Clustering Data Kredit Bank Menggunakan Algoritma *Agglomerative Hierarchical Clustering Average Linkage*

[1]Ginanjar Abdurrahman

[1] Universitas Muhammadiyah Jember

Email: [1] abdurrahmanginanjar@gmail.com

(Naskah masuk: 15 Desember 2108, diterima untuk diterbitkan:29 desember 2018)

ABSTRAK

Data mining adalah pengembangan model yang merepresentasikan penemuan pola menggunakan data historis. Model dapat diaplikasikan pada data untuk prediksi (klasifikasi dan regresi), segmentasi populasi *(clustering)*, dan menentukan hubungan di dalam populasi (asosiasi). Dari beberapa model, salah satunya adalah *clustering* yang didefinisikan sebagai proses mengorganisir objek-objek menjadi satu kelompok yang anggotanya memiliki kemiripan tertentu. Similaritas ada dua, yakni similaritas berdasarkan bentuk dan jarak. *Clustering* mempunyai beberapa karakteristik, yaitu: *partitioning, hierarchical, overlapping, dan hybrid. Hierarchical clustering* adalah salah satu algoritma *clustering* dengan karakteristik setiap data harus termasuk dalam *cluster* tertentu, dan data yang termasuk dalam *cluster* tertentu tidak dapat berpindah ke *cluster lain*. *Hierarchical clustering* ada dua, yaitu *divisive (top to down)* dan *agglomerative (down to top).* Algoritma *agglomerative* ada empat yaitu *single linkage, centroid linkage, complete linkage, dan average linkage.* Salah satu dari algoritma *agglomerative* tersebut adalah *average linkage*. Algoritma ini merupakan algoritma terbaik di antara algoritma *hierarchical* yang lain, tetapi memiliki waktu komputasi tertinggi. Pada penelitian ini akan dilakukan *clustering* terhadap nasabah di suatu bank dengan algoritma *agglomerative hierarchical clustering average linkage*. Atribut data yang digunakan: *status pengecekan, durasi kredit, sejarah kredit, tujuan kredit, besaran kredit, status tabungan, employment, komitmen, status personal, pihak lain, menetap sejak, kepemilikan property, umur, rencana pembayaran lainnya, status rumah, keberadaan kredit, pekerjaan, jumlah tanggungan, telepon rumah, pekerja luar negeri, dan kelas*. Data dalam penelitian ini sebanyak 1000 instances, yang kemudian dijadikan sebagai data training sebanyak 25 %, 50 %, dan 75 %, sedangkan untuk data testing digunakan keseluruhan data.

Kata kunci : Data mining, Dataset, *Clustering*, *Agglomerative Hierarchical Clustering*, *Average Linkage*

ABSTRACT

Data mining is the development of model that represents pattern discovery using historical data. The model can be applied to data for prediction (classification and regression), population segmentation (clustering), and determining relationships within the population (association). Of the several models, one of them is clustering which is defined as the process of organizing objects into one group whose members have similarities. There are two similarities, namely similarity based on shape and distance. Clustering has several characteristics, namely: partitioning, hierarchical, overlapping, and hybrid. Hierarchical clustering is a clustering algorithm with the characteristics of each data must be included in a particular cluster, and data included in a particular cluster cannot moved to another cluster. There are two hierarchical clustering, namely divisive (top to down) and agglomerative (down to top). There are four agglomerative algorithms, namely single linkage, centroid linkage, complete linkage, and average linkage. One of the agglomerative is average linkage. This algorithm is the best hierarchical algorithms, but has the highest computational time. In this study clustering of customers in a bank conducted with the agglomerative hierarchical clustering average linkage. Data attributes used: checking status, credit duration, credit history, credit goals, loan size,

savings status, employment, commitment, personal status, other parties, settled since, property ownership, age, other payment plans, home status, credit availability, employment, number of dependents, landline, overseas workers and class. The data in this study were 1000 instances, which were then used as training data for 25%, 50%, and 75%, while for the testing data the entire data.

Keywords: Data mning, datasets, clustering, agglomerative hierarchical clustering, average linkage.

