Privacy Preserving Data Mining metode k-Anonymity dengan Teknik Clustering

Apsari Ayusya Cantika–2016730012

1 Data Skripsi

Pembimbing utama/tunggal: **Mariskha Tri Adithia**
Pembimbing pendamping: **-**
Kode Topik : **MTA4701**
Topik ini sudah dikerjakan selama : **1** semester
Pengambilan pertama kali topik ini pada : Semester **47 - Ganjil 19/20**
Pengambilan pertama kali topik ini di kuliah : **Skripsi 1**
Tipe Laporan : **B** - Dokumen untuk reviewer pada presentasi dan **review Skripsi 1**

2 Latar Belakang

Di era digital ini, teknik *data mining* semakin banyak digunakan. Teknik ini merupakan teknik yang bertujuan untuk mendapatkan informasi dari data. Oleh karena itu, data harus dirilis agar teknik ini dapat dilakukan. Pada data yang dirilis, terdapat kemungkinan adanya data pribadi seseorang. Jika data dirilis, maka data dapat diakses oleh semua pihak. Saat data dapat diakses oleh semua pihak, pihak yang tidak bertanggung jawab juga dapat melihat data tersebut dan menyalahgunakannya. Akibatnya privasi tidak terlindungi.

Privasi merupakan kemampuan seseorang untuk mengatur bagaimana informasi mengenai dirinya disimpan, dipakai, maupun dihapus. Pada informasi tersebut bisa terdapat informasi atau data yang sensitif. Data sensitif dikenal juga dengan istilah *personally identifiable information* atau PII. PII merupakan informasi mengenai individu yang dikelola oleh perusahaan, termasuk informasi yang dapat dipakai untuk membedakan satu individu dengan individu lainnya. Contoh PII adalah data seperti nama lengkap, nomor kependudukan, dan lain-lain. Hal ini membuat privasi menjadi sesuatu yang penting dan perlu dilindungi. Privasi pada saat dilakukan *data mining* dapat dilindungi dengan *privacy preserving data mining*. *Privacy preserving data mining* adalah bagian dari *data mining* yang bertanggung jawab atas perlindungan privasi dalam proses *data mining*. Dengan adanya *privacy preserving data mining*, informasi bisa didapatkan tanpa perlu merilis data mentah. Sebagian besar metode *privacy preserving data mining* melakukan transformasi pada data yang akan ditambang sehingga data tidak terbuka seluruhnya.

Privacy preserving data mining diklasifikasikan menjadi empat kategori. Pertama, randomisasi. Randomisasi menambahkan distorsi ke data. Namun, distorsi yang ditambahkan harus cukup besar untuk menutupi data, terutama data sensitif. Kedua, *distributed privacy preservation*. Efek dari metode ini adalah perlindungan privasi harus dilakukan sambil menurunkan hasil agregat dari data. Ketiga, dengan cara menurunkan efektivitas dari hasil *data mining*. Metode ini lebih berfokus kepada modifikasi hasil *data mining*. Kategori yang terakhir adalah anonimisasi. Tujuan dari metode anonimisasi ialah membuat data individu sulit dibedakan dari data lainnya.

Salah satu contoh metode anonimisasi adalah *k-anonymity*. Dengan metode *k-anonymity*, data akan sulit dibedakan setidaknya dengan k-1 data lainnya. *k-anonymity* dapat dilakukan dengan beberapa teknik, contohnya *hash*, semantik, dan *clustering*. Metode *k-anonymity* dengan teknik *clustering* memanfaatkan algoritma *clustering* untuk melakukan anonimisasi. Data akan dikelompokkan dengan algoritma *clustering* lalu tiap kelompok atau *cluster* akan digeneralisasi. Setelah generalisasi selesai dilakukan, maka akan didapat hasil anonimisasi seperti pada Tabel 1.

Tabel 1: Tabel Hasil Anonimisasi *k-anonymity* dengan $k = 2$

Education	Race	Sex	Age	Workclass
Bachelors	White	Male	39-42	State-gov
Bachelors	White	Male	39-42	Private
*	White	Male	50-52	Self-emp-not-inc
*	White	Male	50-52	Self-emp-not-inc
*	White	*	37-38	Private
*	White	*	37-38	Private
High	*	Female	28-31	Private
High	*	Female	28-31	Private
Low	Black	*	49-53	Private
Low	Black	*	49-53	Private

Pada penelitian ini, akan dibangun sebuah perangkat lunak yang mengimplementasikan dua algoritma *k-anonymity* dengan teknik *clustering*, yaitu *One Pass k-Means* (OKA) dan *Grading, Centering, Clustering, Generalization* (GCCG). Perangkat lunak ini akan menerima masukan berupa tabel data yang ingin dianonimisasi, pohon klasifikasi, serta nilai k yang diinginkan pengguna. Sedangkan keluaran dari perangkat lunak ini adalah tabel *k-anonymized* yang merupakan tabel hasil anonimisasi.

3 Rumusan Masalah

Rumusan masalah yang akan dibahas pada skripsi ini adalah:

1. Bagaimana cara kerja Algoritma OKA dan GCCG untuk anonimisasi?
2. Bagaimana cara mengimplementasikan Algoritma OKA dan GCCG untuk anonimisasi?
3. Bagaimana cara mengukur performa Algoritma OKA dan GCCG untuk anonimisasi?

4 Tujuan

Tujuan yang ingin dicapai dari skripsi ini:

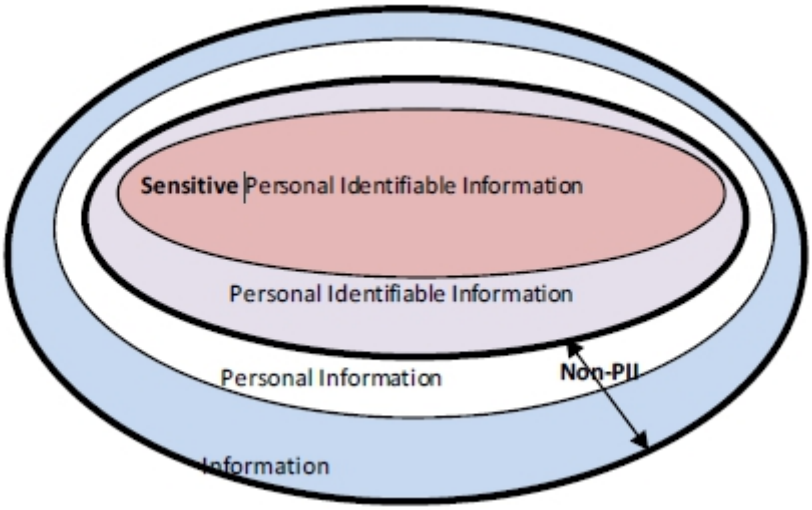
1. Mempelajari Algoritma OKA dan GCCG.
2. Membangun perangkat lunak yang mengimplementasikan Algoritma OKA dan GCCG.
3. Mengukur performa Algoritma OKA dan GCCG dengan mengimplementasikan teknik *data mining*.

