BAB IV
HASIL PENELITIAN

4.1. Deskripsi Hasil Penelitian

Penelitian ini merupakan penelitian untuk melakukan privasi data dimana nantinya data dapat disebar tanpa diketahui secara detail informasinya. Pada penelitian ini dilakukan beberapa langkah agar mendapatkan hasil yang maksimal, yaitu: menarik data dari UCI, menormalkan data dan merubah formatnya agar bisa di-*import* ke *database*, lalu melakukan *clustering* dengan algoritma yang telah ditetapkan (*Systematic Clustering* dan *Greedy K-Member*), dan terakhir di analisis *information lossnya* untuk diketahui algortima mana yang lebih baik digunakan untuk privasi data.

4.1.1. Menarik Data

Hal pertama yang harus dilakukan adalah mencari data set yang tepat, data set yang akan digunakan haruslah representatif. Dalam hal ini, yang memnuhi kriteria adalah data set *Adult* pada UCI *Machine Learning Repository*. Data yang terdapat pada UCI *Machine Learning* adalah sebagai berikut:

```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical,
Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse,
Exec-managerial, Husband, White, Male, 0, 0, 13, United-States,
<=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners,
Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-
cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-
specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-
managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
```

Gambar 4.1. Data Set *Adult* pada UCI *Machine Learning*

4.1.2. Menormalkan Data

Data seperti pada **Gambar 4.1.** belum bisa di *import* ke dalam *database*. Oleh karena itu, data tersebut harus di save dulu di notepad dengan format .txt lalu dipindahkan ke Microsoft Excel dengan format.csv. Data dengan format csv inilah yang nantinya akan di *import* ke MySQL.

	ID	Age	Workclass	Marital_status	Sex	Capital_gain	Capital_loss	Cluster
2		39	State-gov	Never-married	Male	2174	0	
3		50	Self-emp-not-inc	Married-civ-spouse	Male	0	0	
4		38	Private	Divorced	Male	0	0	
5		53	Private	Married-civ-spouse	Male	0	0	
6		28	Private	Married-civ-spouse	Female	0	0	
7		37	Private	Married-civ-spouse	Female	0	0	
8		49	Private	Married-spouse-absen	Female	0	0	
9		52	Self-emp-not-inc	Married-civ-spouse	Male	0	0	
10		31	Private	Never-married	Female	14084	0	
11		42	Private	Married-civ-spouse	Male	5178	0	
12		37	Private	Married-civ-spouse	Male	0	0	
13		30	State-gov	Married-civ-spouse	Male	0	0	
14		23	Private	Never-married	Female	0	0	
15		32	Private	Never-married	Male	0	0	

Gambar 4.2. Data Set *Adult* pada Microsoft Excel dengan Format CSV

Untuk data yang terdapat pada *database* dapat dilihat pada Gambar 4.3 di bawah ini:

	ID	Age	Workclass	Marital_status	Sex	Capital_gain	Capital_loss	Cluster
Edit Copy Delete	2	39	State-gov	Never-married	Male	2174	0	1
Edit Copy Delete	3	50	Self-emp-not-inc	Married-civ-spouse	Male	0	0	1
Edit Copy Delete	4	38	Private	Divorced	Male	0	0	2
Edit Copy Delete	5	53	Private	Married-civ-spouse	Male	0	0	2
Edit Copy Delete	6	28	Private	Married-civ-spouse	Female	0	0	2
Edit Copy Delete	7	37	Private	Married-civ-spouse	Female	0	0	3
Edit Copy Delete	8	49	Private	Married-spouse-absent	Female	0	0	3
Edit Copy Delete	9	52	Self-emp-not-inc	Married-civ-spouse	Male	0	0	3
Edit Copy Delete	10	31	Private	Never-married	Female	14084	0	4
Edit Copy Delete	11	42	Private	Married-civ-spouse	Male	5178	0	4
Edit Copy Delete	12	37	Private	Married-civ-spouse	Male	0	0	4
Edit Copy Delete	13	30	State-gov	Married-civ-spouse	Male	0	0	5
Edit Copy Delete	14	23	Private	Never-married	Female	0	0	5
Edit Copy Delete	15	32	Private	Never-married	Male	0	0	5
Edit Copy Delete	16	40	Private	Married-civ-spouse	Male	0	0	6
Console dit Copy Delete	17	34	Private	Married-civ-spouse	Male	0	0	6