1. PENDAHULUAN

Data mining adalah proses penemuan pola dan hubungan dalam data (Hornick, Marcade & Venkayala, 2007:6). Selain itu, Hornick, Marcade & Venkayala (2007:6) juga mengemukakan bahwa data mining merupakan pengembangan model yang secara khusus merepresentasikan penemuan pola menggunakan data historis. Model dapat diaplikasikan pada data untuk prediksi (klasifikasi dan regresi), segmentasi populasi (*clustering*), dan menentukan hubungan di dalam populasi (asosiasi).

Dari beberapa model yang dapat diaplikasikan pada data, salah satunya adalah segmentasi populasi, dalam hal ini adalah *clustering* (pengelompokan). *Clustering* didefinisikan sebagai suatu proses mengorganisir objek-objek menjadi satu kelompok yang anggotanya memiliki similaritas (kemiripan) tertentu. Similaritas dibedakan menjadi dua, yakni similaritas berdasarkan bentuk dan similaritas berdasarkan jarak.

Clustering mempunyai beberapa karakteristik, yaitu: *partitioning clustering*, *hierarchical clustering, overlapping clustering*, dan *hybrid*. Pada *partitioning clustering,* setiap data harus termasuk pada *cluster* tertentu. Di samping itu, pada *partitioning clustering,* setiap data yang termasuk *cluster* tertentu pada suatu iterasi mempunyai kemungkinan berpindah ke *cluster* lain pada iterasi berikutnya. *Overlapping clustering*, setiap data mempunyai kemungkinan termasuk ke dalam beberapa *cluster*, atau dengan kata lain, data mempunyai nilai keanggotaan pada beberapa *cluster.* Yang ketiga adalah *hierarchical clustering*, yakni setiap data harus termasuk dalam *cluster* tertentu, dan data yang termasuk dalam *cluster* tertentu pada suatu iterasi, tidak dapat berpindah ke *cluster* lain. Karakteristik yang keempat adalah *hybrid*, yakni menggabungkan karakteristik *partitioning*, *overlapping*, dan *hierarchical*.

Algoritma *hierarchical* dibagi menjadi dua, yakni *divisive (top to down)* dan *agglomerative (down to top)*. Di dalam algoritma *agglomerative* terdapat empat macam metode, yaitu: *single linkage*, *centroid linkage*, *complete linkage*, dan *average linkage*. Salah satu algoritma *agglomerative yakni average linkage*, yang selanjutnya dipilih sebagai algoritma di dalam penelitian ini, disebut sebagai algoritma *agglomerative hierarchical clustering (ahc) average linkage.* Algoritma ini dipilih karena merupakan algoritma terbaik di antara algoritma *hierarchical* yang lain walaupun dengan waktu komputasi tertinggi di antara algoritma *hierarchical* yang lain (Barakbah, 2006: 35).

Pada penelitian ini, akan dilakukan *clustering* terhadap nasabah di suatu bank dengan menggunakan algoritma *agglomerative hierarchical clustering (ahc) average linkage.* Adapun atribut data yang digunakan adalah: status pengecekan, durasi kredit, sejarah kredit, tujuan kredit, besaran kredit, status tabungan,

employment, komitmen, status personal, pihak lain, menetap sejak, kepemilikan property, umur, rencana pembayaran lainnya, status rumah, keberadaan kredit, pekerjaan, jumlah tanggungan, telepon rumah, pekerja luar negeri, dan kelas. Data yang digunakan dalam penelitian ini sebanyak 1000 instances, Agar *datasets* dapat digunakan untuk *clustering* data harus dibersihkan terlebih dahulu dari *outlier* dan *extreme value*. Setelah data dibersihkan dari outlier dan extrema value, diperoleh data sebanyak 822 record data. Dari 822 *record data* tersebut kemudian dilakukan partisi data menggunakan metode *remove percentage* untuk data *training*. Partisi yang digunakan sebagai *data training* adalah 25% data, 50% data, 75% data, sedangkan 100% data selanjutnya digunakan sebagai data *testing*..