5 Detail Perkembangan Pengerjaan Skripsi

Detail bagian pekerjaan skripsi sesuai dengan rencana kerja/laporan perkembangan terkahir :

1. **Studi literatur mengenai privasi dan *personally identifiable information* (PII)**
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Menurut Schoeman, privasi dapat didefinisikan menjadi beberapa perspektif yang berbeda. Definisi tersebut dibedakan sebagai berikut:
   - Privasi sebagai hak individu untuk menentukan informasi tentang dirinya sendiri yang dapat dikomunikasikan kepada orang lain.
   - Privasi sebagai ukuran kendali seseorang atas informasi mengenai dirinya sendiri.
   - Privasi sebagai keadaan atau keterbatasan akses ke seseorang.

Jadi, privasi merupakan kemampuan seseorang untuk mengatur bagaimana informasi mengenai dirinya disimpan, dipakai, maupun dihapus. Seiring dengan kemajuan teknologi, informasi tersebut dapat disimpan dalam bentuk digital dan dirilis untuk kebutuhan tertentu. Hal ini mengakibatkan data milik seseorang dapat dilihat oleh pihak yang tidak bertanggung jawab dan disalahgunakan, padahal data dapat mengandung data sensitif (Gambar 1).



Gambar 1: Informasi dan PII

Data sensitif dikenal juga dengan istilah *personally identifiable information* atau PII. PII merupakan informasi mengenai individu yang dikelola oleh perusahaan, termasuk informasi yang dapat dipakai untuk membedakan satu individu dengan individu lainnya. PII terbagi menjadi dua kategori, yaitu PII yang dapat dipakai mengidentifikasi individu secara langsung dan PII yang dapat dipakai mengidentifikasi individu tidak secara langsung. Contoh PII yang dapat mengidentifikasi individu secara langsung adalah:

(a) NIK, tempat tanggal lahir, nama lengkap pada data kependudukan,

(b) Nama ibu kandung dan nomor rekening nasabah pada data perbankan,

(c) Data biometrik seperti data sidik jari, retina mata, dan bentuk wajah,

(d) Nomor induk atau nomor pokok mahasiswa, alamat, dan nomor telepon pada data mahasiswa,

(e) Alamat IP (Internet Protocol) atau alamat MAC (Media Access Control) pada informasi aset,

(f) Nomor registrasi kendaraan seseorang

(g) Foto berisi wajah atau karakteristik unik lainnya

Sedangkan contoh PII yang dapat mengidentifikasi individu tidak secara langsung, ialah:

(a) rekam medis seseorang pada data rumah sakit,

(b) informasi edukasi atau pendidikan,

(c) informasi kepegawaian

(d) informasi letak geografis

(e) informasi finansial

Jika informasi yang ada dapat dipakai untuk mengidentifikasi individu, maka data tersebut merupakan PII.

2. **Studi literatur mengenai teknik** *data mining* **dan** *privacy preserving data mining*
   Status : Ada sejak rencana kerja skripsi.
   Hasil : *Data mining* atau penambangan data adalah proses untuk menemukan pola menarik dan informasi dari sejumlah data yang besar. Proses ini juga sering disebut dengan istilah *knowlegde discovery from data*. Data yang dibutuhkan untuk proses *data mining* bisa berasal dari basis data, *data warehouses*, situs web, repositori informasi, atau pun data yang dialirkan ke sistem secara dinamis. *Data mining* dapat dilakukan pada jenis data apapun selama data tersebut sesuai dengan target yang ingin dicapai. Terdapat beberapa teknik dalam *data mining* yaitu klasifikasi, *clustering*, dan aturan asosiasi.

   Clustering merupakan proses mengelompokkan objek menjadi beberapa kelompok atau *cluster*. Proses ini akan membuat objek dalam sebuah *cluster* memiliki tingkat kemiripan yang tinggi dengan anggota *cluster* yang sama, namun memiliki tingkat kemiripan yang rendah dengan dengan anggota *cluster* lain. Tingkat kemiripan ini bergantung kepada nilai atribut yang dimiliki objek tersebut. Nilai-nilai tersebut nantinya akan dihitung dengan rumus jarak atau kemiripan antar objek.

   Salah satu contoh algoritma *clustering* yang sederhana ialah Algoritma *k-Means*. Pada Algoritma *k-Means*, tahapan *clustering* dilakukan sebagai berikut:

   (1) Ambil k baris pada data secara acak.
   (2) Jadikan k baris tersebut sebagai *centroid* atau titik tengah *cluster*.
   (3) Hitung jarak baris lainnya yang bukan *centroid* dengan semua *centroid*.
   (4) Perbaharui nilai *centroid* dengan menghitung rata-rata nilai semua baris pada *cluster*.
   (5) Ulangi langkah (2) sampai (4) sampai nilai *centroid* dan anggota *cluster* tidak berubah.

   Teknik *data mining* lainnya adalah klasifikasi. Klasifikasi merupakan proses menemukan sebuah model yang mendeskripsikan dan membedakan kelas dari data. Contohnya jika pada data terdapat kelas *iris-setosa, iris-virginica,* dan *iris-versicolor*, maka dengan klasifikasi dapat ditemukan suatu data akan termasuk ke dalam kelas *iris-setosa* atau *iris-virginica* atau *iris-versicolor*.

   Contoh algoritma klasifikasi adalah Algoritma *k-Nearest Neighbor* (kNN). Berikut ini tahapan Algoritma kNN:

   (1) Hitung jarak antara baris yang ingin diketahui kelasnya dengan semua baris pada data
   (2) Urutkan baris berdasarkan jaraknya secara terurut menaik
   (3) Ambil kelas mayoritas pada k baris terdekat

   Pada proses *data mining*, PII dikumpulkan dan ditaruh dalam bentuk digital. Data tersebut dapat diakses oleh program *data mining* dan memungkinkan adanya pelanggaran privasi. Oleh karena itu, untuk melindungi privasi data diperlukan sebuah metode perlindungan. *Privacy preserving data mining* merupakan bagian dari *data mining* yang bertanggung jawab atas perlindungan privasi dalam proses *data mining*. Hal ini berkaitan dengan memperoleh hasil *data mining* tanpa mengungkapkan data sensitif yang mendasarinya. Data sensitif akan dimodifikasi sebelum dilakukan teknik *data mining*. Modifikasi ini dapat menyebabkan hilangnya beberapa informasi dan kemungkinan kegunaan dari hasil *data mining*.

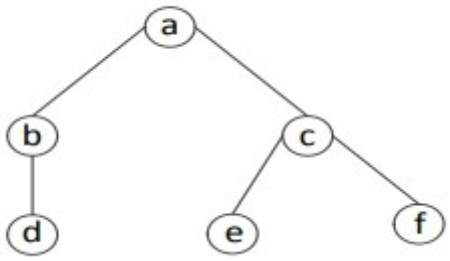
   Metode untuk mencapai *privacy preserving data mining* dapat diklasifikasikan menjadi empat kategori, yaitu randomisasi, anonimisasi, *distributed privacy preservation*, dan menurunkan efektivitas dari hasil *data mining*. Pada penelitian ini, metode akan difokuskan pada metode anonimisasi. Metode ini bertujuan untuk membuat data individu sulit dibedakan dari data lainnya melalui generalisasi dan supresi. Contoh metode yang merupakan metode anonimisasi adalah *k-anonymity*, *l-diversity*, dan *t-closeness*.