Gambar 4.3. Data Set *Adult* pada *Database*

4.1.3. Clustering dengan Algoritma Systematic Clustering

Hasil dari Algoritma *Systematic Clustering* dapat dilihat pada gambar dibawah ini:

Tabel 4.1. *Cluster* pada *Systematic Clustering*

Id	Age	Workclass	Marital	Sex	Capgain	Caploss	Cluster
1	17	Private	Never-married	Female	34095	0	1
2	17	Private	Never-married	Female	0	0	1
3	17	Private	Never-married	Female	0	0	1
4	17	Private	Never-married	Male	1055	0	2
5	17	Private	Never-married	Male	0	0	2
6	17	Private	Never-married	Male	0	0	2
7	17	Private	Never-married	Male	2176	0	3
8	17	Private	Never-married	Male	0	0	3
9	17	Private	Never-married	Male	0	0	3

Pada Tabel 4.1 data sudah di kelompokan berdasarkan *clusternya* dengan nilai k = 3. Maksud dari nilai k = 3 adalah minimal jumlah anggota yang terdapat di *cluster* sebanyak tiga. Algortima *Systematic Clustering* melakukan *clustering* berdasarkan kedekatan atau kesamaan antar atribut. Pada penelitian ini, *clustering* dilakukan berdasarkan umur yang sudah di *sorting* secara *ascending*. Sisa dari atributnya mengikuti umur dan untuk *quasi identifier*nya nantinya akan di lakukan generalisasi.

4.1.4. *Clustering* Dengan Algoritma *Greedy K-Member*

Hasil dari Algoritma *Systematic Clustering* dapat dilihat pada gambar dibawah ini:

Tabel 4.2. *Cluster* pada *Greedy K-Member*

Id	Age	Workclass	Marital	Sex	Capgain	Caploss	Cluster
1	17	Private	Never-married	Female	34095	0	1
2	17	Private	Never-married	Female	0	0	1
9	17	Private	Never-married	Female	0	0	1
4	17	Private	Never-married	Male	1055	0	2
7	17	Private	Never-married	Male	2176	0	2
5	17	Private	Never-married	Male	0	0	2
6	17	Private	Never-married	Male	0	0	3
3	17	Private	Never-married	Male	0	0	3
8	17	Private	Never-married	Male	0	0	3

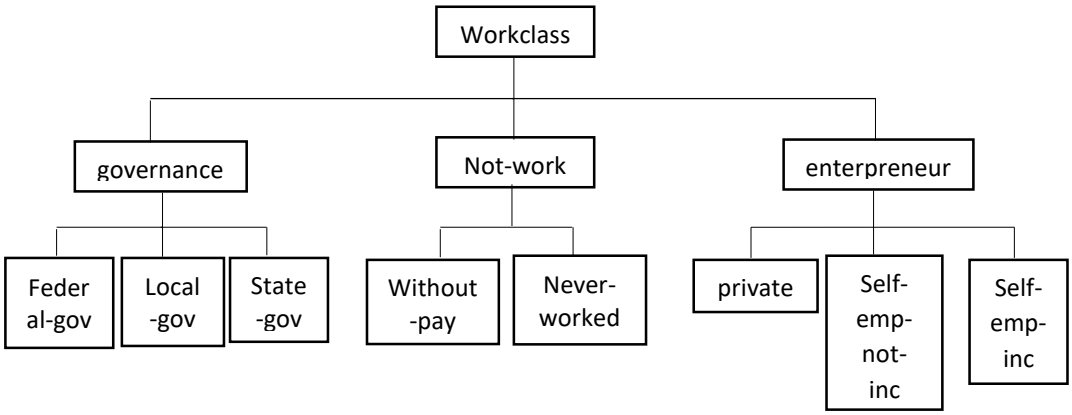
Tabel 4.2 merupakan hasil *clustering* dengan menggunakan Algoritma *Greedy K-Member*. Pada *Greedy K-Member*, untuk dapat membentuk suatu *cluster* maka *record* awal di ambil secara acak. Sebagai contoh, kita ambil *cluster* pertama. Dapat kita lihat pada gambar di atas, *record* pertama yang terpilih untuk *cluster*-1 *adalah record* ke 1. Nilai k pada gambar diatas adalah tiga sehingga membutuhkan dua *record* lagi untuk dapat dimasukan ke dalam *cluster*-1. Untuk mendapatkan dua *record* lagi dicari berdasarkan *find best record* dengan memasukkan satu satu sisa *record* ke dalam *cluster* lalu di generalisasi dan dicari *information loss*nya. *Information loss* terkecil lah yang nantinya akan di masukan ke dalam *cluster*. *Record* selanjutnya yang tergolong ke dalam *cluster*-1 adalah *record* ke 2 dan ke 9.