Untuk mempermudah pekerjaan, berkaitan dengan banyaknya *dataset* yang ada, diperlukan suatu program aplikasi untuk membantu penentuan *clustering* dataset kredit bank. Dalam hal ini, digunakan aplikasi weka 3.6 untuk meng*cluster* dataset, sehingga *clustering dataset* nantinya dapat ditemukan dengan metode *agglomerative hierarchical clustering* berdasarkan perhitungan weka 3.6 tersebut.

2. METODE PENELITIAN

Jenis penelitian ini adalah jenis penelitian kuantitatif, yaitu bernilai secara numerik atau dengan kata lain nilai-nilai peubah ini dinyatakan dalam bilangan real (Abadyo dan Hendro, 1999:3). Metode dalam penelitian ini adalah dengan mengimplementasikan algoritma AHC *Average Linkage* dengan menggunakan bantuan aplikasi *weka 3.9.1.* Adapun langkah-langkah dalam melakukan penelitian ini adalah sebagai berikut.

2.1 Identifikasi Masalah

Identifikasi masalah merupakan tahap awal dalam penelitian, yaitu dengan mengenali masalah yang ada apa saja serta menawarkan solusi yang dapat digunakan untuk menyelesaikan masalah tersebut.

2.2 Studi Literatur

Tahap ini merupakan tahap untuk mencari referensi berupa jurnal penelitian, paper, buku-buku referensi, serta referensi yang lain terkait dengan penelitian untuk melengkapi pengetahuan awal, guna memahami teori yang dapat digunakan untuk menunjang penelitian.

2.3 Dataset

Dataset diambil dari data simulasi yang sudah tersedia pada program *weka 3.9.1* dengan *source data: weka 3.9.1/data/german_credit.arff* dengan *record* sebanyak 1000 *instances*. Agar *datasets* dapat digunakan untuk *clustering* data harus dibersihkan terlebih dahulu dari *outlier* dan *extreme value*. Setelah data dibersihkan dari outlier dan extrema value, diperoleh data sebanyak 822 record data. Dari 822 *record data* tersebut kemudian dilakukan partisi data menggunakan metode *remove percentage* untuk data *training*. Partisi yang digunakan sebagai *data training* adalah 25% data, 50% data, 75% data, sedangkan 100% data selanjutnya digunakan sebagai data *testing*.

2.4 Algoritma AHC *Average Linkage*

Langkah-langkah algoritma AHC *Average Linkage* secara umum

1. Mencari jarak minimum dari dua objek.
2. Gabungkan dua objek dengan jarak minimum menjadi satu *cluster*
3. Cari jarak antar cluster dengan menggunakan rata-rata
4. Ulangi langkah 1-2 sampai semua objek bergabung menjadi satu *cluster*

2.5 Interpretasi dan Evaluasi

Tahap terakhir adalah melakukan interpretasi dan evaluasi. Pada tahap ini, informasi yang dihasilkan dari proses data mining perlu disajikan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.

2.6 Simulasi Perhitungan Manual Algoritma AHC *Average Linkage*

Contoh Kasus:

Jika diketahui matrik jarak antara 5 objek yang disajikan pada Tabel 1.

Tabel 1. Jarak Antara 5 Objek

	A	B	C	D	E
A	0	1	5	6	8
B	1	0	3	8	7
C	5	3	0	4	6
D	6	8	4	0	2
E	8	7	6	2	0

Langkah-langkah dalam algoritma AHC *Average Linkage* secara umum adalah sebagai berikut:

1. Mencari objek dengan jarak minimum

   A dan B mempunyai jarak minimum, yaitu 1, maka objek A dan B bergabung menjadi satu *cluster* (AB).
2. Menghitung jarak antara *cluster* AB dengan objek lainnya.

   $d_{(AB)C} = Average\ (d_{AC}, d_{BC}) = \frac{5+3}{2} = 4$

   $d_{(AB)D} = Average\ (d_{AD}, d_{BD}) = \frac{6+8}{2} = 7$

   $d_{(AB)E} = Average\ (d_{AE}, d_{BE}) = \frac{8+7}{2} = 7.5$

Dengan demikian, terbentuk matrik jarak yang disajikan pada Tabel 2

Tabel 2: Matrik Jarak Baru:

	AB	C	D	E
AB	0	4	7	7.5
C	4	0	4	6
D	7	4	0	2
E	7.5	6	2	0

Mencari jarak terdekat antara dua objek

D dan E mempunyai jarak terdekat, yaitu 2. Maka objek D dan E bergabung menjadi satu *cluster* (DE).

