SKRIPSI

ANALISIS ALGORITMA *ONE PASS K-MEANS* DAN *GRADING, CENTERING, CLUSTERING, GENERALIZATION* UNTUK ANONIMISASI DATA

Apsari Ayusya Cantika

NPM: 2016730012

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

UNDERGRADUATE THESIS

ANALYSIS OF ONE PASS K-MEANS AND GRADING, CENTERING, CLUSTERING, GENERALIZATION ALGORITHM FOR ANONYMIZING DATA

Apsari Ayusya Cantika

NPM: 2016730012

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY

2020

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS ALGORITMA *ONE PASS K-MEANS* DAN *GRADING*, *CENTERING*, *CLUSTERING*, *GENERALIZATION* UNTUK ANONIMISASI DATA

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 11 Juni 2020

Apsari Ayusya Cantika
NPM: 2016730012

LEMBAR PENGESAHAN

ANALISIS ALGORITMA *ONE PASS K-MEANS* DAN *GRADING, CENTERING, CLUSTERING, GENERALIZATION* UNTUK ANONIMISASI DATA

Apsari Ayusya Cantika

NPM: 2016730012

Bandung, 11 Juni 2020

Menyetujui,

Pembimbing

Mariskha Tri Adithia, P.D.Eng

Ketua Tim Penguji	**Anggota Tim Penguji**
Husnul Hakim, M.T.	**Natalia, M.Si.**

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS ALGORITMA *ONE PASS K-MEANS* DAN *GRADING, CENTERING, CLUSTERING, GENERALIZATION* UNTUK ANONIMISASI DATA

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 11 Juni 2020

Apsari Ayusya Cantika
NPM: 2016730012

ABSTRAK

Di era digital ini, teknik *data mining* semakin banyak digunakan. Teknik ini membuat data harus dirilis. Pada data yang dirilis, terdapat kemungkinan adanya data atau informasi pribadi seseorang. Hal ini dapat menyebabkan privasi tidak terlindungi. Privasi merupakan kemampuan seseorang untuk mengatur bagaimana informasi pribadinya disimpan, dipakai, maupun dihapus. Privasi dapat dilindungi dengan *privacy preserving data mining* (PPDM). PPDM merupakan bagian dari *data mining* yang bertanggung jawab atas perlindungan privasi dalam proses *data mining*. Salah satu metode PPDM adalah anonimisasi yang membuat data menjadi anonim. Metode *k-anonymity* merupakan salah satu contoh metode anonimisasi. Tujuan dari penelitian ini adalah melakukan analisis algoritma *k-anonymity* untuk anonimisasi data. Algoritma *k-anonymity* yang akan digunakan adalah Algoritma *One Pass k-Means* (OKA) dan *Grading, Centering, Clustering, Generalization* (GCCG).

Pada penelitian ini, dibangun dua buah perangkat lunak. Perangkat lunak pertama mengimplementasikan Algoritma OKA dan GCCG, sedangkan perangkat lunak kedua mengimplementasikan algoritma *data mining* untuk menguji hasil anonimisasi perangkat lunak pertama. Hasil anonimisasi dari perangkat lunak pertama dan hasil pengujian dari perangkat lunak kedua dipakai untuk analisis. Analisis dilakukan dengan pengujian eksperimental. Pengujian eksperimental dilakukan untuk mendapatkan relasi antara nilai k, *information loss*, waktu eksekusi algoritma yang diimplementasikan, jenis atribut, hasil *data mining*, evaluasi hasil *data mining*, dan jumlah atribut.

Hasil pengujian menunjukan bahwa nilai k memengaruhi *information loss* dan waktu eksekusi kedua algoritma. Semakin tinggi nilai k, nilai *information loss* akan semakin tinggi dan waktu eksekusi semakin cepat. Algoritma GCCG memiliki waktu eksekusi yang lebih cepat dibandingkan Algoritma OKA. Saat data yang dianonimisasi diuji dengan teknik *data mining*, dapat dilihat bahwa jumlah atribut memengaruhi kualitas hasil *clustering*. Semakin banyak jumlah atribut, maka kualitas hasil *clustering* akan semakin menurun. Sementara itu, kualitas hasil klasifikasi tidak dipengaruhi oleh jumlah atribut. Kualitas hasil klasifikasi tidak menentu.

