis kelamin merupakan pemilihan data yang bertujuan untuk mengelompokkan karakteristik penganggur berdasarkan jenis kelamin, sehingga dapat diketahui apakah jenis kelamin yang dominan .

Tabel 4. 5 Pemilihan Fitur Untuk Karakteristik Penganggur Berdasarkan jenis kelamin

No	Atribut	Tipe Data	Klasifikasi Data
1.	Jenis Kelamin	Nominal	1=Laki - laki 2=Perempuan

#### 4.1.6 Fitur Pengangguran Berdasarkan Kompetensi Pendidikan

Pada Tabel 4.6 Tabel Pemilihan Fitur Untuk Karakteristik Penganggur Berdasarkan Kompetensi Pendidikan merupakan pemilihan fitur yang bertujuan

untuk mengelompokan karakteristik pengangur berdasarkan kompetensi pendidikan yang dikuasai, sehingga dapat diketahui pengangur mempunyai kompetensi pendidikan apa saja.

Tabel 4. 6 Tabel Fitur Untuk Karakteristik Pengangur Berdasarkan Kompetensi Pendidikan

No	Atribut	Tipe Data	Klasifikasi Data
1.	Jurusan	Nominal	11=Ilmu Pengetahuan Alam (IPA) 12=Ilmu Sosial dan Politik (IPS) 13=Kedokteran dan Kesehatan 14=Kehutanan 15=Kesenian dan Seni 16=Komunikasi Masa dan Dokumentasi 17=Matematika dan Ilmu Komputer 18=Pelayanan Jasa 19=Perikanan 20=Pertanian 21=Pertukangan, Kerajinan dan Industri 22=Peternakan 23=Psikologi 24=Teknik/Teknologi 25. Transportasi dan Komunikasi 26. Kependidikan dan Keguruan 27. Lainnya 37. SMU IPA 38. SMU IPS 39. SMU Fisika 40. SMU Biologi 41. SMU Bahasa dan Sastra 42. STM Listrik,Mesin,Grafika,Tekstil ,Penerbangan dan Lainny 43= SMEA Ekonomi, Akuntansi dan Tata Negara 44=SMKK Tata Boga, Tata Busana dan Tata Graha 45=SMK Perhotelan, Pariwisata, Perawat 50 = Lain - Lain

## 4.2 Normalisasi Dan Tranformasi Data

Pada Tahap normalisasi data dimana pada tahap ini data sudah dilakukan pemilihan atribut, data yang diperoleh dari BPS masih dalam bentuk Raw data atau data mentah dalam bentuk format \*.sav sehingga masih memerlukan proses lagi yaitu export ke dalam bentuk Microsoft Excell untuk dijadikan ke dalam format \*.csv gunanya untuk dilakukan proses pembedulan data, data yang diperoleh ada beberapa bagian dari poin – poin yang tidak diisi karena poin tersebut tidak relevan dengan pertanyaan sebelumnya maka biasanya jika kosong diisi dengan tanda baca “.” Kemudian untuk dapat dilakukan pengolahan data maka dilakukan pembedulan data dengan mengganti dengan nilai “0”. Pada Gambar 1 merupakan contoh dataset dari data sakernas sebelum dilakukan Normalisasi dan pembedulan data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Table: Data List																	
2	R1005	HUR	IK	UMUR	STATK	R5P1A	R5P1B	R5P1C	R5P1D1	R5P2B	R5P3	R5P4	R5P5	R5P6	R5P7	R5P8A	R5P2D.5	R5P25
3	2	1	2	37	4	2.		2.		1.		2	2	5	2	21.		
4	2	5	1	29	2	5.		2.		1.		2	2	5	2	20.		6
5	2	7	2	24	2	5.		2.		1.		2	2	4	2	35.		6
6	?	1	1	49	?	3.		?		1.		?	?	5	?	56.		
7	2	1	1	71	2	3.		2.		1.		2	2	5	2	56.		3
8	2	2	2	35	2	2.		2.		1.		2	2	4	2	20.		
9	?	3	?	13	1	3.		?		?	?	?	?	3	?			
10	2	1	1	61	2	2.		2.		1.		2	2	5	2	56.		3
11	2	2	2	62	2	3.		2.		1.		2	2	4	2	28.		
12	2	?	1	12	1	2.		2.		2	2	2	2	3	2			
13	?	1	1	40	?	3.		?		1.		?	?	5	?	35.		
14	2	2	2	37	2	3.		2.		1.		2	2	4	2	35.		
15	2	?	2	15	1	?		2.		2	2	2	2	3	2			
16	2	5	1	59	2	2.		2.		1.		2	2	5	2	21.		
17	?	6	?	67	?	1.		?		1.		?	?	4	?	78.		
18	2	1	1	73	2	1.		2.		1.		2	2	5	2	14.		
19	2	2	2	59	2	?		2.		1.		2	2	4	2	20.		
20	?	1	1	35	?	3.		?		1.		?	?	5	?	40.		5
21	2	2	2	33	2	3.		2.		3.		2	2	4	2	14.		

Gambar 4. 1 Print screen dataset dari Data Sakernas sebelum dilakukan Normalisasi dan Pembedulan data

Selanjutnya setelah dataset sudah dilakukan normalisasi data proses selanjutnya adalah dilakukan proses tranformasi data, pada proses ini dimaksudkan untuk memberikan jangkauan nilai rentan nilai tertentu, misalnya pada atribut umur ,jika berumur 0 – 14 tahun maka diberikan kode 1, umur 15 – 24 diberikan kode 2, umur 25 – 64 diberikan kode 3, dan diatas 65 diberikan kode 4, pada tabel

4.1 ,4.2, 4.3, 4.4 ,4.5 dan 4.6 adalah tranformasi data pada masing – masing pemilihan karakteristik pengangguran.

### 4.3 Pembersihan Data / Data Cleaning

Pada proses pembersihan data ini dilakukan karena proses dimana dataset yang digunakan masih campur dan memerlukan penentuan data yang memenuhi konsep Penelitian yaitu konsep Angkatan Kerja ,proses yang dilakukan ialah menghapus data yang tidak berkaitan, adapun proses pembersihan data yang tidak digunakan dalam penelitian yaitu terdapat dalam Tabel 4.7.

Tabel 4. 7 Proses Pembersihan Data

No.	Keterangan	Data tereduksi	Sisa Dataset
1.	Dataset Awal (Data Sakernas)	0	13266
2.	Menghapus umur responden < 15 tahun	1.358	11.908
3.	Menghapus kegiatan seminggu lalu = mengurus rumah tangga, dan pelajar dan jam kerja = 0 Jam	3.240	8.666
4.	Menghapus responden yang tidak bekerja dengan alasan mengurus rumah tangga dan sebagai pelajar	12	8.654

Pada tabel 4.7 menunjukkan dataset awal sebelum dilakukan pembersihan data sebanyak 13.266 setelah dilakukan pembersihan data, dataset memenuhi konsep Angkatan Kerja menjadi 8.654 sample rumah tangga. Setelah melalui tahap pembersihan data, dataset dalam bentuk file \*.CSV kemudian di Export ke dalam bentuk file \*.txt untuk dilakukan proses input data kedalam aplikasi Matlab. Pada Gambar 2 merupakan dataset yang sudah melalui proses normalisasi data, transformasi data dan pembersihan data kemudian sudah diekspor kedalam file \*.txt dan siap untuk diinput kedalam perintah Matlab.

File	Edit	Format	View	Help						
1		0	2	2	5	2				
1		0	2	2	5	2				
1		0	2	2	4	3				
1		0	2	2	5	3				
1		0	2	2	5	3				
1		0	2	2	4	2				
1		0	2	2	5	3				
1		0	2	2	4	2				
1		0	2	2	5	3				
1		0	2	2	4	3				
1		0	2	2	5	2				

Gambar 4. 2 Dataset dalam bentuk format teks \*.txt

#### 4.4 Proses Clustering Data

Pada tahap ini dilakukan setelah melalui tahap Preprocessing data atau penyiapan data, kemudian proses selanjutnya ialah tahap pengklasteran data, tahap pengklasteran ini dilakukan melalui 2 tahap yaitu :

1. Tahap Average Linkage Clustering

Proses pengklasteran data yang bertujuan pengelompokan awal untuk penentuan nilai centroid awal pada Algoritma K-Means.

2. K-means Clustering

Proses pengklasteran data yang dilakukan setelah Average Linkage clustering, dimana nilai centroid awal K-Means didapat dari nilai rata – rata pada masing - masing anggota cluster hasil dari pengelompokan sebelumnya (Average Linkage Clustering).