$d_{(AB)C} = Average\ (d_{AC}, d_{BC}) = \frac{5+3}{2} = 4$

$d_{(AB)(DE)} = Average\ (d_{AD}, d_{AE}, d_{BD}, d_{BE}) = \frac{6+8+8+7}{4} = \frac{29}{4} = 6.75$

$d_{(DE)C} = Average\ (d_{DC}, d_{EC}) = \frac{4+6}{2} = 5$

Dengan demikian matrik jarak barunya disajikan pada Tabel 3

Tabel 3: Matrik Jarak Baru

p-ISSN : 2502-5724; e-ISSN : 2541-5735

	AB	C	DE
AB	0	4	6.75
C	4	0	5
DE	6.75	5	0

4. Mencari jarak terdekat antara dua objek

   AB dan C mempunyai jarak minimum yaitu 4, sehingga objek AB dan C bergabung menjadi satu cluster (ABC).
5. Pada langkah terakhir, cluster ABC bergabung dengan DE sehingga terbentuk cluster tunggal (ABCDE).

3. HASIL PENELITIAN

Pada bagian ini akan dijelaskan mengenai hasil *clustering* menggunakan algoritma *Agglomerative Hierarchical Clustering (AHC)* pada data kredit nasabah dengan bantuan aplikasi *Weka 3.9.1*. Hasil penelitian yang selanjutnya dibahas sebagai berikut.

3.1 Persiapan Data *(Preprocessing)*

Datasets yang digunakan adalah datasets dari program internal *Weka 3.9.1* yang diperoleh dari direktori: "*C:\Weka-3-9\data\g-credit*". Banyaknya record data yang akan digunakan dalam penelitian ini sebanyak 1000 *record* data nasabah kredit, dengan 21 atribut, yakni: *checking status, duration, credit history, purpose, credit amount, savings status, employment, installment commitment, personal status, other parties, residence since, property magnitude, age, other payment plans, housing, existing credits, job, num dependents, own telephone, foreign worker, dan class.*

Agar *datasets* dapat digunakan untuk *clustering* data harus dibersihkan terlebih dahulu dari *outlier* dan *extreme value*. Dari dataset awal yang berjumlah 1000 record data, terdeteksi sebanyak 25 data sebagai *outlier* dan 155 data sebagai *extreme value. Outlier* dan *extreme value* dihilangkan dengan menggunakan *filter* dengan metode *Interquartile Range* sehingga diperoleh sebanyak 822 *record* data yang sudah bersih dari *outlier* maupun *extreme value*.

3.2 Skenario Uji Coba

Dari 822 *record data* kemudian dilakukan partisi data menggunakan metode *remove percentage* untuk data *training*. Partisi yang digunakan sebagai *data training* adalah 25% data, 50% data, 75% data, sedangkan 100% data selanjutnya digunakan sebagai data *testing*. *Remove percentage* yang digunakan adalah *remove percentage* 75% untuk memperoleh data sebanyak 25%, *remove percentage* sebesar 50% untuk memperoleh data sebanyak 50%, dan *remove percentage* sebanyak 25% untuk memperoleh data sebanyak 70%, dan data sebanyak 100% tanpa *remove percentage*. Masing-masing data tersebut, baik data *training* maupun data *testing* selanjutnya dibentuk menjadi 5 *cluster* dengan bantuan program aplikasi data *mining Weka 3.9.1* menggunakan algoritma *Agglomerative Hierarchical Clustering Average Linkage*.

3.3 Hasil dan Pembahasan

Berikut ini adalah hasil uji coba menggunakan bantuan perangkat lunak *Weka 3.9.1.*.