Kata-kata kunci: Privasi, *Data Mining, Privacy Preserving Data Mining, K-Anonymity, One Pass k-Means,* GCCG

ABSTRACT

In this digital age, the usage of data mining becomes more frequent. This technique makes data needs to be released. There might be personal data or information in data that are released. This thing can make privacy unprotected. Privacy is someone's ability to manage how their personal information is stored, used, or deleted. Privacy can be protected with privacy preserving data mining (PPDM). PPDM is part of data mining that is responsible for protecting privacy on the data mining process. One of the PPDM methods is anonymization that makes the data anonymous. K-anonymity is an example of anonymization methods. The purpose of this study is to analyze the k-anonymity algorithm for anonymizing data. The k-anonymity algorithms that will be used in this study are One Pass k-Means (OKA) and Grading, Centering, Clustering, Generalization (GCCG).

In this study, two software was built. The first software implements OKA and GCCG Algorithm. Meanwhile, the second software implements data mining algorithm to test the first software's anonymization result. The first software's anonymization result and second software's data mining results were used for analysis. The analysis is done with a test. The test was done to get relations between k value, information loss, the implemented algorithm's execution time, type of attribute, data mining result, evaluation of data mining result, and amount of attribute.

The test result showed that k value affects both algorithm's anonymization quality and execution time. The higher k values, the higher information loss will be and execution time will be faster. GCCG Algorithm has a faster execution time than OKA Algorithm. When anonymized data tested with data mining techniques, it can be seen that amount of attribute affects the quality of clustering result. The more amount of attribute, the quality of clustering result will be lower. Meanwhile, the quality of classification result is not affected by the amount of attribute. The quality of classification result is uncertain.

Keywords: Privacy, Data Mining, Privacy Preserving Data Mining, k-Anonymity, One Pass k-Means, GCCG

Dipersembahkan untuk Tuhan YME, keluarga, para dosen, teman-teman yang telah memberi dukungan dalam pembuatan skripsi ini, serta diri sendiri.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa, karena dengan rahmat dan karunia-Nya, penulis dapat menyelesaikan penyusunan skripsi berjudul "Analisis Algoritma *One Pass k-Means dan Grading, Centering, Clustering, Generalization* untuk Anonimisasi Data". Skripsi ini dibuat dan diajukan untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Universitas Katolik Parahyangan. Selain itu, penulisan skripsi ini bertujuan untuk memberikan pengetahuan kepada pembaca mengenai *privacy preserving data mining* metode *k-anonymity* dengan teknik *clustering*. Selama penulisan skripsi ini, penulis menyadari bahwa penulisan skripsi ini dapat selesai karena bantuan dan dukungan beberapa pihak. Oleh karena itu, penulis mengungkapkan rasa terima kasih kepada:

1. Ibu Mariskha Tri Adithia, S.Si., M.Sc., PDEng. selaku dosen pembimbing yang telah membimbing dan mendukung penulis selama proses penyusunan skripsi ini.
2. Bapak Husnul Hakim, S.Kom., M.T., dan Ibu Natalia, M.Si. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun.
3. Keluarga yang selalu memberikan doa dan dukungan.
4. Ko Edrick dan Ko Cornel, yang selalu menjadi tempat penulis bertanya mengenai skripsi.
5. Teman-teman seperjuangan skripsi, khususnya teman dengan pembimbing sama (Kevin Arnold dan Chris Eldon) dan teman-teman grup kacang (Ferdian, JL, Cahyadi, Kikil, R, dan JY).
6. Sahabat baik penulis, yaitu Sasa, Vina, Edo, KN, serta Enrico yang selalu mendukung dan menyemangati penulis.