3. **Studi literatur mengenai metode *k-anonymity* dengan teknik *clustering***
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Metode *k-anonymity* merupakan salah satu metode untuk melakukan proses anonimisasi. Metode ini membuat suatu data menjadi sulit dibedakan dengan $k-1$ data lainnya. *k-anonymity* dapat dilakukan dengan beberapa teknik, seperti *hash*, semantik, dan *clustering*. Metode *k-anonymity* dengan teknik *clustering* merupakan metode untuk melakukan anonimisasi yang memanfaatkan algoritma *clustering*. Data yang akan dianonimisasi dibagi ke dalam kelompok-kelompok atau *cluster* dengan teknik *clustering*. Setelah terbentuk kelompok-kelompok, data pada setiap kelompok akan digeneralisasi. Generalisasi berarti data akan diubah menjadi bentuk yang lebih umum. Jika data berupa numerik, maka data akan dibuat menjadi rentang data terkecil sampai data terbesar pada kelompoknya masing-masing. Contoh generalisasi data numerik ada pada (Table 2).

Tabel 2: Tabel Contoh Generalisasi Atribut Numerik

Data pada Sebuah *Cluster*	
Sebelum Generalisasi	Setelah Generalisasi
5	[5-10]
7	[5-10]
10	[5-10]

Data bernilai kategori harus digeneralisasi dengan bantuan pohon klasifikasi / *classification tree*. Pohon klasifikasi mendefinisikan hierarki generalisasi data kategori. Jika data dan *centroid* bernilai sama, maka data tidak diubah. Jika data berbeda dengan *centroid* maka data dan *centroid* akan diubah ke bentuk yang lebih general. Bentuk general ini didapat dari *closest common ancestor* pada pohon klasifikasi. Contoh pohon klasifikasi dapat dilihat pada (Gambar 2 dan contoh generalisasi atribut kategori dapat dilihat pada (Table 3).



Gambar 2: Contoh Pohon Klasifikasi

Tabel 3: Tabel Contoh Generalisasi Atribut Kategori

Data pada Sebuah *Cluster*	
Sebelum Generalisasi	Setelah Generalisasi
c	a
c	a
d	a

Meskipun generalisasi tidak menghilangkan informasi apapun (hanya mengganti nilainya), namun cara ini tetap akan menghasilkan *information loss*. *Information loss* adalah ukuran yang menunjukkan banyaknya informasi yang hilang setelah data dianonimisasi. Semakin rendah nilainya maka semakin baik. Nilai *information loss* dihitung per *cluster* dan atribut. *Information Loss* untuk atribut numerik dihitung dengan rumus sebagai berikut:

$$IL = \frac{X_{max} - X_{min}}{Max - Min} \tag{1}$$

- X_{max} merupakan nilai maksimum pada *cluster* tersebut,
- X_{min} merupakan nilai minimum pada *cluster* tersebut,
- Max merupakan nilai maksimum dari seluruh baris pada data,
- Min merupakan nilai minimum dari seluruh baris pada data.

Sedangkan untuk atribut kategori, *information loss* dihitung dengan rumus sebagai berikut:

$$IL = \frac{size(c)}{Size} \tag{2}$$

- c merupakan nilai atribut setelah generalisasi,
- $size(c)$ merupakan jumlah *descendant leaf nodes* pada pohon klasifikasi,
- $Size$ merupakan jumlah semua node *leaf*.

Total *information loss* dihitung dengan rumus:

$$IL = \frac{\sum_{i=1}^{m} IL_i}{m} \tag{3}$$

- m merupakan banyaknya baris pada data
- IL_i merupakan nilai *information loss* tiap baris pada data. Nilai ini didapat dari jumlah *information loss* semua atribut pada baris tersebut.

Pada *k-anonymity*, atribut (kolom) dibagi ke dalam tiga kategori. *Identifier, quasi-identifier,* dan *sensitive identifier*. *Identifier* merupakan atribut yang dapat dipakai untuk mengidentifikasi seseorang secara langsung, seperti nama. Atribut bertipe *identifier* biasanya akan dihilangkan sebelum data dianonimisasi. *Quasi-identifier* (QI) adalah atribut yang nilainya secara tidak langsung dapat dipakai mengidentifikasi seseorang. Jika atribut bertipe ini digabungkan dengan beberapa informasi eksternal maka atribut dapat dipakai untuk mengidentifikasi seseorang. Secara teoritis, semua atribut selain *identifier* pada data merupakan *quasi-identifier*. Contoh atribut bertipe ini ialah kode pos, tanggal lahir, dan lain-lain. Atribut bertipe ini merupakan atribut yang nantinya akan dianonimisasi. *Sensitive identifiers* adalah atribut yang berkaitan dengan informasi privasi sensitif seperti gaji dan informasi kesehatan. Atribut *sensitive identifiers* bergantung pada konteks, sehingga pada proses anonimisasi atribut ini tidak akan diubah.

4. **Studi literatur mengenai Algoritma OKA**
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Algoritma *One Pass k-Means* (OKA) adalah turunan dari salah satu algoritma *clustering*, yaitu Algoritma *k-Means*. Perbedaan OKA dengan k-Means terletak pada perulangannya. OKA hanya melakukan satu kali perulangan saja, tidak seperti k-Means yang melakukan perulangan berkali-kali. Algoritma ini terbagi ke dalam dua tahap, yaitu *clustering* dan *adjustment*. Algoritma OKA tahapan *clustering* dapat dilihat pada (Algoritma 1) dan tahapan *adjustment* pada (Algoritma 2).

Algorithm 1: Algoritma *Clustering One Pass k-Means* (OKA) - Tahap *Clustering*

input : data, k
output: clustered_data
urutkan data sesuai *quasi-identifier*
$p \leftarrow \lfloor \frac{n}{k} \rfloor$
pilih p baris dari data secara acak untuk menjadi *centroid*
$C_i \leftarrow r_i$ for i ← 1 sampai p
hapus p baris dari data
while *data* $\neq \emptyset$ **do**
    r ← baris pertama pada data
    hitung jarak r ke semua p
    tambahkan r ke C_i terdekat
    perbaharui nilai *centroid* C_i
    hapus r dari data
end

Algorithm 2: Algoritma *Clustering One Pass k-Means* (OKA) - Tahap *Adjustment*

input : clustered_data, k
output: adjusted_data
$R \leftarrow \emptyset$
for *setiap cluster pada clustered_ data dengan ukuran* $> k$ **do**
    urutkan baris pada *cluster* berdasarkan jaraknya terhadap *centroid* secara terurut menaik
    while *ukuran cluster pada clustered_ data* $> k$ **do**
        $r \leftarrow$ data terjauh dari *centroid*
        masukkan r ke R
        keluarkan r dari clustered$_data$
    end
end
while $R \neq \emptyset$ **do**
    if *ada cluster dengan ukuran* $< k$ **then**
        masukkan r ke *cluster* dengan ukuran < k terdekat
    else
        masukkan r ke *cluster* terdekat
    end
end

Pada tahap *clustering*, data akan diurutkan terlebih dahulu berdasarkan *quasi-identifier*. Pengurutan data nantinya akan sama dengan tahap *grading* dan *centering* pada Algoritma GCCG. Setelah data terurut diambil *centroid* secara acak, dan data selain *centroid* akan dikelompokkan dengan *centroid* terdekat. Terakhir, nilai *centroid* akan diperbaharui dengan mengambil baris yang memiliki jarak paling dekat ke semua baris pada *cluster* tersebut.