##### 4.4.1 Implementasi metode Average Linkage dengan menggunakan Matlab

Pada tahap implementasi ini, file input sudah berupa file \*.txt yang merupakan hasil export dari file Microsoft Excell dalam bentuk \*.csv, langkah selanjutnya adalah melakukan load file dan menentukan jumlah cluster yang akan

dihasilkan, pada Gambar 4.3 penentuan jumlah cluster sebanyak 2 ( $c=2$ ). Nilai dari “c” dan dataset yang sudah dimasukkan kedalam variable “x” diinputkan kedalam fungsi Hierarchi\_K\_Means untuk dilakukan tahap clustering menggunakan Average Linkage, seperti pada Gambar 4.3 :

```

1 - clear;
2 - clc;
3 - x=load('pengangguran.txt');
4 - x=x(:,1:6);
5 - c=2;
6 - y=Hierarchi_K_Means(x,c);

```

Gambar 4. 3 Perintah load dataset pada Matlab

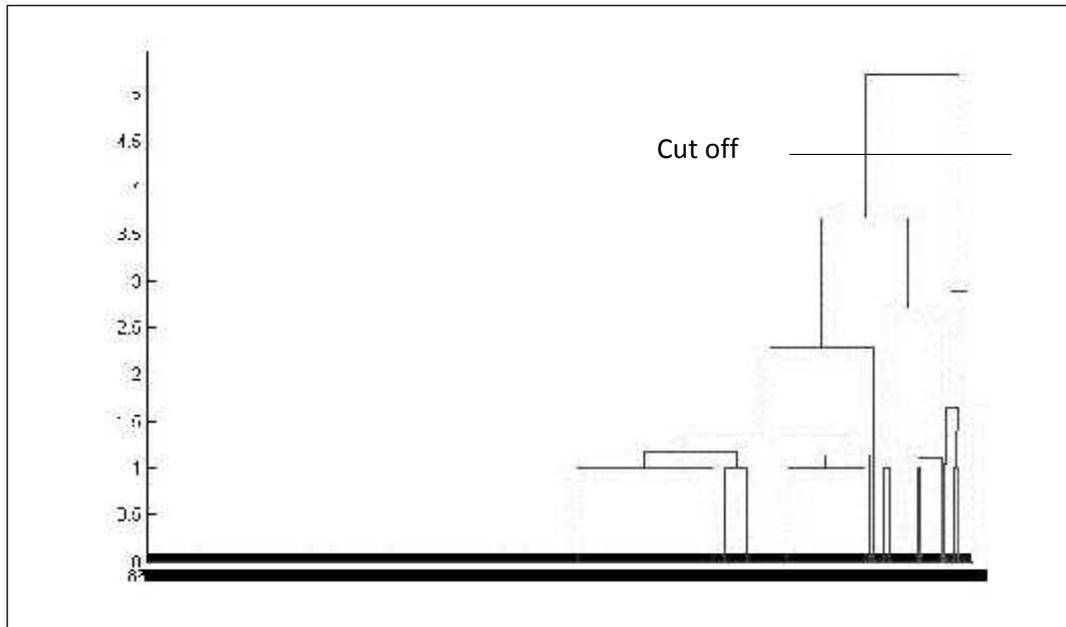
```

function [waktukomputasi, pasatbaca, zpphen, ukur1, ukur2, tamol1, tamol2, g, h]=Hierarchi_K_Means(x,c)
2 % x adalah matrix metadata
3 % c adalah jumlah cluster
4 starttime=cputime;
5 % untuk membaca jarak antar data
6 y=pdist(x,'euclidean');
7 % Hierarchical clustering linkage, complete, average dan centroid
8 z=linkage(y,'average');
9 % untuk mengetahui pada cluster manakah sebuah data tersebut berada [index
10 % cluster]
11 t=cluster(z,'maxclust',c);
12 dendrogram(x,t);
13 % untuk menampilkan data yang masuk kedalam cluster ke 1,2,3
14 aa=find(t==1);
15 bb=find(t==2);
16 data1=x(aa,:);
17 data2=x(bb,:);

```

Gambar 4. 4 Perintah Average Linkage Clustering dengan menggunakan Matlab

Kemudian pada fungsi Hierarchi\_K\_Means Pada Gambar 4.4 adalah mengukur ketidakmiripan antar cluster dengan menggunakan Euclidean ditunjukkan pada baris ke – 6 yaitu  $y=pdist(x,'Euclidean')$  dan perintah Average Linkage clustering dengan menggunakan Average linkage ditunjukkan pada baris ke-8 yaitu  $z=linkage(y,'average')$ . Kemudian hasil dari pengelompokan tersebut akan menghasilkan grafik dendogram pada Gambar 4.5



Gambar 4. 5 Grafik Dendogram average linkage

Untuk menghasilkan jumlah cluster sebanyak 2 maka akan dilakukan Cut off pada penggabungan cluster paling atas yang ditunjukkan pada Gambar 4.5 Grafik dendogram, kemudian pada perintah yang terletak pada baris 16 dan 17 pada Gambar 4.4 hasil dari 2 buah pengelompokan yaitu cluster 1 yang dimasukkan pada variable “data1”, dan cluster 2 yang dimasukkan pada variable “data2”

#### 4.4.2 Implementasi K-Means clustering dengan menggunakan Matlab

Pada tahap ini merupakan tahap lanjutan dari Average Linkage clustering yaitu setelah hasil dari pengelompokan diperoleh pada “data1” untuk cluster 1 dan “data2” untuk cluster 2 langkah selanjutnya ialah clustering dengan menggunakan Algoritma K-Means, dari variable “data1” dan “data2” hasil clustering sebelumnya kemudian dilakukan rata – rata untuk mendapatkan nilai centroid awal yang akan dimasukkan pada algoritma k-means, pada Gambar 4.6 merupakan perintah pada Matlab untuk pembangkitan nilai awal centroid untuk K-means :

```

30 - mean1=mean(data1)
31 - mean2=mean(data2)
32 - pusatbaru=[mean1;mean2];
33 - [idx, ctrs]=kmeans(x,c,'start',pusatbaru);
34 - j=[x idx];
35
36 - ff=find(idx==1);
37 - gg=find(idx==2);
38
39 %untuk menampilkan data yang masuk ke dalam setiap cluster
40 %yang kemudian di cluster
41 - tampil1=x(ff,:);
42 - csvwrite('cluster1.dat',tampil1);
43 - type cluster1.dat;
44 - tampil2=x(gg,:);
45 - csvwrite('cluster2.dat',tampil2);
46 - type cluster2.dat;

```

Gambar 4. 6 Implementasi K-means dan penentuan centroid awal K-Means

Pada Gambar 4.6 “mean1” merupakan centroid awal pertama dari nilai rata – rata pada cluster 1 dan “mean2” merupakan centroid awal kedua dari nilai rata – rata pada cluster 2 kemudian masing – masing nilai centroid dimasukan kedalam variable “pusatbaru” kemudian dimasukan pada fungsi K-means clustering ditunjukkan pada baris ke-33 yaitu [idx ctrs]=kmeans(x,c,'start',pusatbaru) ,kemudian output atau hasil dari tahap k-means clustering ini dimasukan ke dalam variable “tampil1” untuk hasil dari cluster 1 dan tampil2 untuk hasil dari cluster 2.

#### 4.5 Evaluasi Dan Validasi Average Linkage dan K-Means clustering

Pada tahap pengujian validasi cluster ini adalah tahap penting dalam penelitian, sebab dari uji validasi ini akan bertujuan selain untuk menentukan jumlah cluster yang ideal untuk proses clustering dalam analisa faktor – faktor pengangguran tapi juga untuk membandingkan metoda gabungan Average Linkage clustering dan K-Means clustering dengan metode K-Means clustering murni, apakah dengan penggabungan dua metode Average Linkage dan K-Means dapat menghasilkan nilai validasi Silhoutte lebih baik dari pada hasil validasi dari Algoritma K-Means saja.

```

50 % Validasi Cluster
51 - SI = silhouette(x,idx,'Euclidean')
52 - SI_cluster = [mean(SI(idx==1)) mean(SI(idx==2))]
53 - SI_semua = mean(SI_cluster)
54 - jarak='euclidean';

```

Gambar 4. 7 Implementasi Uji Validasi dengan metode Silhoutte Index dengan menggunakan Matlab

Pada Gambar 4.7 merupakan Script perintah pada Matlab untuk melakukan tahap uji coba validasi cluster dengan Silhoutte Index menggunakan pengukuran jarak Euclidean Distance. Nilai *Silhouette Index* mendekati 1 menunjukkan bahwa data tersebut tepat berada pada klaster tersebut. SI bernilai 0 atau mendekati 0 maka posisi data berada pada perbatasan dua klaster, sedangkan nilai negatif menandakan jarak rata-rata antar objek jauh. Artinya Semakin dekat SI kepada 1, maka semakin baik pengelompokkan data Sebaliknya, semakin dekat SI kepada -1, maka semakin buruk pengelompokkan datanya. Proses uji validasi ini dilakukan melalui beberapa tahap yaitu :

1. Lakukan clustering untuk menghasilkan 2 sampai 8 cluster.
2. Lakukan pengukuran nilai Silhoutte index dari cluster 2 sampai cluster 8.
3. Buat Grafik nilai Silhoutte dari cluster 2 sampai cluster 8.
4. Nilai paling besar mendekati 1 adalah nilai optimal cluster dan jumlah cluster yang paling ideal untuk metoda clustering.