4.1.5. K-*Anonymity*

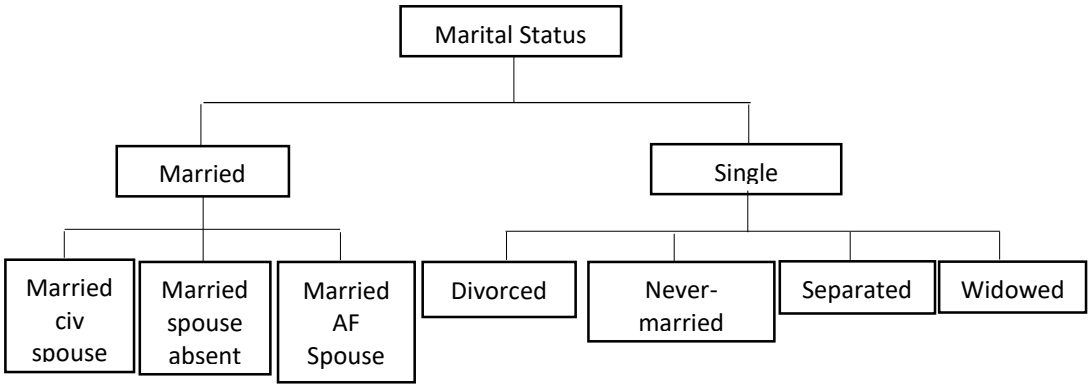
K-*Anonymity* merupakan proses menganonimkan atribut *quasi identifier*. Penganoniman disini dilakukan dengan menggunakan generalisasi. Pada *Systematic Clustering* digunakan proses generalisasi *local recording* karena data di*cluster*

berdasarkan kedekatan informasi. Sedangkan pada *Greedy K-Member* digunakan proses generalisasi *global recording* karena *cluster* dilakukan secara acak.

Generalisasi menggunakan teknik *taxonomy tree*. Bentuk *taxonomy tree* pada *local rerording* adalah sebagai berikut:



Gambar 4.4. *Taxonomy Tree* untuk *Workclass*



Gambar 4.5. *Taxonomy Tree* untuk *Marital Status*

Gambar 4.4 dan Gambar 4.5 merupakan *taxonomy tree* yang akan digunakan pada generalisasi. Kita ambil contoh pada *workclass*, jika pada satu *cluster* isi clusternya hanya teridiri dari *without-pay* dan *never-worked* maka *workclass* pada *cluster* tersebut akan digeneralisasi menjadi *not-work*. Demikian pula dengan yang

lainnya. Sedangkan untuk atribut *sex* karena hanya punya dua turunan, yaitu *female* dan *male*, maka misalnya jika dalam satu *cluster* isi dari *sex* adalah *male* semua maka generalisasinya tetap menjadi *male*. Akan tetapi jika dalam satu *cluster* terdapat *female* dan *male* maka akan di generalisasi menjadi *sex*. Untuk lebih jelasnya, dapat dilihat pada Tabel 4.3 untuk data yang sudah di generalisasi pada Algoritma *Systematic Clustering*:

Tabel 4.3. Generalisasi pada *Systematic Clustering*

Id	Age	Workclass	Marital	Sex	Capgain	Caploss	Cluster
1	17	Private	Never-married	Sex	34095	0	1
2	17	Private	Never-married	Sex	0	0	1
3	17	Private	Never-married	Sex	0	0	1
4	17	Private	Never-married	Male	1055	0	2
5	17	Private	Never-married	Male	0	0	2
6	17	Private	Never-married	Male	0	0	2
7	17	Private	Never-married	Male	2176	0	3
8	17	Private	Never-married	Male	0	0	3
9	17	Private	Never-married	Male	0	0	3