Data yang telah dikelompokkan akan disesuaikan pada tahap *adjusment*. Pada tahap *adjustment* data terjauh dari *centroid* pada kelompok dengan jumlah anggota lebih dari k akan dipindahkan ke kelompok dengan jumlah anggota kurang dari k dengan *centroid* terdekat. Semua perhitungan jarak antar baris dilakukan dengan *Gower Distance* karena perhitungan jarak ini cocok untuk atribut campuran (terdiri dari atribut numerik dan kategori). Berikut ini rumus *Gower Distance* untuk atribut numerik:

$$dist = \frac{\mid a_1 - a_2 \mid}{Max - Min} \tag{4}$$

- a_1 merupakan nilai atribut dari baris-1,
- a_2 merupakan nilai atribut dari baris-2,
- Max merupakan nilai maksimum dari atribut ini,
- Min merupakan nilai minimum dari atribut ini.

Sedangkan untuk atribut kategori dilakukan dengan rumus:

$$dist = \begin{cases} 1, & \text{if } a_1 = a_2 \\ 0, & \text{if } a_1 \neq a_2 \end{cases} \tag{5}$$

- a_1 merupakan nilai atribut dari baris-1,
- a_2 merupakan nilai atribut dari baris-2.

Perhitungan jarak total dihitung dengan menjumlahkan jarak semua atribut.

Setelah dua tahap ini dilakukan, *cluster* yang terbentuk akan dianonimisasi dengan cara generalisasi. Setelah generalisasi dilakukan, kualitas tabel anonim atau hasil anonimisasi dapat dilihat dari nilai *information loss*.

Masukan yang diperlukan untuk algoritma ini adalah data yang akan dianonimisasi dan nilai k (konstanta *k-anonymity*). Sedangkan keluaran yang dihasilkan adalah data yang sudah teranonimisasi.

5. **Studi literatur mengenai Algoritma** *Clustering using Representative*
   Status : Ada sejak rencana kerja skripsi.
   Hasil :

   Algoritma *Clustering using Representative* (CURE) merupakan algoritma *clustering* yang efektif dipakai pada set data yang berukuran besar. Pada algoritma ini, teknik *hierarchical clustering* dan *partitional clustering* digabungkan. Pada awalnya sebagian data akan dikelompokkan menggunakan *hierarchical clustering*. Setelah itu, data lainnya akan dikelompokkan menggunakan *partitional clustering*. Algoritma ini memiliki kelebihan efisien untuk data set besar dan lebih kuat untuk menghadapi pencilan. Namun algoritma ini memiliki kelemahan, yaitu hanya dapat dipakai untuk data bertipe numerik saja. Tahapan Algoritma CURE dapat dilihat pada (Algoritma 3).

   Algorithm 3: Algoritma *Clustering using Representative* (CURE)

   input : data, k, x
   output: clustered_data
   1 ambil sebagian data secara acak untuk membuat initial_cluster
   2 buat initial_cluster dengan *hierarchical clustering*
   3 geser *centroid* semua *cluster* sebanyak x%
   4 kelompokkan data ke *centroid* terdekat

   Pada awalnya, data akan diambil sebagian secara acak. Kemudian data tersebut akan dijadikan *cluster* inisial dengan menggunakan algoritma *hierarchical clustering*. Setelah *cluster* terbentuk, *centroid* setiap *cluster* akan digeser sebanyak x%. Data yang belum dikelompokkan akan dikelompokkan ke *centroid* terdekat.

6. **Studi literatur mengenai Algoritma** *Average-link Agglomerative*
   Status : Baru ditambahkan pada semester ini.
   Hasil : Algoritma CURE yang sebelumnya dipelajari memiliki kelemahan, yaitu hanya dapat dipakai untuk data bertipe numerik. Oleh karena itu, dipelajari algoritma lainnya yang cocok untuk atribut

campuran (numerik dan kategori), yaitu Algoritma *Average-link Agglomerative*. Algoritma ini merupakan salah satu algoritma *hierarchical clustering*. Tahapan Algoritma *Average-link Agglomerative* dapat dilihat di (Algoritma 4).

Algorithm 4: Algoritma *Average-Link Agglomerative*

input : data, k
output: clustered_data
while *semua cluster pada clustered_data tamemilikianggota<k* **do**
    hitung rata-rata jarak antar semua baris pada data
    gabungkan baris yang memiliki jarak rata-rata terdekat
    perbaharui rata-rata jarak antar baris
end

Pada algoritma ini, data akan dikelompokkan dengan data lainnya yang memiliki rata-rata jarak terdekat. Perhitungan jarak dihitung menggunakan *Gower Distance* sama seperti pada Algoritma OKA. Saat dipakai untuk anonimisasi dengan metode *k-anonymity*, algoritma ini menghasilkan *cluster* dengan jumlah anggota yang tidak seimbang. Hal ini mengakibatkan hampir semua data yang dianonimisasi menjadi bentuk paling general dan membuat nilai *information loss* menjadi tinggi.

7. **Studi literatur mengenai Algoritma GCCG**
   Status : Baru ditambahkan pada semester ini.
   Hasil : Algoritma GCCG merupakan algoritma *k-anonymity* dengan teknik *clustering*. Algoritma ini menghasilkan *cluster* dengan jumlah yang seimbang daripada Algoritma CURE. Pada algoritma ini dilakukan empat tahapan, yaitu *grading, centering, clustering,* dan *generalization*. Pertama data akan diberi *grade* kemudian diurutkan berdasarkan *grade* tersebut. *Grade* didapat dari perhitungan berdasarkan nilai atribut suatu data. Berikut ini rumus untuk menghitung grade atribut numerik:

$$grade = \frac{a}{\sum_{i=1}^{n} b_i} \tag{6}$$

- a merupakan nilai atribut pada baris yang sedang dihitung *grade*-nya,
- n merupakan banyaknya baris pada data,
- b_i nilai atribut baris ke-i.

Grade untuk atribut kategori dihitung dengan rumus:

$$grade = \frac{\text{count}(x)}{n} \tag{7}$$

- count(x) merupakan banyaknya baris yang memiliki nilai x pada atribut yang sedang dihitung,
- n merupakan banyaknya seluruh baris pada atribut yang sedang dihitung.

Setelah data terurut, data akan dikelompokkan dan akhirnya digeneralisasi. Masukan yang diperlukan algoritma ini adalah dataset, k (konstanta k-anonymity), dan pohon klasifikasi setiap atribut kategori. Setelah algoritma selesai dilakukan, keluaran yang dihasilkan adalah data yang sudah teranonimisasi. Algoritma dapat dilihat pada (Algoritma 5).