Penulis menyadari bahwa penelitian ini masih jauh dari kata sempurna. Oleh karena itu, penulis memohon maaf jika terdapat kekurangan pada penelitian ini. Penulis juga mengharapkan kritik dan saran yang membangun untuk menyempurnakan penelitian ini. Semoga penelitian ini dapat bermanfaat bagi segenap pihak yang berkepentingan.

Bandung, Juni 2020

Penulis

DAFTAR ISI

Kata Pengantar xv

Daftar Isi xvii

Daftar Gambar xix

Daftar Tabel xxi

Daftar Kode Program xxiii

1 Pendahuluan 1
- 1.1 Latar Belakang . . . 1
- 1.2 Rumusan Masalah . . . 2
- 1.3 Tujuan . . . 2
- 1.4 Batasan Masalah . . . 2
- 1.5 Metodologi . . . 3
- 1.6 Sistematika Pembahasan . . . 3

2 Landasan Teori 5
- 2.1 Privasi dan *Personally Identifiable Information* (PII) . . . 5
- 2.2 *Data Mining* . . . 6
  - 2.2.1 *Clustering* . . . 6
  - 2.2.2 Klasifikasi . . . 7
- 2.3 Evaluasi Hasil *Clustering* dan Klasifikasi . . . 7
  - 2.3.1 Koefisien Silhouette . . . 7
  - 2.3.2 Tingkat Akurasi Klasifikasi . . . 8
- 2.4 Koefisien Korelasi Spearman . . . 8
- 2.5 *Library* Python . . . 8
  - 2.5.1 *Library Sorting* . . . 8
  - 2.5.2 *Library Data Mining (KMeans dan DecisionTreeClassifier)* . . . 9
  - 2.5.3 *Library* Evaluasi *Data Mining (silhouette_score* dan *accuracy_score)* . . . 9
  - 2.5.4 *Library* Koefisien Korelasi Spearman . . . 9
- 2.6 *Privacy Preserving Data Mining* (PPDM) . . . 9
- 2.7 *k-Anonymity* . . . 10
  - 2.7.1 Algoritma *One Pass k-Means* (OKA) . . . 11
  - 2.7.2 Algoritma *Grading, Centering, Clustering, Generalization* (GCCG) . . . 13
- 2.8 *Information Loss* . . . 13
- 2.9 Penyimpanan Data . . . 14
  - 2.9.1 JSON . . . 14
  - 2.9.2 CSV . . . 15

3 Analisis 17
- 3.1 Analisis Masalah . . . 17

3.2 Studi Kasus . . . 18
3.3 Gambaran Umum Perangkat Lunak . . . 22
3.3.1 Diagram Aktivitas . . . 23
3.3.2 Diagram Kelas . . . 26

4 Perancangan **29**
4.1 Perancangan Antarmuka . . . 29
4.1.1 Perancangan Antarmuka Perangkat Lunak Anonimisasi . . . 29
4.1.2 Perangcangan Antarmuka Perangkat Lunak Pengujian . . . 32
4.2 Diagram Kelas Lengkap . . . 34
4.3 Masukan Perangkat Lunak . . . 40

5 Implementasi dan Pengujian **43**
5.1 Implementasi Antarmuka . . . 43
5.1.1 Implementasi Antarmuka Perangkat Lunak Anonimisasi . . . 43
5.1.2 Implementasi Antarmuka Perangkat Lunak Pengujian . . . 46
5.2 Pengujian . . . 47
5.2.1 Pengujian Fungsional . . . 47
5.2.2 Pengujian Eksperimental . . . 50

6 Kesimpulan dan Saran **61**
6.1 Kesimpulan . . . 61
6.2 Saran . . . 62

Daftar Referensi **63**

A *File* Masukan untuk Pengujian Fungsional **65**
A.1 *File CSV* . . . 65
A.2 *File* JSON . . . 65