Pada proses validasi cluster ini dilakukan dengan menguji hasil clustering dari karakteristik pengangguran dan faktor – faktornya.

#### 4.5.1 Uji Coba Validasi Cluster Tingkat Pengangguran

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, kemudian membandingkan hasil validasi dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni, serta mengelompokkan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang

mempunyai karakteristik pengangguran, hasil pengelompokan dari 2 sampai 8 cluster pada Tabel 4.8

Tabel 4. 8 Hasil Uji Validasi Silhoutte Index dataset karakteristik Pengangguran

Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK Kmeans	0.7012	0.6755	0.6978	0.7563	0.7728	0.7266	0.7787	0.78
K-Means	0.3725	0.5331	0.6978	0.6817	0.5704	0.6523	0.6829	0.69

Hasil uji validasi karakteristik pengangguran dari Tabel 4.8 nilai maksimum pada algoritma ALK (Average Linkage K-means) pada cluster ke – 8 sebesar 0.7787, hal itu menunjukkan bahwa cluster optimal dari algoritma ALK terletak pada cluster 8. Apabila dirata – rata keduanya menghasilkan ALK = 0.729843 dan K-Means = 0.598671, hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari K-Means clustering murni, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik, jadi dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 21,91 %

Tabel 4.9 Nilai simpangan baku karakteristik pengangguran

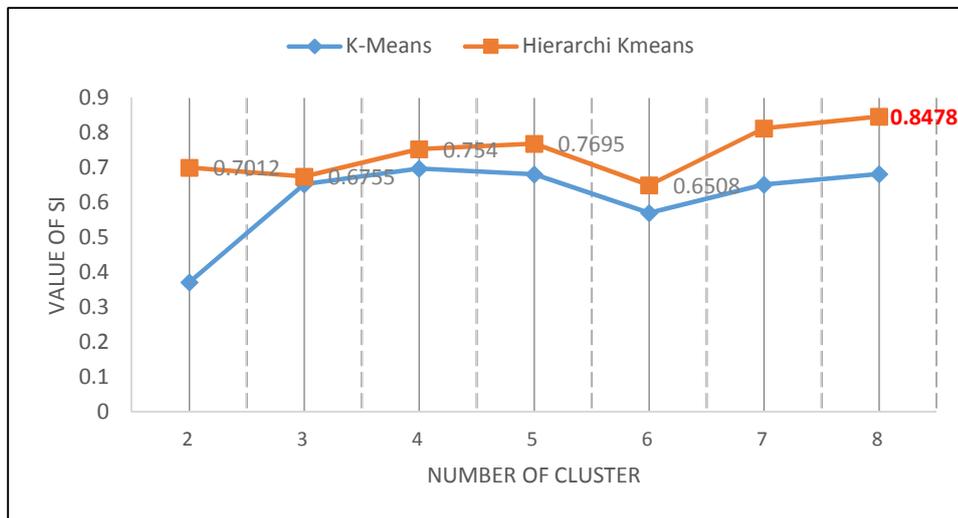
Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.73	0.040297
K-Means	0.599	0.117575

Setelah diketahui nilai rata – rata dari masing – masing pengujian dan cluster yang optimal dari 2 sampai 8 cluster, kemudian menghitung simpangan baku atau standar deviasi untuk mengetahui keragaman(variasi) data dari uji validasi silhouette index. Pada Tabel 4.9 nilai standar deviasi dari gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian ALK dan K-means lebih stabil hasil clusteringnya.

Tabel 4. 9 Perbandingan waktu Komputasi dengan dataset karakteristik Pengangguran

Jumlah Cluster	2	3	4	5	6	7	8
ALK Kmeans	59.20	58.17	64.23	56.12	60.8	56.34	57.34
K-Means	4.18	4.29	4.62	4.98	5.60	5.68	6.04

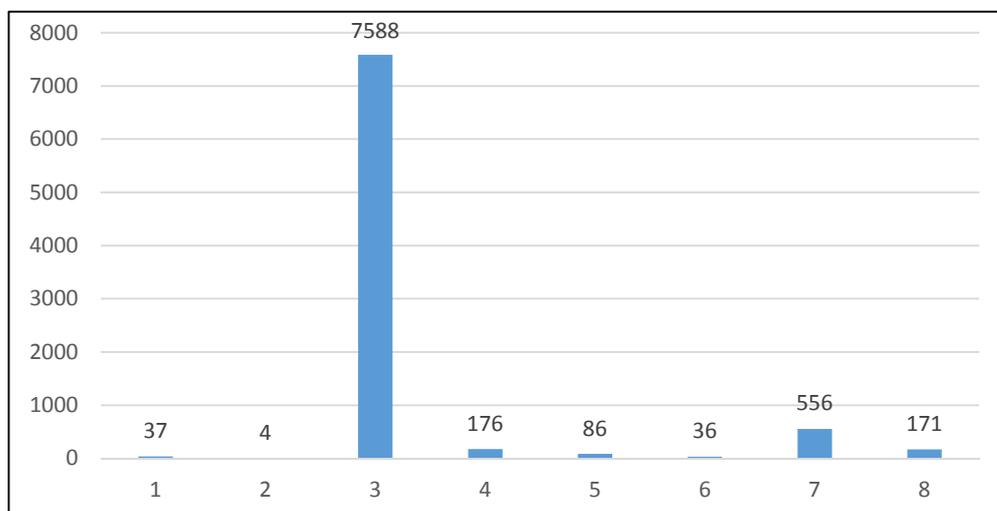
Pada Tabel 4.10 merupakan nilai perbandingan waktu komputasi dengan data set fitur utama pengangguran pada kedua algoritma , dimana algoritma Average Linkage dan K-Means mempunyai komputasi lebih lama dari K-Means biasa.



Gambar 4. 8 Hasil Uji Coba validasi Silhoutte index attribute karakteristik

Hasil dari Tabel 4.8 ditampilkan dalam grafik pada Gambar 4.8 yang menunjukkan grafik gabungan Average Linkage dan K-Means lebih tinggi dari pada grafik K-Means biasa.

Pada Gambar 4.8. Merupakan jumlah anggota pada masing – masing cluster dari hasil pengelompokan dan penentuan jumlah cluster optimal sebanyak 8 cluster.



Gambar 4. 9 Diagram Anggota Cluster Karakteristik Pengangguran

Pada grafik pada gambar 4.9 Jumlah anggota cluster Mempunyai karakteristik masing – masing yaitu ditampilkan pada Tabel 4.11 :

Tabel 4. 10 Tabel Analisa Hasil Pengelompokan Karakteristik Pengangguran

Cluster	Jml Data	%	Karakteristik Pengangguran
1	37	0.43%	Kelompok Angkatan Kerja yang mempunyai karakteristik pengangguran terbuka, seluruh responden tidak bekerja, tidak mencari pekerjaan karena putus asa
2	4	0.046%	Kelompok Angkatan Kerja yang mempunyai karakteristik pengangguran terbuka seluruh sample tidak bekerja karena belum mulai pekerjaan
3	5.171	59.75%	Kelompok Angkatan Kerja yang mempunyai karakteristik sebagai pekerja Normal
	2.417	27.93%	Kelompok Angkatan Kerja yang mempunyai karakteristik pekerja paruh waktu atau pekerja yang jam kerja dibawah 35 jam
4	176	2 %	mempunyai karakteristik bervariasi yaitu pekerja Normal sebanyak 79 dan setengah pengangguran sebanyak 63, dan pekerja paruh waktu sebanyak 34
5	86	0,99 %	Kelompok Angkatan Kerja mempunyai karakteristik penganggur terbuka dengan alasan lain- lain atau tidak mampu melakukan pekerjaan

Tabel 4.11 Tabel Lanjutan

Cluster	Jml Data	%	Karakteristik Pengangguran
6.	36	0.41 %	Kelompok Angkatan Kerja mempunyai karakteristik penganggur terbuka dengan alasan sudah mempunyai pekerjaan tapi sementara tidak bekerja
7.	556	6.42 %	Kelompok Angkatan Kerja mempunyai karakteristik penganggur terbuka tidak mampu melakukan pekerjaan
8.	45	0.52 %	Kelompok Angkatan Kerja mempunyai karakteristik pekerja Normal yang mempunyai jam kerja > 35 jam selama seminggu
	126	1.46 %	Kelompok Angkatan Kerja mempunyai karakteristik pekerja paruh waktu yang mempunyai jam kerja < 35 jam selama seminggu

#### 4.5.2 Pengklasteran dan Uji Coba Validasi Cluster Karakteristik Pengangguran dengan wilayah

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, membandingkan dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni serta mengelompokkan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang mempunyai karakteristik pengangguran berdasar wilayah, hasil pengelompokan dari 2 sampai 8 cluster dicatat pada Tabel 4.12.