Algorithm 5: Algoritma *Grading, Centering, Clustering, Generalization* (GCCG)

input : data, k, classification_tree
output: clustered_data

```
n ← banyak baris pada data
for i = 1 to n do
    hitung grade dari baris ke i
end
urutkan data berdasarkan grade secara terurut menurun
for i = 1 to (n−1)/k do
    pilih baris i sebagai centroid
    kelompokkan baris i dengan k − 2 baris lainnya yang memiliki jarak terdekat ke centroid
    hapus centroid dan baris dari data
end
kelompokkan data yang tersisa dalam satu cluster
for setiap cluster do
    generalisasi cluster menggunakan classification_tree
end
```

8. **Analisis masalah perangkat lunak yang akan dibangun**
   Status : Ada sejak rencana kerja skripsi
   Hasil : Privasi pada teknik *data mining* dapat dilakukan dengan *privacy preserving data mining*. Salah satu metode *privacy preserving data mining* adalah anonimisasi dengan *k-anonymity*. Metode ini akan membuat data sulit dibedakan dengan $k-1$ data lainnya. Metode ini dapat dilakukan dengan berbagai teknik, salah satunya *clustering*. Anonimisasi dengan *k-anonymity* dengan teknik *clustering* dicapai dengan mengelompokkan data dengan jumlah anggota kelompok sebanyak k. Pengelompokkan data dilakukan dengan memanfaatkan algoritma *clustering*. Nilai k yang merupakan konstanta *k-anonymity*, dapat ditentukan oleh pengguna. Setelah data terkelompok, setiap kelompok akan digeneralisasi. Pada perangkat lunak yang akan dibangun, akan digunakan dua algoritma *k-anonymity* dengan teknik *clustering*, yaitu Algoritma OKA dan Algoritma GCCG.

   Kedua algoritma ini dipilih karena kedua algoritma ini dapat dipakai untuk atribut campuran, tidak seperti Algoritma CURE. Selain itu kompleksitas algoritma ini juga cukup baik. Algoritma GCCG memiliki waktu eksekusi dan *information loss* yang lebih baik dari Algoritma KACA dan *Incognito*. Algoritma *Incognito* diketahui memiliki kompleksitas eksponensial. Sedangkan Algoritma OKA memiliki kompleksitas $O(\frac{n^2}{k})$. Kompleksitas dan *information loss* Algoritma OKA diketahui lebih baik daripada Algoritma *Greedy k-Member*. Selain itu, perhitungan jarak yang dipakai pada penelitian ini adalah *Gower Distance*. Hal ini dikarenakan perhitungan jarak *Gower Distance* merupakan perhitungan jarak yang cocok untuk atribut campuran.

   Tahapan yang dilakukan untuk anonimisasi adalah sebagai berikut:

   - Sebelum anonimisasi dilakukan, PII akan dipakai untuk membagi atribut pada data menjadi tiga kategori. *Identifier, quasi-identifier,* dan *sensitive identifier*. Setelah atribut data dibagi, maka data akan diolah. Atribut *identifier* akan dihilangkan dari data.
   - Data akan diumpankan ke algoritma *clustering* yang dipakai, yaitu Algoritma OKA dan Algoritma GCCG.
   - Data yang sudah terbagi ke dalam *cluster* akan digeneralisasi
   - Setelah generalisasi selesai, kualitas tabel anonim dapat dilihat dengan menghitung nilai *information loss*.

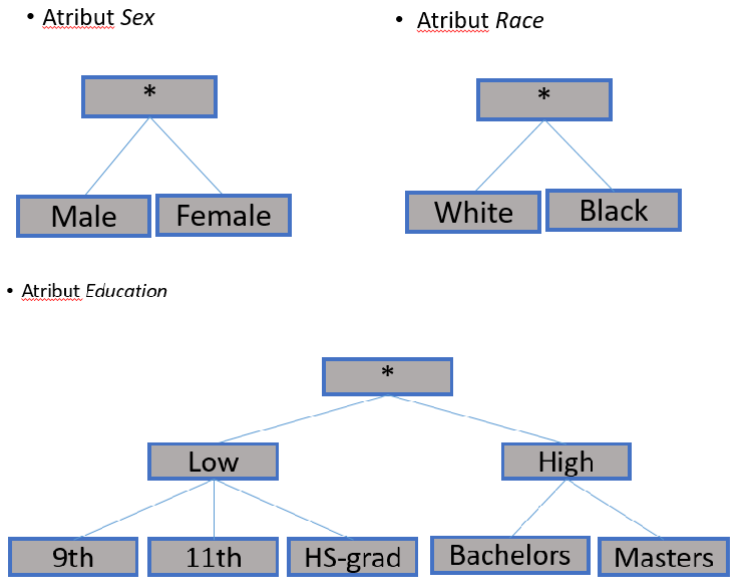
Hasil anonimisasi dapat diuji dengan menggunakan teknik *data mining*. Tabel hasil anonimisasi akan ditambang dengan algoritma *clustering*, kemudian hasil penambangan dapat dibandingkan.

Penelitian ini akan menghasilkan perangkat lunak untuk menganonimisasi data dan membandingkan data (tabel privat dan tabel anonimisasi / tabel anonimisasi Algoritma OKA dan tabel anonimisasi Algoritma GCCG, dan lain-lain).

Agar lebih memahami Algoritma OKA dan Algoritma GCCG untuk anonimisasi, dilakukan studi kasus terhadap set data kecil. Masukan yang dipakai adalah (Tabel 4) sebagai data atau tabel privat, (Gambar 3 sebagai pohon klasifikasi, $k = 2$, *sensitive identifiers*= Workclass dan QI= Sex, Race, Education, Age.

Tabel 4: Tabel Privat untuk Studi Kasus

Race	Sex	Age	Education	Workclass
White	Male	39	Bachelors	State-gov
White	Male	50	Bachelors	Self-emp-not-inc
White	Male	38	HS-grad	Private
Black	Male	53	11th	Private
Black	Female	28	Bachelors	Private
White	Female	37	Masters	Private
Black	Female	49	9th	Private
White	Male	52	HS-grad	Self-emp-not-inc
White	Female	31	Masters	Private
White	Male	42	Bachelors	Private



Gambar 3: Pohon Klasifikasi untuk Studi Kasus

(a) Algoritma OKA
Berikut ini langkah-langkah studi kasus yang dilakukan dengan menggunakan Algoritma OKA:

(1) Baris pada data diurutkan sesuai dengan *grade* yang telah dihitung. Pada (Tabel 5) dapat dilihat hasil perhitungan *grade* dan pada (Tabel 6) dapat dilihat hasil pengurutan.

Tabel 5: Tabel Hasil Perhitungan *Grade*

ID	Race	Sex	Age	Education	Workclass	Grade
1	White	Male	39	Bachelors	State-gov	1.8
2	White	Male	50	Bachelors	Self-emp-not-inc	1.8
3	White	Male	38	HS-grad	Private	1.6
4	Black	Male	53	11th	Private	1.1
5	Black	Female	28	Bachelors	Private	1.2
6	White	Female	37	Masters	Private	1.4
7	Black	Female	49	9th	Private	0.9
8	White	Male	52	HS-grad	Self-emp-not-inc	1.6
9	White	Female	31	Masters	Private	1.4
10	White	Male	42	Bachelors	Private	1.8

Tabel 6: Tabel Hasil Pengurutan

ID	Race	Sex	Age	Education	Workclass	Grade
1	White	Male	39	Bachelors	State-gov	1.8
2	White	Male	50	Bachelors	Self-emp-not-inc	1.8
10	White	Male	42	Bachelors	Private	1.8
3	White	Male	38	HS-grad	Private	1.6
8	White	Male	52	HS-grad	Self-emp-not-inc	1.6
6	White	Female	37	Masters	Private	1.4
9	White	Female	31	Masters	Private	1.4
5	Black	Female	28	Bachelors	Private	1.2
4	Black	Male	53	11th	Private	1.1
7	Black	Female	49	9th	Private	0.9

(2) Dipilih lima baris untuk menjadi *centroid*, yaitu baris dengan ID=$\{2, 3, 8, 5, 4\}$.