B Kode Program Perangkat Lunak Anonimisasi **69**
B.1 Kelas Abstrak KAnonymizer . . . 69
B.2 *Interface* Clusterizer . . . 70
B.3 Kelas OKAAnonymizer . . . 70
B.4 Kelas GCCGAnonymizer . . . 72
B.5 Kelas UIController . . . 73

C Kode Program Perangkat Lunak Pengujian **77**
C.1 Kelas UIController . . . 77

DAFTAR GAMBAR

2.1 Informasi dan PII 5
2.2 Contoh Pohon Klasifikasi 10

3.1 Pohon Klasifikasi untuk Studi Kasus 19
3.2 Diagram Aktivitas Proses Anonimisasi 24
3.3 Diagram Aktivitas Proses Pengujian 25
3.4 Diagram Kelas Perangkat Lunak yang Akan Dibangun 26

4.1 Tampilan Antarmuka Masukan Perangkat Lunak Anonimisasi 29
4.2 Tampilan saat *Dropdown* Perangkat Lunak Anonimisasi Ditekan 30
4.3 Tampilan Antarmuka Data Perangkat Lunak Anonimisasi 31
4.4 Tampilan Antarmuka Hasil Perangkat Lunak Anonimisasi 31
4.5 Tampilan Antarmuka *Menu* untuk Menyimpan Tabel 32
4.6 Tampilan Antarmuka Masukan Perangkat Lunak Pengujian 32
4.7 Tampilan saat *Dropdown* Perangkat Lunak Pengujian Ditekan 33
4.8 Tampilan Antarmuka Data Perangkat Lunak Pengujian 34
4.9 Tampilan Antarmuka Hasil Perangkat Lunak Pengujian 34
4.10 Diagram Kelas Lengkap 35

5.1 Tampilan Utama Perangkat Lunak Anonimisasi 43
5.2 Tampilan *Tab* "Data" Perangkat Lunak Anonimisasi 45
5.3 Tampilan *Dropdown* Perangkat Lunak Anonimisasi 46
5.4 Tampilan *Tab* "Hasil" Perangkat Lunak Anonimisasi 46
5.5 Tampilan Pilihan untuk Menyimpan Hasil Anonimisasi pada *Tab* "Hasil" 47
5.6 Tampilan Utama Perangkat Lunak Pengujian 47
5.7 Tampilan *Tab* "Data" Perangkat Lunak Pengujian 48
5.8 Tampilan *Dropdown* Perangkat Lunak Pengujian 48
5.9 Tampilan *Tab* "Hasil" Perangkat Lunak Pengujian 49
5.10 Hasil Anonimisasi Data Perangkat Lunak Anonimisasi dengan Algoritma OKA . . 49
5.11 Hasil Anonimisasi Data Perangkat Lunak Anonimisasi dengan Algoritma GCCG . 50
5.12 Grafik Perbandingan *Information Loss* Data "ADULT" Atribut Kategori 52
5.13 Grafik Perbandingan Waktu Eksekusi Data "ADULT" Atribut Kategori 53
5.14 Grafik Perbandingan *Information Loss* Data "ADULT" Atribut Campuran 53
5.15 Grafik Perbandingan Waktu Eksekusi Data "ADULT" Atribut Campuran 54
5.16 Grafik Perbandingan *Information Loss* Data "Heart Disease" 55
5.17 Grafik Perbandingan Waktu Eksekusi Data "Heart Disease" 55
5.18 Grafik Perbandingan *Information Loss* Algoritma OKA Menggunakan Ketiga Data 56
5.19 Grafik Perbandingan Waktu Eksekusi Algoritma OKA Menggunakan Ketiga Data 56
5.20 Grafik Perbandingan *Information Loss* Algoritma GCCG Menggunakan Ketiga Data 57
5.21 Grafik Perbandingan Waktu Eksekusi Algoritma GCCG Menggunakan Ketiga Data 57
5.22 Grafik Perbandingan Koefisien Silhouette Hasil *Clustering* Ketiga Data 58
5.23 Grafik Perbandingan Hasil *Clustering* Data Sebelum dan Sesudah Anonimisasi . . 58
5.24 Grafik Perbandingan Tingkat Akurasi Hasil Klasifikasi Ketiga Data 59