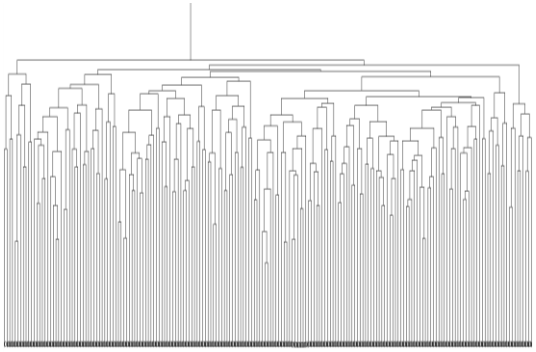
1. Data *training* sebanyak 25%, dibentuk 5 *cluster* disajikan pada Tabel 4 berikut.

Tabel 4: Hasil clustering data *training* sebanyak 25%

Cluster ke-	Jumlah anggota cluster
1	195 (95%)
2	3 (1%)
3	2 (1%)
4	4 (2%)
5	1 (0%)

Dengan menggunakan data *testing* sebanyak 205 (25%) diperoleh 4 *cluster* yakni cluster-1 berjumlah 195 (95%), *cluster*-2 berjumlah 3 (1%), *cluster*-3 berjumlah 2 (1%), cluster-4 berjumlah 4 (2%). Pada *cluster*-5 hanya berjumlah 1 anggota (1%),, sehingga dalam hal ini, tidak terbentuk *cluster*, dengan demikian jumlah cluster yang terbentuk hanyalah 4 *cluster*.

Visualisasi Hasil Clustering dengan alogritma *Agglomerative Hierarchical Clustering* jika disajikan dengan dendogram dari data *training* dengan melibatkan 205 data disajikan pada Gambar 1 berikut



Gambar 1: Dendogram untuk visualisasi hasil *clustering* dengan *Algoritma AHC* untuk data testing sebanyak 205 (25%) *record* data

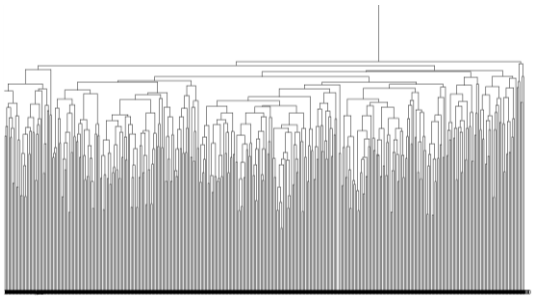
2. Data *training* sebanyak 50%, dibentuk 5 *cluster* disajikan pada Tabel 5.

Tabel 5: Hasil clustering data *training* sebanyak 50%

Cluster ke-	Jumlah anggota cluster
1	393 (96%)
2	4 (1%)
3	6 (1%)
4	7 (2%)
5	1 (0%)

Dengan menggunakan data *testing* sebanyak 411 (50%) diperoleh 4 *cluster* yakni cluster-1 berjumlah 393 (95%), *cluster*-2 berjumlah 4 (1%), *cluster*-3 berjumlah 6 (1%), cluster-4 berjumlah 7 (2%). Pada *cluster*-5 hanya berjumlah 1 (1%), sehingga dalam hal ini, tidak terbentuk *cluster*, dengan demikian jumlah *cluster* yang terbentuk hanyalah berjumlah 4.

Visualisasi Hasil Clustering dengan alogritma *Agglomerative Hierarchical Clustering* jika disajikan dengan dendogram dari data *training* dengan melibatkan 411 data disajikan pada Gambar 2 berikut



Gambar 2 Dendogram untuk visualisasi hasil *clustering* dengan *Algoritma AHC* untuk data testing sebanyak 411 (50%) *record* data

3. Data *training* sebanyak 75%, dibentuk 5 *cluster* disajikan pada Tabel 6.

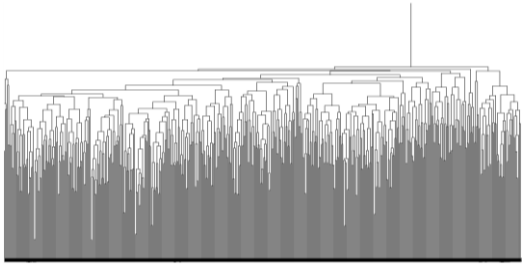
Tabel 6: Hasil clustering data *training* sebanyak 75%