(3) Kemudian hitung jarak setiap baris ke setiap *centroid* dan kelompokkan baris dengan *centroid* terdekat. Hasil perhitungan jarak dan *cluster* dapat dilihat pada (Tabel 7).

Tabel 7: Tabel Hasil Perhitungan Jarak dan *Cluster*

ID	c1 (ID=2)	c2(ID=3)	c3(ID=8)	c4(ID=5)	c5(ID=4)	*cluster*
1	1.44	3.04	2.52	3.44	3.56	1
10	1.32	2.16	2.4	2.56	2.4	1
6	3.52	3.04	3.6	2.36	3.6	4
9	3.76	3.28	3.84	2.12	3.84	4
7	4.04	2.44	4.12	1.84	2.12	5

(4) Nilai *centroid* kemudian diperbaharui dengan menghitung baris yang memiliki jarak paling dekat ke semua baris pada *cluster* tersebut. Sehingga didapatkan *centroid* baru $C = \{10, 3, 8, 9, 4\}$

(5) Baris yang memiliki jarak terjauh dari *centroid*-nya pada *cluster* dengan anggota lebih dari k, dikeluarkan dari *cluster* tersebut. Terdapat dua *cluster* dengan anggota lebih dari k, yaitu *cluster* 1 dan 4. Pada *cluster* 1, baris dengan ID=2 merupakan baris yang paling jauh dari *centroid*. Sedangkan pada *cluster* 4, baris dengan ID=5 merupakan baris terjauh.

(6) Baris yang sudah dikeluarkan dari *centroid* dihitung jaraknya ke *centroid* dari *cluster* yang anggotanya kurang dari k. Baris akan dikelompokkan bersama *centroid* yang memiliki jarak terdekat. Tahap ini merupakan tahap *adjustment*. Pada kasus ini *cluster* yang anggotanya kurang dari k adalah *cluster* 2 dan 3. Hasil *adjustment* dapat dilihat pada (Tabel **??**).

Tabel 8: Tabel Hasil *Adjustment*

ID	Jarak dengan C2	Jarak dengan C3	*Cluster*
2	2.48	1.08	3
5	3.4	4.96	2

(7) Setelah semua perhitungan dilakukan, didapatkan pembagian *cluster* seperti pada (Table 9).

Tabel 9: Tabel Hasil *Clustering*

ID	Race	Sex	Age	Education	Workclass	Cluster
10	White	Male	42	Bachelors	Private	1
1	White	Male	39	Bachelors	State-gov	1
9	White	Female	31	Masters	Private	2
6	White	Female	37	Masters	Private	2
3	White	Male	38	HS-grad	Private	3
5	Black	Female	28	Bachelors	Private	3
8	White	Male	52	HS-grad	Self-emp-not-inc	4
2	White	Male	50	Bachelors	Self-emp-not-inc	4
4	Black	Male	53	11th	Private	5
7	Black	Female	49	9th	Private	5

(8) Setelah terbagi ke dalam *cluster*, data akan digeneralisasi. Akhirnya didapatkan tabel anonim. Tabel anonim dari hasil generalisasi dapat dilihat pada (Tabel 10).

Tabel 10: Tabel Anonim Algoritma OKA

Race	Sex	Age	Education	Workclass
White	Male	[39 - 42]	Bachelors	Private
White	Male	[39 - 42]	Bachelors	State-gov
White	Female	[31 - 37]	Masters	Private
White	Female	[31 - 37]	Masters	Private
*	*	[28 - 38]	*	Private
*	*	[28 - 38]	*	Private
White	Male	[50 - 52]	*	Self-emp-not-inc
White	Male	[50 - 52]	*	Self-emp-not-inc
Black	*	[49 - 53]	Low	Private
Black	*	[49 - 53]	Low	Private

(b) Algoritma GCCG

Berikut ini langkah-langkah studi kasus yang dilakukan dengan menggunakan Algoritma GCCG.

(1) Pertama, setiap baris pada data akan dihitung *grade*-nya. Tahap ini merupakan tahapan *grading*. Hasil perhitungan *grade* dapat dilihat pada (Tabel 5).

(2) Setelah *grading*, dilakukan tahap *centering*. Pada tahap ini data akan diurutkan berdasarkan *grade*. Hasil pengurutan ini dapat dilihat pada (Tabel 6).

(3) Dari tabel yang sudah terurut, diambil baris pertama dan $k-1$ baris terdekat dari baris tersebut untuk dijadikan satu *cluster*. Jadi pada awalnya, baris dengan ID=1 diambil dan dikelompokkan bersama baris dengan ID=3 yang merupakan baris terdekat. Hal ini dilakukan terus sebanyak $\frac{n-1}{k}$. Dari tahap ini, didapatkan empat *cluster* C=$\{(ID = 1, ID = 10), (ID = 2, ID = 8), (ID = 3, ID = 6), (ID = 9, ID = 5)\}$.

(4) Baris yang belum dikelompokkan akan langsung dijadikan satu *cluster*, sehingga terdapat satu *cluster* tambahan X=$(ID = 4, ID = 7)$.

(5) Setelah semua tahap *clustering*, dilakukan tahap generalisasi. Dari hasil generalisasi didapatkanlah tabel anonim. Tabel anonim dapat dilihat pada (Tabel 11).

Tabel 11: Tabel Anonim Algoritma GCCG

Race	Sex	Age	Education	Workclass
White	Male	[39 - 42]	Bachelors	State-gov
White	Male	[39 - 42]	Bachelors	Private
White	Male	[50 - 52]	*	Self-emp-not-inc
White	Male	[50 - 52]	*	Self-emp-not-inc
White	*	[37 - 38]	*	Private
White	*	[37 - 38]	*	Private
*	Female	[28 - 31]	High	Private
*	Female	[28 - 31]	High	Private
Black	*	[49 - 53]	Low	Private
Black	*	[49 - 53]	Low	Private

Terdapat dua diagram aktivitas pada penelitian ini, yaitu diagram aktivitas proses anonimisasi data dan diagram aktivitas proses analisis. Sedangkan pengujian dengan teknik *data mining* akan menggunakan perangkat lunak lainnya yang melakukan *clustering*, seperti Weka. Diagram aktivitas untuk proses anonimisasi dapat dilihat pada (Gambar 4).

Detail proses anonimisasi pada (Gambar 4) adalah sebagai berikut:

(1) Perangkat lunak menerima masukan dari pengguna berupa data yang akan dianonimisasi dan pohon_klasifikasi.

(2) Perangkat lunak menampilkan kembali data yang sudah dimasukkan pengguna untuk ditinjau ulang.

(3) Perangkat lunak menerima masukan dari pengguna mengenai tipe dari setiap atribut (numerik / kategori).

(4) Perangkat lunak menerima masukan dari pengguna berupa nilai k (konstanta *k-anonymity* dan algoritma yang ingin dipakai pengguna

(5) Perangkat lunak mengolah atribut sesuai dengan jenisnya. Jika atribut merupakan *identifier* maka atribut akan langsung dihilangkan dari tabel.