Sedangkan untuk *Greedy K-Member* yang menggunakan *global recording*, cara generalisasinya berbeda. Pada Algoritma *Greedy K-Member* generalisasi dilakukan saat proses meng*cluster*. Karena menggunakan *global recording* maka secara otomatis akan langsung di generalisasi ke tingkat tertingginya, tidak perlu dikelompok-kelompokan seperti *local recording*. Pada algoritma ini proses generalisasi dilakukan per-*cluster*. Untuk lebih jelasnya, dapat dilihat pada Tabel 4.4 untuk data yang sudah di generalisasi pada Algoritma *Greedy K-Member*:

Tabel 4.4. Generalisasi pada *Greedy K-Member*

Id	Age	Workclass	Marital	Sex	Capgain	Caploss	Cluster
1	17	Private	Never-married	Female	34095	0	1
2	17	Private	Never-married	Female	0	0	1
9	17	Private	Never-married	Female	0	0	1
4	17	Private	Never-married	Male	1055	0	2
7	17	Private	Never-married	Male	2176	0	2
5	17	Private	Never-married	Male	0	0	2
6	17	Private	Never-married	Male	0	0	3
3	17	Private	Never-married	Male	0	0	3
8	17	Private	Never-married	Male	0	0	3

4.2. Analisis Data Penelitian

Berdasarkan algoritma yang digunakan, maka dapat diketahui *information loss* dan *running time*nya. Hasil yang didapatkan adalah sebagai berikut:

Tabel 4.5. Hasil *Systematic Clustering*

K	Mulai Proses	Selesai Proses	Total	*Information Loss*
3	11:16:05	11:47:12	31:07	11843,9
4	12:08:17	12:25:26	17:09	12774,8
5	12:35:50	12:48:50	13:00	13554,5
6	12:59:23	13:21:20	22:03	13933.3
7	13:31:56	13:47:11	15:55	14344,1
8	15:12:05	15:28:11	16:06	14483,2
9	15:45:21	15:58:09	12:88	14611,1
10	16:07:44	16:18:51	11:07	14671,7
11	16:31:09	16:45:48	14:39	14705,2

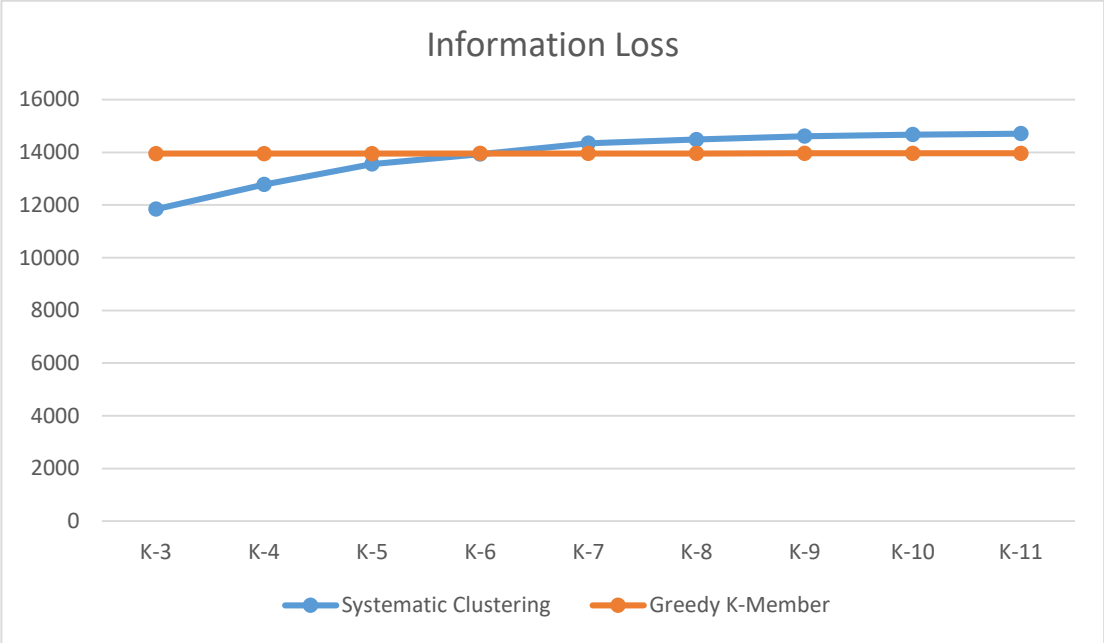
Sedangkan hasil untuk Algoritma *Greedy K-Member* adalah:

Tabel 4.6. Hasil *Greedy K-Member*

K	Mulai Proses	Selesai Proses	Total	*Information Loss*
3	10:24:43	10:59:23	34:13	13950.83
4	09:38:20	10:16:39	36:19	13950.85
5	11:18:03	11:55:20	37:17	13950.85

6	11:56:02	12:35:14	39:12	13955.67
7	12:35:33	13:10:34	35:01	13953.9
8	13:10:58	13:47:18	37:20	13956.33
9	13:48:03	14:25:21	37:18	13962.29
10	14:27:23	15:02:46	35:23	13962.23
11	15:05:23	15:40:28	35:05	13962.09

Supaya lebih mudah dalam melakukan analisis, maka penulis juga memaparkan hasilnya dalam bentuk grafik seperti dibawah ini:

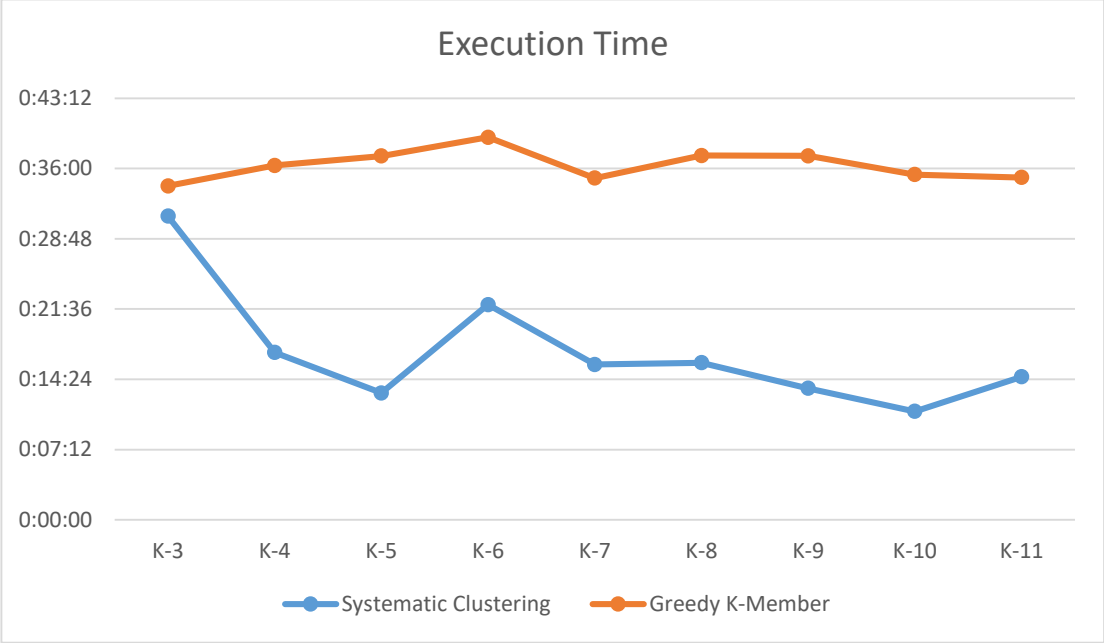


Gambar 4.6. Grafik Perbandingan *Information Loss* Algoritma *Systematic Clustering* dan *Greedy K-Member*

Berdasarkan Gambar 4.6 dapat dilihat jika perbandingan *information loss* antara *Systematic Clustering* dan *Greedy K-Member* tidak terlalu berbeda jauh. Perbedaan yang signifikan di antar dua algoritma tersebut adalah Agoritma *Systematic Clustering information loss* terus meningkat seiring dengan bertambahnya nilai K, sedangkan pada Algoritma *Greedy K-Member*, nilai *information loss* terlihat stagnan atau tidak terlalu

banyak mengalami peningkatan dari awal. Selain itu, pada *Systematic Clustering* titik jenuh terjadi ketika K=7, sedangakan pada *Greedy K-Member* titik jenuh terjadi dari awal yaitu ketika K=3. Dari data sebanyak 7415 data, *information loss Systematic Clustering* memiliki nilai terendah yaitu sebesar 11843,9 ketika K=3.