Tabel 4. 11 Hasil Uji Validasi Silhoutte Index dataset Pengangguran dengan wilayah

Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK Kmeans	0.661	0.619	0.621	0.657	0.6502	0.6014	0.5632	0.66
K-Means	0.335	0.324	0.482	0.560	0.5773	0.6141	0.539	0.61

Hasil uji validasi pengangguran berdasarkan wilayah dari tabel 4.12 nilai maksimum pada setiap pengujian menunjukkan pada algoritma ALK (Average

Linkage K-means) pada 2 cluster sebesar 0.6612, artinya bahwa cluster optimal dari algoritma ALK terletak pada cluster 2, dan apabila dirata – rata keduanya menghasilkan  $ALK = 0.624928571$  dan  $K\text{-Means} = 0.490528571$ , hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari K-Means clustering murni, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik maka dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 27,4 %.

Tabel 4. 12 Nilai Standar Deviasi Silhoutte Index data set wilayah

Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.624929	0.035131
K-Means	0.490529	0.116711

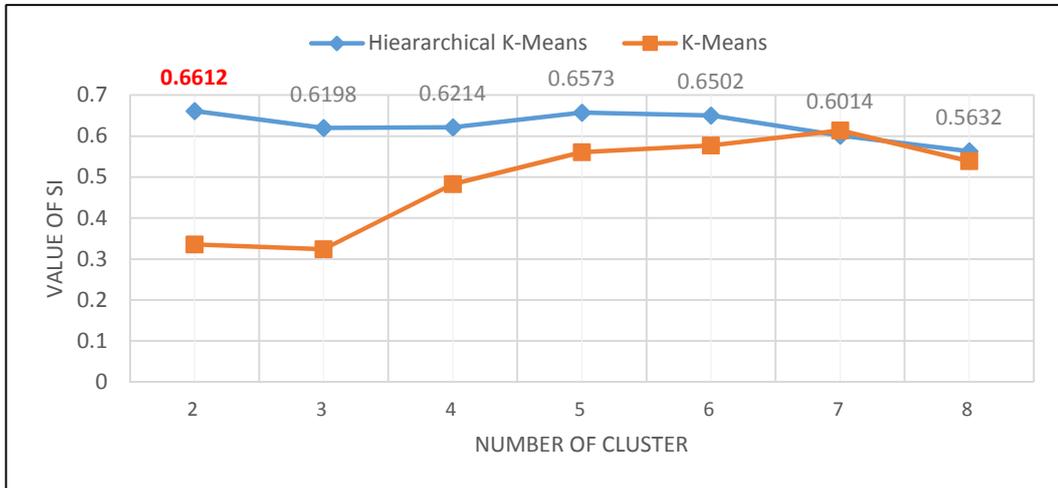
Pada Tabel 4.13 menunjukkan perhitungan nilai standar deviasi dari nilai Silhoutte index pada dataset wilayah, gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian ALK dan K-means lebih stabil hasil clusteringnya.

Tabel 4. 13 Perbandingan waktu komputasi data set wilayah

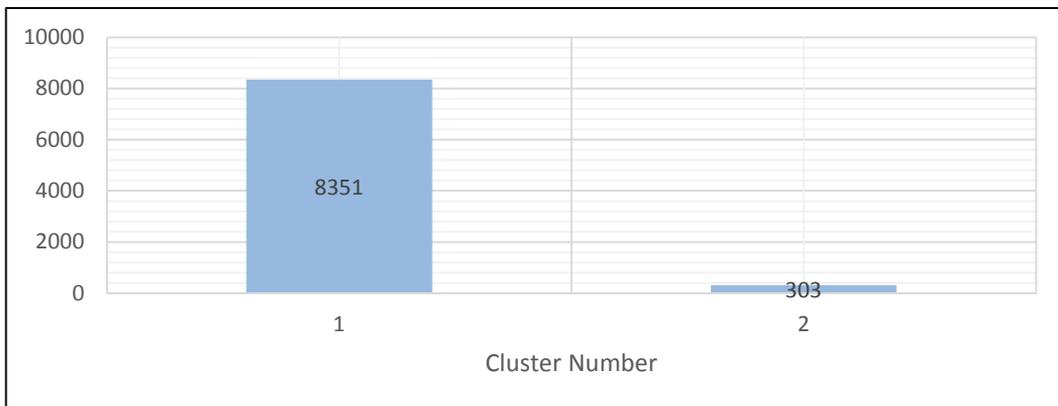
Jumlah Cluster	2	3	4	5	6	7	8
ALK K-Means	62.7344	62.0469	63.023	62.86	64.023	63.56	62.03
K-Means	4.0313	4.6406	4.9375	6.0156	6.1563	5.7031	5.7813

Pada Tabel 4.14 merupakan nilai perbandingan waktu komputasi dengan data set fitur wilayah dari kedua algoritma , dimana algoritma Average Linkage dan K-Means mempunyai komputasi lebih lama dari K-Means biasa.

Dari hasil validasi pada Tabel 4.13 ditampilkan dalam grafik pada Gambar 4.11. yang menunjukkan grafik gabungan Average Linkage dan K-Means lebih tinggi dari pada grafik K-Means biasa.



Gambar 4. 11 Hasil Uji Coba Silhoutte index Atribut Karakteristik Pengangguran dengan wilayah



Gambar 4. 10 Diagram Jumlah Anggota Cluster Berdasarkan Wilayah

Pada grafik pada gambar 4.11. Jumlah anggota cluster Mempunyai karakteristik masing – masing yaitu ditampilkan pada Tabel 4.15 :

Tabel 4. 14 Tabel Analisa Pengelompokan Pengangguran berdasarkan wilayah

Cluster	Jumlah data	%	Wilayah	Karakteristik Pengangguran
1	3.854		Pedesaan	Kelompok Angkatan Kerja yang mempunyai karakteristik pengangguran terbuka sebanyak 265 sample, Pekerja paruh waktu sebanyak 1.522 sample, Pekerja Normal sebanyak 2.067 sample
	3.476		Perkotaan	Kelompok Angkatan Kerja yang mempunyai karakteristik pengangguran terbuka 327 sample dengan alasan putus asa, Pekerja Normal sebanyak 3.149 sample
2	125		Perkotaan	Kelompok Angkatan Kerja yang mempunyai karakteristik penganggur terbuka sebanyak 81 , Setengah Penganggur sebanyak 30, Pekerja Paruh Waktu sebanyak 14
	99		Pedesaan	Kelompok Angkatan Kerja yang mempunyai karakteristik penganggur terbuka sebanyak 46 sample, Setengah Penganggur 33 sample, Pekerja Paruh Waktu sebanyak 20 sample

#### 4.5.3 Uji Coba Validasi Cluster Pengangguran Berdasarkan Pendidikan Terakhir

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, membandingkan dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni serta mengelompokan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang mempunyai karakteristik pengangguran dengan pendidikan terakhir, hasil pengelompokan dari 2 sampai 8 cluster dicatat pada Tabel 4.16

Tabel 4. 15 Hasil Uji Validasi Silhoutte Index Pengangguran dengan pendidikan terakhir

Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK								0.6136
Kmeans	0.6136	0.4881	0.5226	0.5111	0.5943	0.571	0.5867	
K-Means	0.6136	0.4644	0.4977	0.469	0.5487	0.4657	0.3966	0.6136

Hasil uji validasi pengangguran dari Tabel 4.16 nilai maksimum pada setiap pengujian menunjukkan pada algoritma ALK (Average Linkage K-means) pada 2 cluster sebesar 0.6136, hal itu menunjukkan bahwa cluster optimal dari algoritma ALK terletak pada cluster 2, dan apabila dirata – rata nilai SI keduanya menghasilkan ALK = 0.555342857 dan K-Means = 0.493671429, hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari K-Means clustering murni, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik maka dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 12,49 %.