(6) Perangkat menjalankan proses anonimisasi dengan algoritma yang sebelumnya sudah dipilih

(7) Perangkat lunak menghitung *information loss* yang dihasilkan serta waktu eksekusi algoritma.

(8) Perangkat lunak menampilkan tabel hasil anonimisasi kepada pengguna beserta dengan informasi tambahan lainnya, seperti *information loss* dan waktu eksekusi.

(9) Jika pengguna ingin menyimpan hasil anonimisasi, maka perangkat lunak akan membuat *file* untuk menyimpan hasil anonimisasi. Jika pengguna tidak ingin menyimpan hasil anonimisasi, maka proses selesai.

Sedangkan diagram aktivitas proses analisis dapat dilihat pada (Gambar 5).

Detail proses analisis pada (Gambar 5) adalah sebagai berikut:

(1) Perangkat lunak menerima masukan dari pengguna berupa data yang akan dibandingkan. Data dapat berupa tabel privat maupun tabel hasil anonimisasi.

(2) Perangkat lunak menerima masukan dari pengguna berupa atribut apa saja yang akan dibandingkan.

(3) Perangkat lunak menampilkan hasil perbandingan data yang diinginkan oleh pengguna.

Diagram kelas untuk perangkat lunak yang akan dibangun dapat dilihat pada (Gambar 6).

Perangkat lunak ini akan memiliki kelas dan *interface* sebagai berikut:

(a) *KAnonymizer* merupakan kelas abstrak yang nantinya akan di-*extend* oleh kelas yang mengimplementasikan algoritma *k-anonymity*. Kelas ini memiliki empat atribut, yaitu:

- k, nilai konstanta *k-anonymity*.
- data, tabel privat/data yang ingin dianonimisasi.
- *classification_tree*, *tree* yang menyatakan pohon klasifikasi untuk generalisasi.
- *anonymize_data*, data hasil anonimisasi.

Kelas ini juga memiliki tiga *method*, yaitu:

- *anonymized()* merupakan *method* untuk melakukan anonimisasi data. Pada kelas ini, *method anonymized()* masih abstrak.
- *get_anonymized_data()* merupakan *method* yang mengembalikan tabel hasil anonimisasi.
- *count_information_loss()* merupakan *method* untuk menghitung nilai *information loss*.

(b) *Clusterizer* merupakan *interface* yang merepresentasikan pembuat *cluster*. *Interface* ini akan diimplementasi oleh *OKAAnonymizer* dan *GCCGAnonymizer* yang merupakan *k-anonymizer* dengan teknik *clustering*. *Interface* ini memiliki satu *method*, yaitu *clustering* yang merupakan *method* untuk melakukan teknik *clustering*.

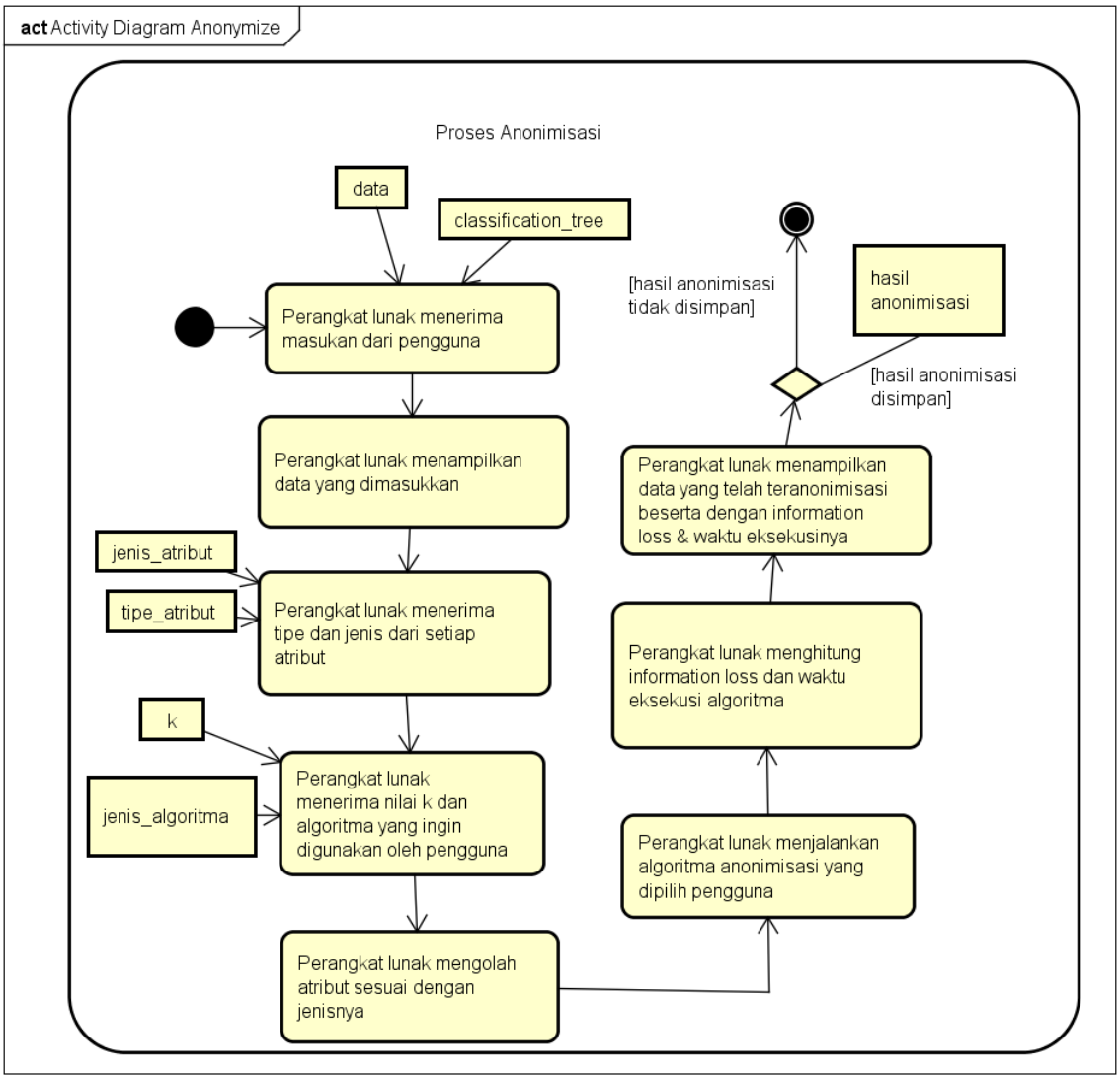
(c) *OKAAnonymizer* merupakan kelas yang mengimplementasikan Algoritma OKA untuk anonimisasi dengan teknik *clustering*. Oleh karena itu kelas ini mengimplementasi *interface Clusterizer* dan *extend* kelas *KAnonymizer*. Kelas ini memiliki satu atribut tambahan, yaitu *sorted_data* yang merupakan data hasil pengurutan. Kelas ini juga memiliki tiga *method* selain *method* yang diturunkan dari *KAnonymizer* dan *Clusterizer*, yaitu:

- *sort()* merupakan *method* untuk mengurutkan data.
- *adjust()* merupakan *method* untuk melakukan tahap *adjustment* pada Algoritma OKA.
- *get_sorted_data()* merupakan *method* yang mengembalikan data yang sudah diurutkan.