Selain berdasarkan *information loss*, waktu eksekusi juga harus di perhitungkan. Berikut ditampilkan grafik untuk *running time* kedua algoritma.

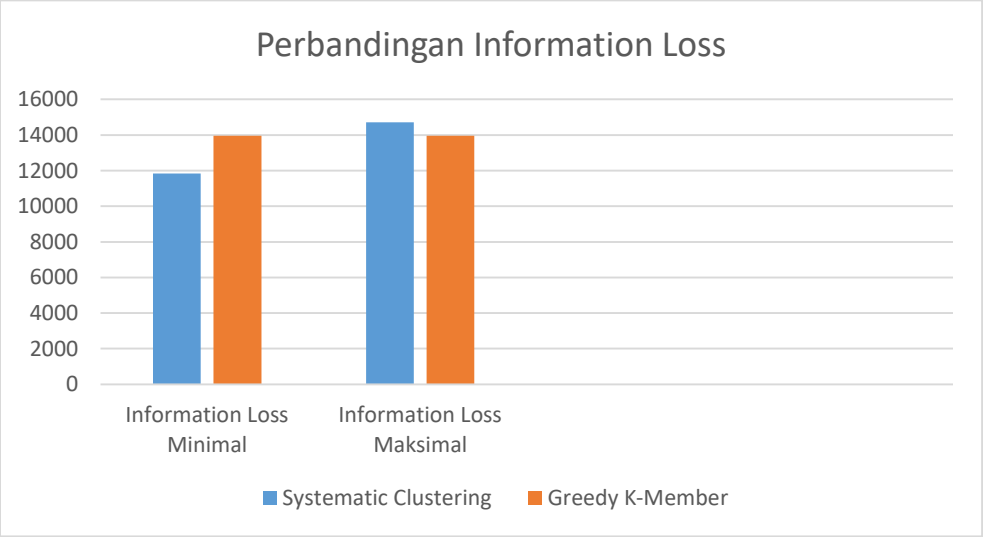


Gambar 4.7. Grafik Perbandingan *Execution Time* Algoritma *Systematic Clustering* dan *Greedy K-Member*

Pada Gambar 4.7 dapat dilihat jika waktu yang dibuhkan oleh Algoritma *Greedy K-Member* lebih banya dibandingkan *Systematic Clustering*. Hal ini terjadi karena pada Algoritma *Greedy K-Member* membandingkan satu persatu *record* yang ada untuk dapat membentuk suatu *cluster* sehingga membutuhkan waktu yang lebih lama.

4.3. Pembahasan

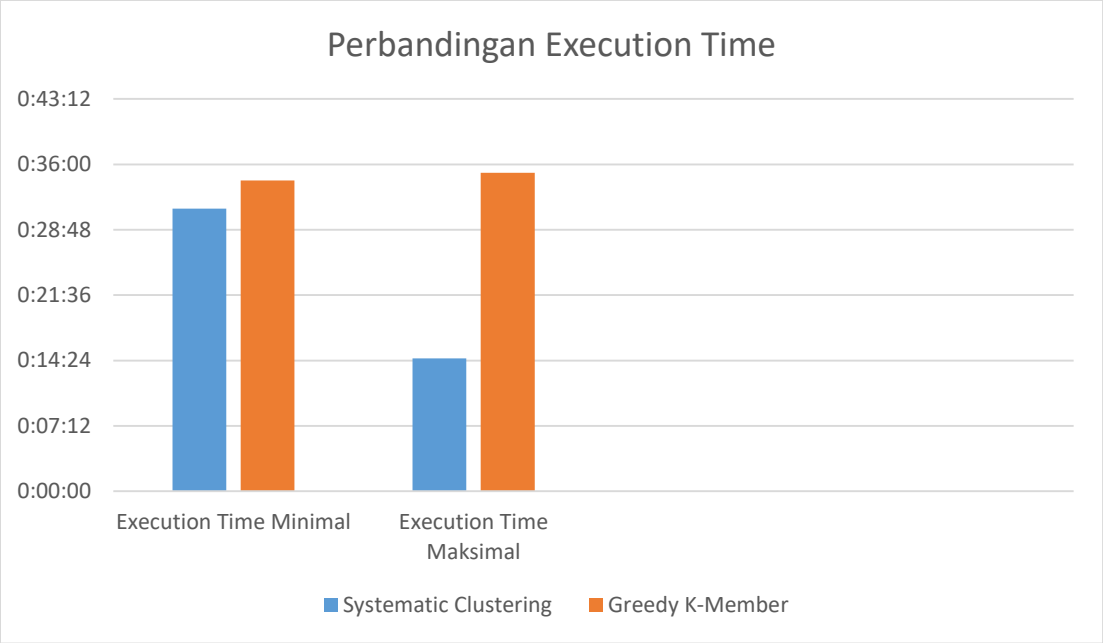
Berdasarkan analisis, maka dapat ditentukan algoritma mana yang lebih baik. Pada Algoritma *Systematic Clustering*, *Information Loss Total Minimal* berada pada K-3 bernilai 11843,9, sedangkan *Information Loss Total Maksimum* berada pada K-11 bernilai 14705,2. Pada Algortima *Greedy K-Member*, *Information Loss Total Minimal* berada pada K-3 bernilai 13950,83, sedangkan *Information Loss Total Maksimum* berada pada K-11 bernilai 13962,09. Untuk lebih memudahkan melihat perbandingannya, maka penulis juga memaparkan dalam bentuk grafik yang dapat dilihat pada Gambar 4.8.