Tabel 4. 16 Nilai Standar Deviasi Silhoutte Index Data Set Pendidikan Terakhir

Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.555343	0.047769
K-Means	0.493671	0.06964

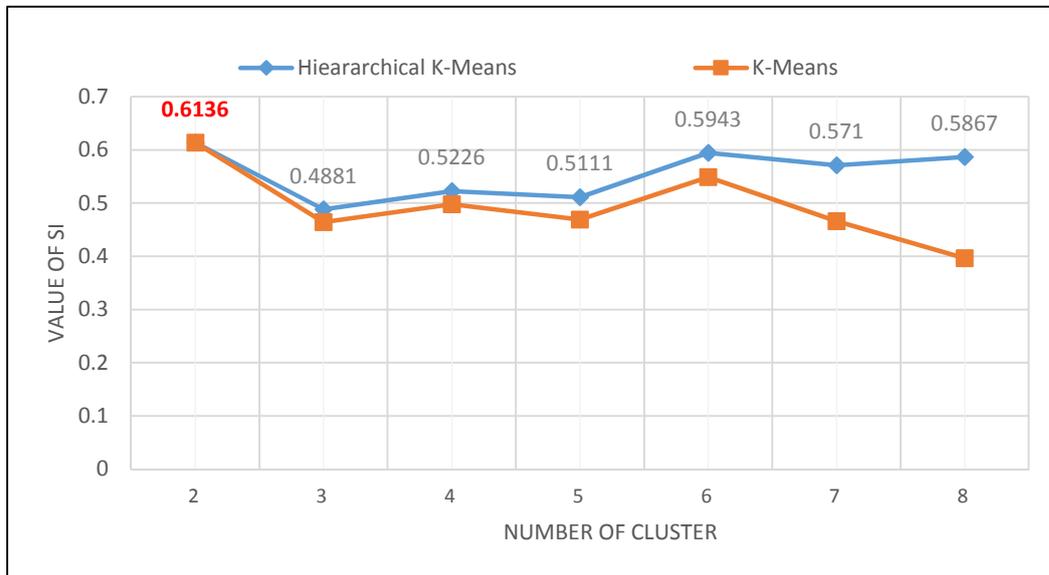
Pada Tabel 4.17 menunjukkan perhitungan nilai standar deviasi dari nilai Silhoutte index pada dataset pendidikan terakhir, gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian gabungan ALK dan K-means lebih stabil hasil clusteringnya.

Tabel 4.17 Perbandingan waktu komputasi data set pendidikan terakhir

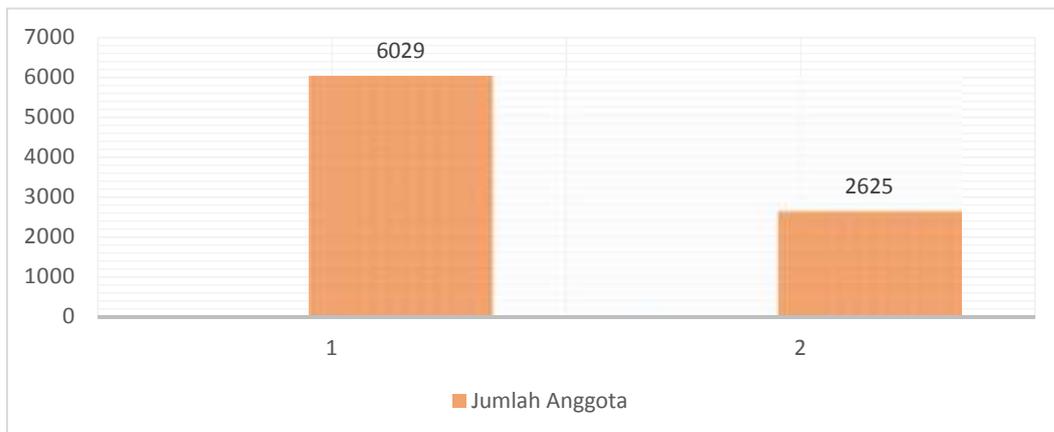
Jumlah Cluster	2	3	4	5	6	7	8
ALK K-Means	73.51	74.35	73.23	76.62	75.2	74.23	78.8
K-Means	4.6875	5.0781	5.5469	6.3438	5.75	5.875	6.2188

Pada Tabel 4.17 merupakan nilai perbandingan waktu komputasi dengan data set fitur pendidikan terakhir dari kedua algoritma , dimana algoritma Average Linkage dan K-Means mempunyai komputasi lebih lama dari K-Means biasa.

Hasil dari validasi pada Tabel 4.16 ditampilkan dalam grafik pada Gambar 4.12 yang menunjukkan grafik gabungan Average Linkage dan K-Means lebih tinggi dari pada grafik K-Means biasa.



Gambar 4. 12 Attribut Karakteristik Pengangguran dengan pendidikan Terakhir



Gambar 4. 13 Diagram Jumlah Anggota pada Karakteristik pengangguran berdasarkan pendidikan terakhir

Pada Gambar 4.13 menunjukkan Jumlah anggota cluster Mempunyai karakteristik masing – masing kemudian dianalisa dan ditampilkan pada Tabel 4.19 :

Tabel 4. 18 Tabel Analisa Hasil Pengelompokan Pengangguran berdasarkan Pendidikan

Cluster	Jumlah data	%	Jenis Pengangguran	Karakteristik Pengangguran berdasar pendidikan
1	3314		karakteristik Pekerja Normal	Berpendidikan rendah dari SD sampai SMP
	552		Pengangguran terbuka	Berpendidikan rendah dari SD sampai SMP
	38		Setengah Penganggur	berpendidikan maksimal SMK
	2.125		Pekerja paruh waktu	berpendidikan rendah maksimal SMP
2	1.981		karakteristik Pekerja Normal	mempunyai karakteristik pendidikan tinggi SMA/SMK tertinggi S2/S3,Pengangguran
	51		Pengangguran terbuka	berpendidikan SMA/SMK sebanyak 21, berpendidikan SMA Aliyah sebanyak 23, pendidikan D IV/S1 sebanyak 4, pendidikan DIII sebanyak 3 sample
	25		Setengah Penganggur	berpendidikan SMA/SMK sampai S1
	452		Pekerja paruh waktu	berpendidikan SMA/SMK sampai S1

#### 4.5.4 Uji Validasi Cluster Pengangguran berdasarkan Jenis Kelamin

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, membandingkan dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni serta mengelompokan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang mempunyai karakteristik pengangguran dengan jenis kelamin, hasil pengelompokan dari 2 sampai 8 cluster dicatat pada Tabel 4.19

Tabel 4. 19 Hasil Uji Validasi Silhoutte Index Pengangguran dengan Jenis Kelamin

Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK Kmeans	0.669	0.6264	0.6403	0.6746	0.637	0.6191	0.5794	0.6746
K-Means	0.6268	0.4587	0.5145	0.4547	0.6153	0.5324	0.5365	0.6268

Hasil uji validasi pengangguran dari Tabel 4.19 nilai maksimum pada setiap pengujian menunjukkan pada algoritma ALK (Average Linkage K-means) pada cluster 5 sebesar 0.6746, hal itu menunjukkan bahwa cluster optimal dari algoritma ALK terletak pada cluster 5, dan apabila maing – masing nilai SI dirata – rata keduanya menghasilkan ALK = 0.635114286 dan K-Means = 0.534128571, hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari pada nilai shilhoutte dari K-Means clustering murni, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik maka dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 18,9 %.

Tabel 4. 20 Nilai Standar Deviasi Silhoutte Index Data Set Jenis Kelamin

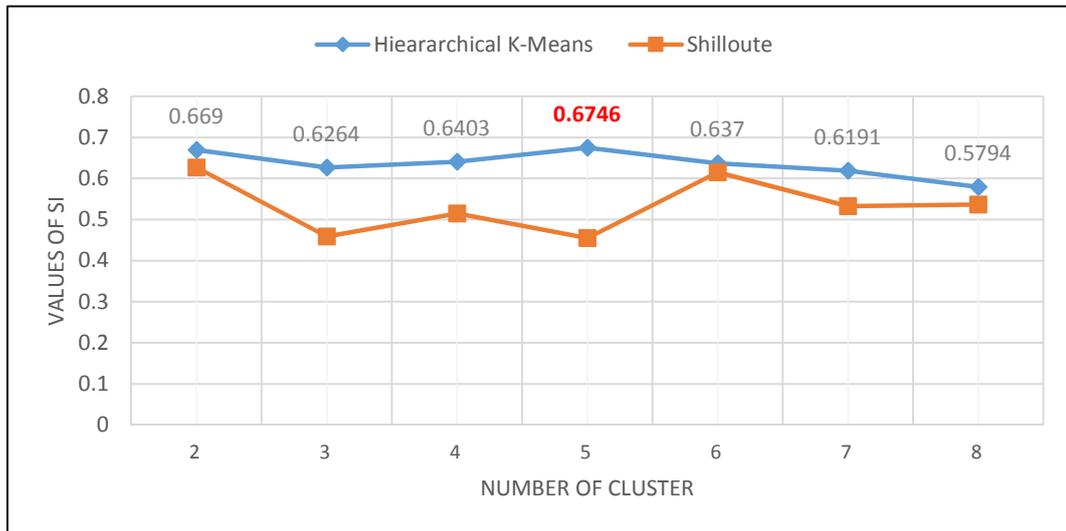
Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.635114	0.032079
K-Means	0.534129	0.067785

Pada Tabel 4.20 menunjukkan perhitungan nilai standar deviasi dari nilai Silhoutte index pada dataset jenis kelamin, gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian gabungan ALK dan K-means lebih stabil hasil clusteringnya.