5.25 Grafik Perbandingan Hasil Klasifikasi Data Sebelum dan Sesudah Anonimisasi . . 59

DAFTAR TABEL

1.1 Tabel Hasil Anonimisasi *k-anonymity* dengan $k = 2$ 2

2.1 Tabel Contoh Generalisasi Atribut Numerik 10
2.2 Tabel Contoh Generalisasi Atribut Kategori 10

3.1 Tabel Privat untuk Studi Kasus 18
3.2 Tabel Hasil Perhitungan *Grade* 19
3.3 Tabel Hasil Pengurutan 20
3.4 Tabel Hasil Perhitungan Jarak dan *Cluster* 20
3.5 Tabel Hasil *Adjustment* 20
3.6 Tabel Hasil *Clustering* 21
3.7 Tabel Anonim Algoritma OKA 21
3.8 Tabel Anonim Algoritma GCCG 23

5.1 Tabel Koefisien Korelasi Spearman 51
5.2 Sampel Data Privat Pertama ("ADULT" Atribut Kategori) 52
5.3 Sampel Data Privat Kedua ("ADULT" Atribut Campuran) 52
5.4 Sampel Data Privat Ketiga ("Heart Disease") 54

DAFTAR KODE PROGRAM

2.1 Contoh Struktur JSON 15
2.2 Contoh CSV 15

4.1 Format *File* .csv untuk Masukan Perangkat Lunak Anonimisasi dan Pengujian . . 40
4.2 Contoh *file* .csv untuk Masukan Perangkat Lunak Anonimisasi dan Pengujian . . . 40
4.3 Spesifikasi *file* .json untuk masukan perangkat lunak anonimisasi 40
4.4 Contoh *file* .json untuk masukan perangkat lunak anonimisasi 41

5.1 Contoh *File* .csv 43
5.2 Contoh *file* .json untuk masukan perangkat lunak anonimisasi 44

A.1 *File* .csv yang digunakan untuk pengujian 65
A.2 *File* .json yang digunakan untuk pengujian 65

B.1 KAnonymizer.py 69
B.2 Clusterizer.py 70
B.3 OKAAnonymizer.py 70
B.4 GCCGAnonymizer.py 72
B.5 UIController.py 73

C.1 UIController.py 77

BAB 1

PENDAHULUAN

1.1 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 [1]. 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 [2]. Contoh PII adalah data seperti nama lengkap, nomor kependudukan, dan lain-lain. Hal ini membuat privasi menjadi sesuatu yang penting dan perlu dilindungi pada saat *data mining* dilakukan. Privasi pada saat dilakukan *data mining* dapat dilindungi dengan *privacy preserving data mining* (PPDM). PPDM adalah bagian dari *data mining* yang bertanggung jawab atas perlindungan privasi dalam proses *data mining* [3]. Dengan adanya PPDM, informasi bisa didapatkan tanpa perlu merilis data mentah. Sebagian besar metode PPDM melakukan transformasi pada data yang akan ditambang sehingga data tidak terbuka seluruhnya.

PPDM diklasifikasikan menjadi beberapa kategori, contohnya randomisasi dan anonimisasi[3]. Randomisasi merupakan metode PPDM yang menambahkan distorsi pada data. Namun, distorsi yang ditambahkan harus cukup besar untuk menutupi data, terutama data sensitif. Sedangkan anonimisasi adalah metode PPDM yang membuat data menjadi anonim. 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 [4]. *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 1.1. Dari tabel, dapat dilihat bahwa baris pertama dan baris kedua memiliki nilai yang sama pada setiap atribut sehingga kedua baris sulit dibedakan. Begitu juga dengan baris-baris berikutnya. Setiap baris sulit dibedakan dengan $k-1$ baris lainnya.