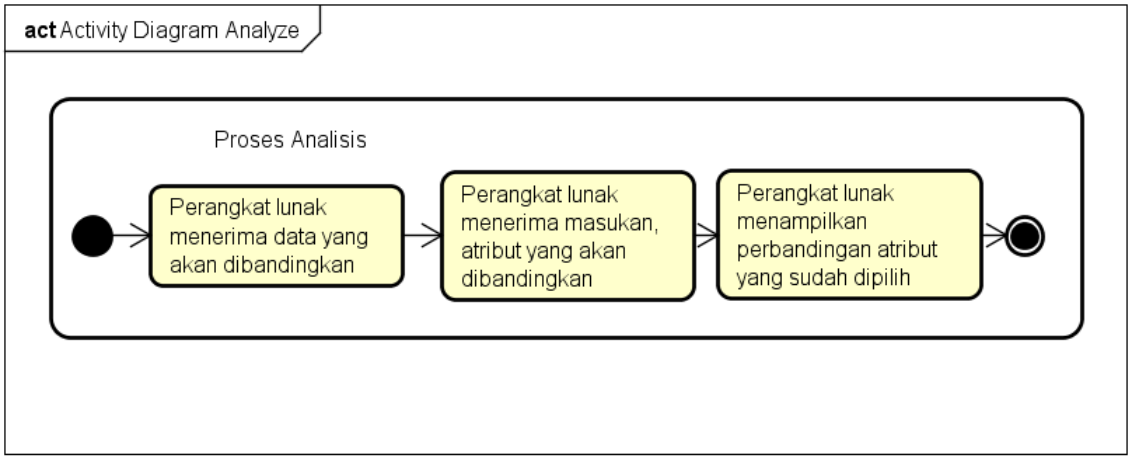
(d) *GCCGAnonymizer* merupakan kelas yang mengimplementasikan Algoritma GCCG untuk anonimisasi dengan teknik *clustering*. Oleh karena itu kelas ini mengimplementasi *interface Clusterizer* dan *extend* kelas *KAnonymizer*. Kelas ini memiliki satu atribut tambahan, yaitu *sorted_data* yang merupakan data hasil pengurutan. Kelas ini juga memiliki dua *method* selain *method* yang diturunkan dari *KAnonymizer* dan *Clusterizer*, yaitu:

- *grading()* yang merupakan *method* untuk menghitung *grade* tiap baris pada data.
- *get_sorted_data()* merupakan *method* yang mengembalikan data yang sudah diurutkan.

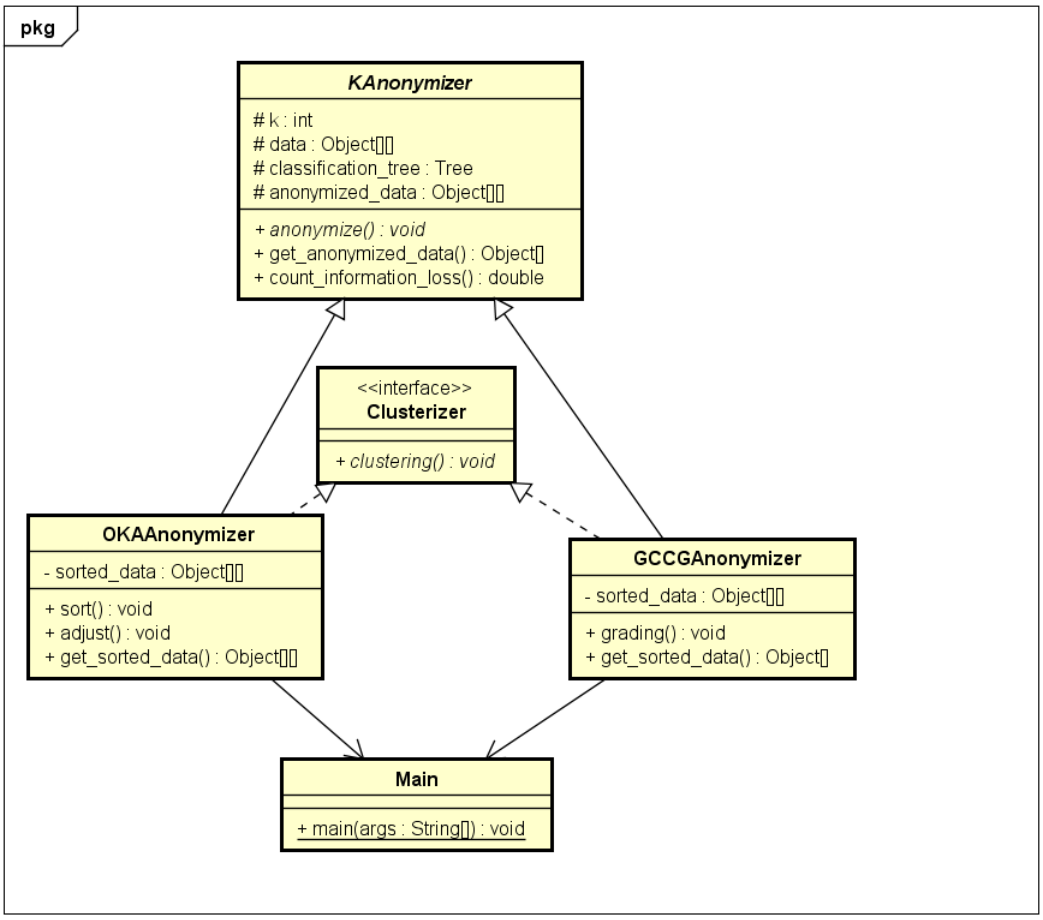
(e) *Main* merupakan kelas yang dipanggil terlebih dahulu dan menjembatani antara pengguna dengan kelas-kelas yang menyelesaikan masalah anonimisasi.



Gambar 4: Diagram Aktivitas Proses Anonimisasi



Gambar 5: Diagram Aktivitas Proses Analisis



Gambar 6: Diagram Kelas Perangkat Lunak yang Akan Dibangun

6 Pencapaian Rencana Kerja

Langkah-langkah kerja yang berhasil diselesaikan dalam Skripsi 1 ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai privasi dan *personally identifiable information* (PII).
2. Melakukan studi literatur mengenai teknik *data mining* dan *privacy preserving data mining*.
3. Melakukan studi literatur mengenai metode *k-Anonymity* dengan teknik *clustering*.
4. Melakukan studi literatur mengenai Algoritma OKA.
5. Melakukan studi literatur mengenai Algoritma CURE.
6. Melakukan studi literatur mengenai Algoritma *Average-link Agglomerative*.
7. Melakukan studi literatur mengenai Algoritma GCCG.
8. Analisis masalah perangkat lunak yang akan dibangun.
9. Menulis dokumen skripsi (sebagian bab 1, bab 2 dan bab 3).

7 Kendala yang Dihadapi

Kendala - kendala yang dihadapi selama mengerjakan skripsi :

- Banyak tugas mata kuliah lain yang harus dikerjakan.

Bandung, 22/11/2019

Apsari Ayusya Cantika

Menyetujui,

Nama: Mariskha Tri Adithia
Pembimbing Tunggal