Tabel 4. 21 Perbandingan waktu komputasi dengan data set jenis kelamin

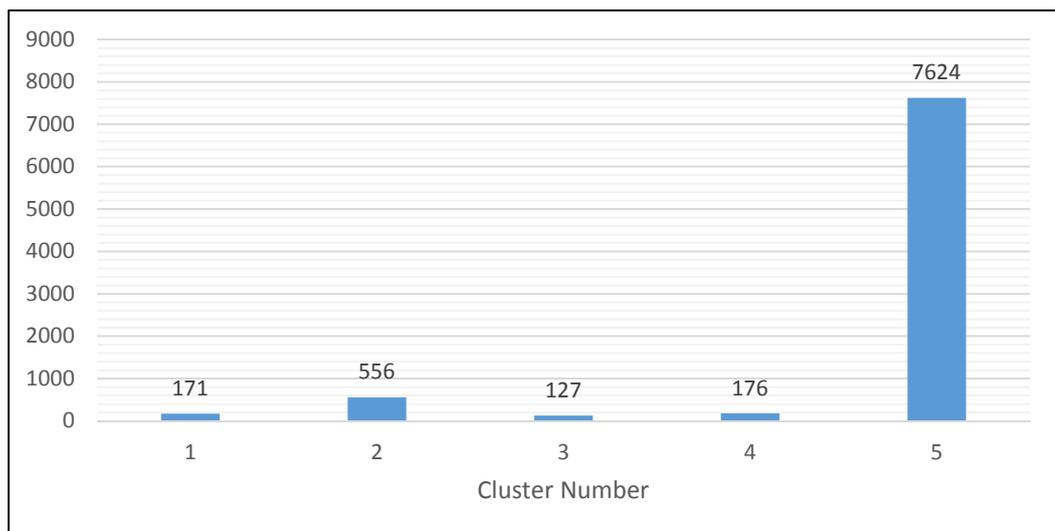
Jumlah Cluster	2	3	4	5	6	7	8
ALK K-Means	51.42	49.14	53.43	58.2	49.89	52.6	55.3
Shilloute	4.8	5.1406	5.4688	6.625	5.7344	7.0156	5.7969

Pada Tabel 4.21 merupakan nilai perbandingan waktu komputasi dengan data set fitur jenis kelamin dari kedua algoritma , dimana algoritma Average Linkage dan K-Means mempunyai komputasi lebih lama dari K-Means biasa.



Gambar 4. 14 Hasil Uji Coba Silhouette index karakteristik pengangguran dengan Jenis Kelamin

Gambar 4.14 yang menunjukkan grafik gabungan Average Linkage dan K-Means lebih tinggi dari pada grafik K-Means biasa.



Gambar 4. 15 Diagram Jumlah Anggota pada Karakteristik pengangguran berdasarkan Jenis Kelamin

Pada Gambar 4.15 menunjukkan Jumlah anggota cluster Mempunyai karakteristik masing – masing kemudian dianalisa dan ditampilkan pada Tabel 4.22

Tabel 4. 22 Tabel Analisa Pengelompokan Pengangguran berdasarkan Jenis Kelamin

Cluster	Jumlah data	%	Jenis Kelamin	Karakteristik Pengangguran berdasar pendidikan
1	118		Laki – laki	karakteristik Pekerja Normal 30 sample, Pekerja Paruh waktu 88 sample
	38		Perempuan	karakteristik Pekerja Normal 15 sample, Pekerja Paruh waktu 38 sample.
2	323		Laki – laki	mempunyai karakteristik Penganggur Terbuka sebanyak 323 sample
	233		Perempuan	mempunyai karakteristik Penganggur Terbuka sebanyak 233 sample
3	111		Laki-laki	mempunyai karakteristik penganggur terbuka sebanyak 111 sample
	17		Perempuan	mempunyai karakteristik penganggur terbuka sebanyak 17 sample
4	134		Laki-laki	mempunyai karakteristik Pekerja Normal sebanyak 59, pekerja paruh waktu 27, Pengangguran Setengah sebanyak 48,
	42		Perempuan	mempunyai karakteristik Pekerja Normal sebanyak perempuan 20 sample, pekerja paruh waktu perempuan 7 sample, Pengangguran Setengah sebanyak perempuan sebanyak 15 sample.
5	4377		Laki-laki	mempunyai karakteristik Pekerja Normal sebanyak 3.297, Pekerja Paruh Waktu sebanyak 1.046, Pengangguran Terbuka sebanyak 34
	3247		Perempuan	mempunyai karakteristik Pekerja Normal sebanyak, perempuan 1.874 sample, Pekerja Paruh Waktu sebanyak 1.371. Pengangguran Terbuka sebanyak perempuan 2 sample

#### 4.5.5 Pengklasteran Karakteristik Pengangguran berdasarkan Umur

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, membandingkan dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni serta mengelompokan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang mempunyai karakteristik pengangguran dengan umur, hasil pengelompokan dari 2 sampai 8 cluster dicatat pada Tabel 4.23.

Tabel 4. 23 Hasil Uji Validasi Silhoutte Index dataset Pengangguran dengan umur

Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK Kmeans	0.4757	0.4295	0.402	0.4306	0.4257	0.4213	0.3984	0.4757
K-Means	0.429	0.4127	0.3415	0.4097	0.405	0.3621	0.3766	0.429

Hasil uji validasi pengangguran dari Tabel 4.23 nilai maksimum pada setiap pengujian menunjukkan pada algoritma ALK (Average Linkage K-means) pada cluster 2 sebesar 0.4757, hal itu menunjukkan bahwa cluster optimal dari algoritma ALK terletak pada cluster 2, dan apabila masing – masing nilai SI dirata – rata keduanya menghasilkan ALK = 0.426171 dan K-Means = 0.390943, hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari K-Means clustering murni, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik maka dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 9.01 %.

Tabel 4. 24 Nilai Standar Deviasi dari Silhoutte Index dengan data set umur

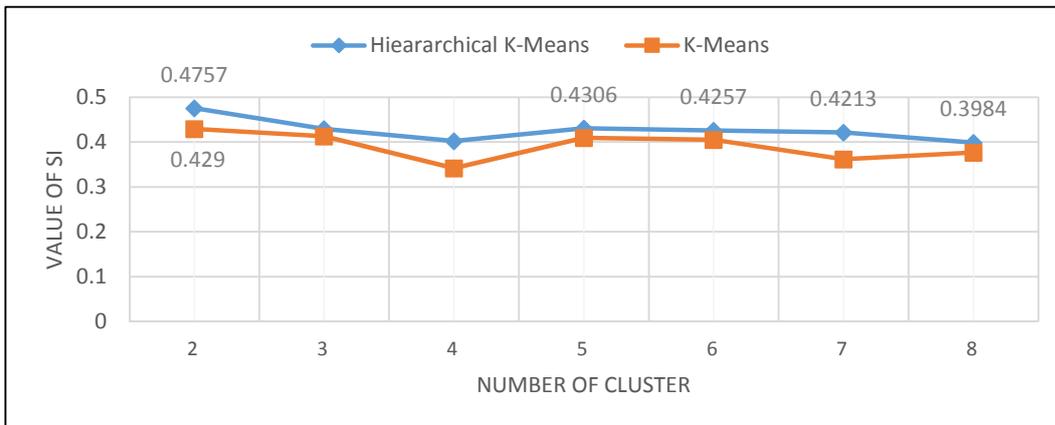
Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.426171	0.025375
K-Means	0.390943	0.0315

Pada Tabel 4.24 menunjukkan perhitungan nilai standar deviasi dari nilai Silhouette index pada dataset jenis kelamin, gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian gabungan ALK dan K-means lebih stabil hasil clusteringnya

Tabel 4. 25 Perbandingan waktu komputasi dengan data set umur

Jumlah Cluster	2	3	4	5	6	7	8
ALK K-Means	69.07	73.42	70.36	69.88	75.2	72.9	70.1
K-Means	4.9688	5.2656	5.5156	6.5781	5.7344	6	6.2344

Pada Tabel 4.25 merupakan nilai perbandingan waktu komputasi dengan data set fitur umur dari kedua algoritma , dimana algoritma Average Linkage dan K-Means mempunyai komputasi lebih lama dari K-Means biasa.