Proses anonimisasi akan menghasilkan *information loss*. Nilai *information loss* merupakan ukuran yang menunjukkan banyaknya informasi yang hilang setelah data dianonimisasi. Nilai ini dapat dipakai untuk analisis hasil anonimisasi. Selain itu, hasil anonimisasi juga dapat diuji dengan teknik *data mining* seperti *clustering* yang mengelompokkan data ke dalam kelompok atau *cluster*, dan klasifikasi yang mencari kelompok yang sesuai untuk data. Hasil *clustering* dan klasifikasi dapat dipakai untuk analisis dengan melihat nilai evaluasinya, yaitu tingkat akurasi klasifikasi dan

Tabel 1.1: 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

koefisien silhouette *clustering*. Tingkat akurasi menunjukkan seberapa akurat data diklasifikasikan, sedangkan koefisien silhouette menunjukan seberapa baik data dikelompokkan.

Pada penelitian ini, 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. Setelah proses anonimisasi dilakukan, akan dilakukan analisis terhadap hasil anonimisasi. Hasil anonimisasi juga diuji dengan perangkat lunak yang mengimplementasikan algoritma *data mining*, dan akan dilakukan analisis terhadap hasil *data mining* tersebut.

1.2 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 perbandingan performa Algoritma OKA dan GCCG untuk anonimisasi?

1.3 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. Membandingkan performa Algoritma OKA dan GCCG.

1.4 Batasan Masalah

Batasan masalah untuk penelitian ini adalah sebagai berikut:

1. Perangkat lunak yang dibuat hanya dapat menerima *file* masukan yang valid (tidak ada kesalahan penulisan data di dalam *file*), karena perangkat lunak yang dibangun hanya sebagai alat untuk membantu penelitian ini, sehingga perangkat lunak tidak dibangun untuk digunakan secara umum.

2. Jumlah baris maksimal pada tabel privat untuk dianonimisasi perangkat lunak adalah 100 baris, karena perhitungan jarak antardata memiliki kompleksitas kuadratik.

1.5 Metodologi

Metodologi yang digunakan dalam penelitian ini adalah:

1. Melakukan studi literatur mengenai *privacy preserving data mining*, metode *k-anonymity* dengan teknik *clustering*, dan teknik *Data Mining*
2. Melakukan studi literatur mengenai Algoritma OKA dan Algoritma GCCG
3. Analisis masalah perangkat lunak yang akan dibangun
4. Merancang perangkat lunak yang akan dibangun
5. Membangun perangkat lunak yang mengimplementasikan Algoritma OKA dan GCCG
6. Melakukan pengujian fungsional dan eksperimental
7. Analisis hasil pengujian dan mengambil kesimpulan

1.6 Sistematika Pembahasan

Laporan penelitian tersusun ke dalam enam bab secara sistematis sebagai berikut:

- Bab 1 Pendahuluan
  Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian dan sistematika pembahasan.
- Bab 2 Landasan Teori
  Berisi landasan teori mengenai dasar-dasar dari privasi dan PII, *data mining*, *library* Python, koefisien korelasi, *privacy preserving data mining*, *k-anonymity* dengan teknik *clustering* dan penyimpanan data.
- Bab 3 Analisis
  Berisi analisis masalah, studi kasus dan gambaran umum perangkat lunak (diagram aktivitas dan diagram kelas).
- Bab 4 Perancangan
  Berisi perancangan perangkat lunak yang akan dibangun, meliputi perancangan antarmuka, diagram kelas lengkap dan masukan perangkat lunak.
- Bab 5 Implementasi dan Pengujian
  Berisi implementasi antarmuka perangkat lunak, pengujian fungsional, pengujian eksperimental serta kesimpulan dari pengujian.
- Bab 6 Kesimpulan dan Saran
  Berisi kesimpulan dari awal hingga akhir penelitian dan saran untuk penelitian berikutnya.