Cluster ke-	Jumlah anggota cluster
1	557 (90%)
2	30 (5%)
3	3 (0%)
4	20 (3%)
5	6 (1%)

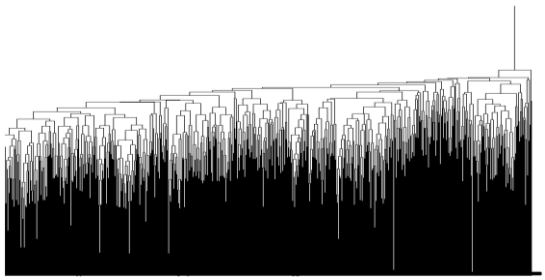
Dengan menggunakan data *testing* sebanyak 616 (50%) diperoleh 5 *cluster* yakni *cluster*-1 berjumlah 557 (90%), *cluster*-2 berjumlah 30 (5%), *cluster*-3 berjumlah 3 (0%), *cluster*-4 berjumlah 20 (3%). Pada *cluster*-5 berjumlah 6 (1%),

Visualisasi Hasil *Clustering* dengan alogritma *Agglomerative Hierarchical Clustering* jika disajikan dengan dendogram dari data *training* dengan melibatkan 616 data disajikan pada Gambar 3 berikut

p-ISSN : 2502-5724; e-ISSN : 2541-5735



Gambar 3 Dendogram untuk visualisasi hasil *clustering* dengan *Algoritma AHC* untuk data testing sebanyak 616 (75%) *record* data



Gambar 4 Dendogram untuk visualisasi hasil *clustering* dengan *Algoritma AHC* untuk data testing sebanyak 822 *record* data.

4. Data *testing* sebanyak 100%, dibentuk 5 *cluster* dapat dilihat pada Tabel 7.

Tabel 7: Hasil clustering data *training* sebanyak 100%

Cluster ke-	Jumlah anggota cluster
1	806 (98%)
2	5 (1%)
3	9 (1%)
4	1 (0%)
5	1 (0%)

Dengan menggunakan data *testing* sebanyak 822 (100%) diperoleh 3 *cluster* yakni *cluster*-1 berjumlah 806 (98%), *cluster*-2 berjumlah 5 (1%), *cluster*-3 berjumlah 9 (1%). Pada *cluster*-4 dan *cluster*-5 masing-masing hanya beranggotakan 1 (0%), sehingga dalam hal ini *cluster* tidak terbentuk. Dengan demikian jumlah *cluster* yang terbentuk hanyalah 3.

Visualisasi Hasil Clustering dengan alogritma *Agglomerative Hierarchical Clustering* jika disajikan dengan dendogram dari data testing dengan melibatkan 822 data disajikan pada Gambar 4 berikut.

4. KESIMPULAN

1. Algoritma *clustering agglomerative hierarchical clustering average linkage* merupakan algoritma *unsupervised learning* dengan menggabungkan dua *cluster* dan seterusnya menjadi *cluster* baru berdasarkan *similarity* yang ditentukan menggunakan jarak *Euclidean* dengan kriteria rata-rata jarak seluruh individu dalam *cluster* yang lain.
2. Dengan menggunakan data *testing* sebanyak 822 (100%) diperoleh 3 *cluster* yakni *cluster*-1 berjumlah 806 (98%), *cluster*-2 berjumlah 5 (1%), *cluster*-3 berjumlah 9 (1%). Pada *cluster*-4 dan *cluster*-5 masing-masing hanya beranggotakan 1 (0%), sehingga dalam hal ini *cluster* tidak terbentuk. Dengan demikian jumlah *cluster* yang terbentuk hanyalah 3.

DAFTAR PUSTAKA

Barakbah, A.R., 2006. *Clustering: workshop data mining 18-20 juli 2006.* Jurusan Teknologi Informasi Politeknik Elektronika Negeri Semarang.

Hornick, F.M., Marcade, E. & Venkayala, S. 2007. *Java data mining: strategy, standard, and practice: a practical guide for architecture, design, and*

implementation. San Fransisco: Elsevier.