Gambar 4. 16 Hasil Uji Coba Silhouette index karakteristik pengangguran berdasarkan Umur

Jumlah anggota cluster Mempunyai karakteristik masing – masing kemudian dianalisa dan ditampilkan pada Tabel 4.26 :

Tabel 4. 26 Tabel Analisa Hasil Pengelompokan Pengguran berdasarkan umur

Cluster	Angkatan Kerja	Umur	Jumlah data
1	Pekerja Normal	15-24 tahun	545
		25-44 tahun	2555

Tabel 4.26 Tabel lanjutan

Cluster	Angkatan Kerja	Umur	Jumlah data
1	Pekerja Normal	45-64 tahun	4
	Penganggur Terbuka	15-24 tahun	132
		25-44 tahun	96
	Setengah Penganggur	15-24 tahun	21
		25-44 tahun	30
		45-64 tahun	6
	Pakerja Paruh Waktu	15-24 tahun	285
		25-44 tahun	943
		45-64 tahun	6
2	Pekerja Normal	45-64 tahun	1994
		65 tahun keatas	197
	Penganggur Terbuka	45-64 Tahun	164
		65 tahun keatas	327
	Setengah Pengangguran	65 tahun keatas	6
	Pekerja Paruh Waktu	65 tahun keatas	1350
	Setengah Pengangguran	65 tahun keatas	6
	Pekerja Paruh Waktu	65 tahun keatas	1350

#### 4.5.6 Pengklasteran Tingkat Pengangguran Berdasarkan Kompetensi Keahlian

Pada proses pengklasteran akan menghasilkan 2 sampai 8 cluster yang bertujuan untuk mengetahui jumlah cluster optimal dari semua pengujian, membandingkan dari Algoritma Gabungan Average Linkage dan K-Means dengan K-Means murni serta mengelompokan dari cluster yang sudah ditentukan untuk menghasilkan tingkat pengangguran dari faktor yang mempunyai karakteristik pengangguran dengan kompetensi keahlian, hasil pengelompokan dari 2 sampai 8 cluster dicatat pada Tabel 4.27.

Tabel 4. 27 Hasil Uji Validasi Silhoutte Index dataset karakteristik Pengangguran dengan Kompetensi Keahlian

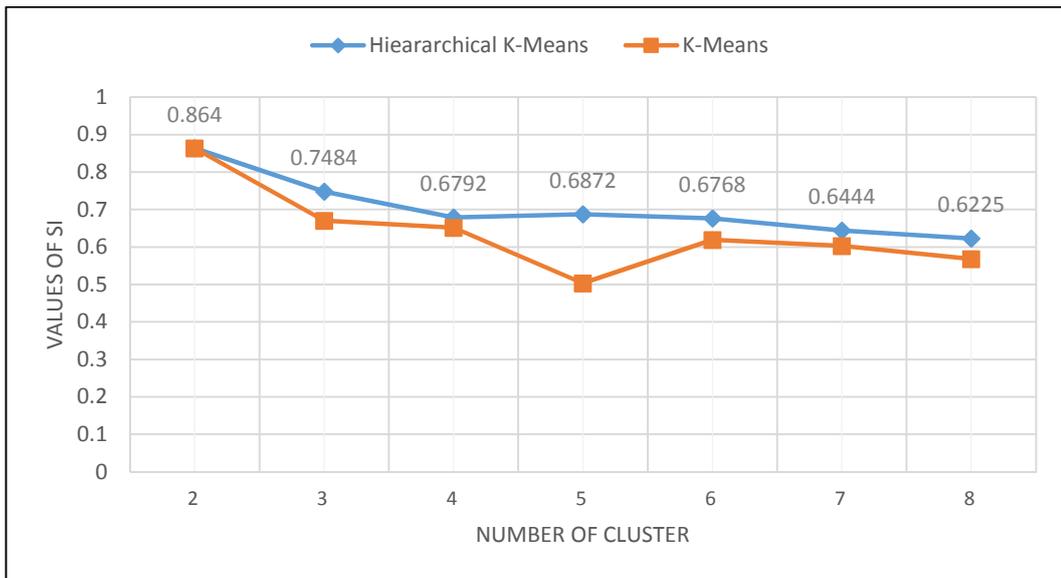
Jumlah Cluster	2	3	4	5	6	7	8	Max
ALK								0.864
Kmeans	0.864	0.7484	0.6792	0.6872	0.6768	0.6444	0.6225	
K-Means	0.864	0.6703	0.6518	0.5037	0.6197	0.6028	0.5679	0.864

Hasil uji validasi pengangguran berdasar kompetensi keahlian dari tabel 4.27 nilai maksimum pada setiap pengujian menunjukkan pada algoritma ALK (Average Linkage K-means) pada cluster 2 sebesar 0.864, hal itu menunjukkan bahwa cluster optimal dari algoritma ALK terletak pada cluster 2, dan apabila masing – masing nilai SI dirata – rata keduanya menghasilkan ALK = 0.703214286 dan K-Means = 0.640029, hasil nilai Silhoutte index pada gabungan Algoritma Average Linkage clustering dan K-means Clustering didapatkan nilai yang lebih tinggi dari K-Means clustering, jika makin besar nilai SI (Silhoutte) maka kualitas clusternya lebih baik maka dari kedua nilai tersebut kualitas cluster mengalami peningkatan sebesar 9.87%.

Tabel 4. 28 Nilai Standar Deviasi dari Silhoutte Index Data Set Kompetensi Keahlian

Algoritma	Mean	Standar Deviasi
ALK+K-Means	0.703214	0.081033
K-Means	0.640029	0.113141

Pada Tabel 4.28 menunjukkan perhitungan nilai standar deviasi dari nilai Silhoutte index pada dataset kompetensi keahlian, gabungan Average Linkage dan K-Means lebih kecil dari pada K-Means biasa, hal itu menunjukkan dari hasil pengujian gabungan ALK dan K-means lebih stabil hasil clusteringnya



Gambar 4. 17 Hasil Uji Coba Silhouette index karakteristik pengangguran berdasarkan Kompetensi Keahlian

Jumlah anggota cluster Mempunyai karakteristik masing – masing kemudian dianalisa dan ditampilkan pada tabel 4.30 :

Tabel 4. 29 Tabel Analisa Hasil Pengelompokan Pengguran berdasarkan umur

Cluster	Angkatan Kerja	Jurusan	Jumlah data
1	Pekerja Normal	Tidak Ada Isian	3322
		Admin Pemerintahan, perkantoran, manajemen Keuangan dan Bisnis	101
		Agama Dan Ilmu Kehutanan	25
		Arsitek Dan Perencanaan	6
		Bahasa sastra dan ekonomi	123
		Hukum, Humaniora dan IPS	67
		Kedokteran dan kesehatan	64
		Kehutanan dan pelayanan jasa	4
		Komunikasi Masa, Dokumentasi, Matematika dan Ilkom	39
		Perikanan	2
	Penganggur Terbuka	Tidak Ada Isian	554
		Admin Perkantoran, manajemen keuangan, agaman dan ilmu kehutanan	3
		Bahasa sastra, ekonomi, hukum dan IPS	12
		Kedokteran dan kesehatan	3
Kehutanan dan kesenian		2	

		Matematika dan Ilkom	2	
	Setengah Penganggur	Agama, ilmu kehutanan, Ilmu Ekonomi dan IPS	4	
		Kedokteran dan kesehatan	1	
		Matematika Ilkom	1	
		Tidak Ada Isian	38	
	Pekerja Paruh Waktu	Administrasi Perkantoran, Keuangan,Manajemen bisnis, ekonomi	21	
		Bahasa sastra, IPS, Agama dan Ilmu Kehutanan	30	
		Hukum dan Kehakiman	5	
		Kedokteran dan Kesehatan	10	
		Komunikasi Masa, dokumentasi dan Kesenian	3	
		Matematika dan Ilkom	4	
	2	Pekerja Normal	Bahasa sastra, Kependidikan dan keguruan	152
			Pertanian dan Peternakan	15
Petukangan kerajinan dan industri			4	
Psikologi			5	
SMEA Ekonomi dan SMU IPS			651	
SMK Perhotelan,Tata boga,Biologi,SMU IPA			372	
STM Listrik dan Teknik			344	
Transportasi			1	
Lainnya			11	
Tidak Ada Isian			24	
Bahasa dan sastra,Kependidikan dan Keguruan			4	
Pertanian			2	
Penganggur Terbuka			SMEA Ekonomi dan SMU IPS	58
		SMK Perhotelan dan SMU IPA	35	
		STM Listrik dan Teknik	37	
		Transportasi	1	
		Tidak Ada Isian	6	
		SMU IPS dan Teknik	4	
		Bahasa, sastra dan kependidikan dan Keguruan	67	
Setengan Pengangguran		Pertanian dan Peternakan	4	
Pekerja Paruh Waktu		Lainnya	3	
		Psikologi	1	
		SMEA Ekonomi,SMU IPS,SMK Perhotelan,SMKK Tata Boga	203	
		SMU IPA,SMU Biologi	56	
		STM Listrik Dan Teknik	46	
		Tidak Ada Isian	8	