Gambar 4.8. Grafik Perbandingan *Information Loss* Maksimal dan Minimal Algoritma *Systematic Clustering* dan *Greedy K-Member*

Selain berdasarkan *Information Loss*, akan dibandingkan juga berdasarkan *running time*. Pada Algoritma *Systematic Clustering*, *running time* minimal yang dibutuhkan untuk memproses data adalah 31 menit 7 detik, sedangkan *running time* maksimal yang dibutuhkan untuk memproses data adalah 14 menit 39 detik. Pada

Algortima *Greedy K-Member*, *running time* minimal yang dibutuhkan untuk memproses data adalah 34 menit 13 detik, sedangkan *running time* maksimal yang dibutuhkan untuk memproses data adalah 35 menit 5 detik. Untuk lebih memudahkan melihat perbandingannya, maka penulis juga memaparkan dalam bentuk grafik yang dapat dilihat pada Gambar 4.9.



Gambar 4.9. Grafik Perbandingan *Execution Time* Maksimal dan Minimal Algoritma *Systematic Clustering* dan *Greedy K-Member*

Berdasarkan *information loss* dan *running time*, maka dapat dilihat jika Algoritma *Systematic Clustering* lebih baik jika dibandingkan dengan Algortima *Greedy K-Member*. Penggunaan dua atribut sensitif pada penelitian ini tidak mempengaruhi *information loss* maupun *running time*. Akan tetapi, dua atribut sensitif disini mempengaruhi keanoniman suatu data.

4.4. Aplikasi Hasil Penelitian

Hasil dari penelitian ini dapat diterapkan di banyak perusahaan atau organisasi yang mengahruskannya untuk melakukan publikasi data tetapi masih ada data yang bersifat rahasia. Berikut merupakan beberapa lembaga yang memerlukan publikasi data, yaitu:

1. Rumah Sakit

Rumah sakit memiliki banyak data yang bersifat rahasia karena berkaitan dengan penyakit seseorang. Penyakit tentu saja sesuatu yang bersifat sensitif. Dengan menggunakan penganoniman data sebagaimana pada penelitian ini, maka pihak rumah sakit masih dapat mempublikasikan data yang ada ke publik karenan penganoniman ini sedniri meminimalisir peluang penggabungan dua tabel untuk mengetahui identitas pasien.

2. Kepolisian

Tidak jauh berbeda dengan rumah sakit, lembaga kepolisian juga memiliki banyak data yang bersifat rahasia, seperti jenis kriminal yang pernah dilakukan seseorang. Selain itu, lembaga kepolisian dianggap perlu mempublikasikan datanya untuk diketahui khalayak umum sebagai contoh daftar kejahatan yang terjadi dalam kurun waktu tertentu. Hal tersebut bertujuan untuk mengedukasi masyarakat untuk waspada terhadap tindakan kejahatan yang terjadi di lingkungan tempat tinggalnya. Dengan diterapkannya hasil penelitian ini, kepolisian dapat meminimalkan kebocoran informasi yang bersifat sensitif.