## **BAB V**

### **KESIMPULAN DAN SARAN**

Dari hasil penelitian yang telah dilakukan dalam Tesis ini, dapat diambil kesimpulan sebagai berikut :

Setelah proses uji validasi cluster dengan silhouette index pada masing – masing data set faktor pengangguran dilakukan, didapatkan jumlah cluster yang ideal untuk dilakukan analisa pada masing – masing faktor pengangguran. Dari hasil uji validasi tersebut menghasikan beberapa kelebihan dan kelemahan pada masing – masing algoritma yaitu Algoritma gabungan dari Average Linkage dan K-Means menghasilkan rata – rata nilai Silhoutte index lebih tinggi dari pada K-Means murni, sehingga jika semakin tinggi nilai SI maka semakin baik kualitas clusternya maka kualitas cluster yang didapatkan dari Algoritma gabungan Average Linkage dan K-Means clustering lebih baik dari pada K-Means murni. Peningkatan nilai rata – rata dari kedua algoritma tersebut mencapai 9,01 % sampai 27,4 %. Untuk hasil Silhoutte dari algoritma ALK (Average Linkage K-Means) lebih stabil nilainya jika dibanding dengan K-Means murni, itu dibuktikan dengan nilai standar deviasi dari hasil uji validasi silhouette index selalu lebih kecil dari pada K-Means murni, akan tetapi kelemahan yang didapat adalah waktu komputasi yang lebih lama dari K-Means Murni, hal itu terjadi karena data set diolah menggunakan 2 algoritma yang berbeda.

Dari hasil analisa cluster dari masing – masing faktor pengangguran, pengangguran terbesar di wilayah perkotaan, kemudian berdasarkan pendidikan yang ditamatkan pengangguran terbesar mempunyai pendidikan rendah setingkat SD sampai tingkat SMP akan tetapi ada juga dijumpai pengangguran yang mempunyai pendidikan tinggi mulai SMK sampai perguruan tinggi S1. Pengangguran berdasarkan jenis kelamin terbesar pengangguran berjenis kelamin Laki – laki. Pengangguran berdasarkan umur, pengangguran terbesar pada rentan umur 45 – 65 keatas (Lansia) mencapai 491 sample, kemudia rentan umur 15 – 44 tahun mencapai 228 sample. sedangkan pengangguran berdasarkan kompetensi

keahlian atau latar belakang pendidikannya pengangguran terbesar adalah mereka yang tidak mempunyai latar belakang pendidikan (Tamatan SD) sebanyak 554 sample, kemudian yang mempunyai latar pendidikan keguruan, kependidikan dan SMK ekonomi, perhotelan, dll mencapai 208 sample, tapi ada juga latar belakang kedokteran dan kesehatan mencapai 2 sample.

## **SARAN**

Adapun saran dari penelitian ini adalah adanya penelitian yang lebih lanjut untuk membandingkan metode Algoritma Average Linkage K-Means dengan metode clustering lainnya guna menghasilkan kualitas cluster yang lebih baik.

Dari hasil analisa cluster tersebut dapat memberikan kebijakan yang berkaitan dengan masalah pengangguran, yaitu pemerintah daerah propinsi Jawa Timur dalam hal ini bisa Gubernur atau Dinas Terkait untuk menciptakan lapangan pekerjaan bagi tenaga kerja pendidikan rendah, seperti membuka bursa kerja khusus untuk pendidikan rendah, kemudian lebih menggiatkan bursa kerja di SMK untuk lebih banyak menyerap tenaga kerja dari tamatan SMK atau sederajat. Balai latihan kerja juga perlu ditingkatkan dan disesuaikan dengan kebutuhan dunia kerja saat ini.

## DAFTAR PUSTAKA

- Febti Eka Pratiwi, Ismaini Zain. (2014), "klasifikasi pengangguran terbuka menggunakan CART (Classification and regression tree) di propinsi Sulawesi Utara," *JURNAL SAINS DAN SENI POMITS*, Vol. 3, No.1.
- Larose, Daniel T. (2005), *Discovering Knowledge in Data: An Introduction to Data Mining*, John Willey & Sons, Inc., New York.
- BPS. (2014), *Laporan Eksekutif Keadaan Angkatan Kerja Di Jawa Timur Tahun 2013 – 2014*, BPS Propinsi Jawa Timur., Jawa Timur.
- Pramudiono, I, (2006), Apa itu Data Mining? Dalam <http://datamining.japati.net/cgi-bin/indodm.cgi>. Diakses tanggal 26 Februari 2011.
- Han, Kamber. (2001), *Data Mining: Concepts and Techniques 2nd ed*, Morgan Kaufmann Publishers., USA
- Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U. (2008), "Predicting expression patterns from regulatory sequence in Drosophila segmentation", *Nature* 451(7178): 535-540., USA
- Oyelade, Oladipupo, Obagbuwa. (2010), "Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance", *International Journal of Computer Science and Information Security*, Volume 7.
- Sun Jigui, Liu Jie, Zhao Lianyu. (2008), "Clustering algorithms Research", *Journal of Software*, Vol 19, No 1, pp.48-61.
- Izenman, Alan Julian. (2008), "Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning", *Springer*, New York.
- Tan, P., Steinbach, M., Kumar, V. (2006), *Introduction to Data Mining*, Pearson Education., New York.

Petrovic,S. (2006),”A Comparison between the Silhouette index and The Davies-Bouldin Index labeling IDS Cluster”. In:Proceeding of the 11 th Nordic Workshop of Secure IT Systems. 53-64

Rousseeuw, P. J. (1987), “Silhouettes : A Graphical Aid To The Interpretation And Validation of Cluster Analysis”, *Journal of Computational And Applied Mathematics*, 20, 53-56, New York.

Yuhefizar, Budi Santosa, I Ketut Eddy P, Yoyon K. Suprpto,”*Combination of Cluster Method for Segmentation of Web Visitors*”, TELKOMNIKA, Vol.11, No.1, pp. 207~214, March 2013

## BIOGRAFI PENULIS



Penulis lahir di Tulungagung, 3 Februari 1985. Penulis adalah anak ke – 3 dari 3 bersaudara dari pasangan ayahanda Suyitno dan Ibu Hj.Kustinah (Semoga Allah Melimpahkan rahmad dan kesehatan kepada beliau). Penulis sejak kecil mengenyam pendidikan di Tulungagung di SDN Wonorejo 1, kemudian melanjutkan pendidikan di SLTPN 2 Tulungagung dan lulus SMU tahun 2003 di SMU PGRI 1 Tulungagung. Kemudian di tahun yang sama penulis melanjutkan studi S1 di STIKI (Sekolah Tinggi Informatika Dan Komputer Indonesia) Malang jurusan Teknik Informatika dan meraih gelar sarjana computer pada tahun 2008. Kemudian penulis bekerja di beberapa perusahaan yaitu sebagai staff programmer di Global Net Malang, kemudian penulis juga sempat mengajar di Wearness Malang sebagai Dosen pengajar jurusan Teknik informatika, kemudian pada tahun 2009 bekerja sebagai Guru Swasta di SMKN 1 Boyolangu sebagai guru Rekayasa Perangkat Lunak, dan pada tahun 2011 Penulis diterima sebagai Pegawai Negeri Sipil (PNS) di SMKN 1 Trenggalek sebagai Guru TIK. Pada tahun 2014 penulis mendapatkan kesempatan untuk mendapatkan Beasiswa pendidikan S2 Chief Information Officer (CIO) dari Kementerian Komunikasi dan Informatika dengan jurusan Teknik Elektro dengan bidang Telematika CIO di Institut Teknik Sepuluh Nopember (ITS) Surabaya. Akhirnya penulis dapat menyelesaikan pendidikan S2 pada bulan September 2016. Dan setelah lulus penulis kembali melanjutkan karier untuk mengemban ilmu dan pengetahuan di Trenggalek sebagai Pengajar SMK. Penulis sangat tertarik pada bidang rekayasa perangkat lunak, komputasi cerdas, dan Data mining, penulis juga gemar browsing internet, medsos seperti facebook, untuk sekedar chatting di waktu senggang untuk berdiskusi masalah tesis dapat kontak dengan penulis melalui email [edogawa27@gmail.com](mailto:edogawa27@gmail.com), facebook melalui <http://www.facebook.com/edogawa27>