SKRIPSI

PENERAPAN ALGORITMA GREEDY K-MEMBER CLUSTERING UNTUK ANONIMISASI DATA PADA LINGKUNGAN BIG DATA

Stephen Jordan

NPM: 2016730018

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2021

UNDERGRADUATE THESIS

APPLICATION OF GREEDY K-MEMBER CLUSTERING ALGORITHM FOR DATA ANONYMIZATION IN BIG DATA ENVIRONMENT

Stephen Jordan

NPM: 2016730018

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

2021

LEMBAR PENGESAHAN

PENERAPAN ALGORITMA GREEDY K-MEMBER CLUSTERING UNTUK ANONIMISASI DATA PADA LINGKUNGAN BIG DATA

Stephen Jordan

NPM: 2016730018

Bandung, 29 Januari 2021

Menyetujui,

Pembimbing Utama	Pembimbing Pendamping
Mariskha Tri Adithia, S.Si, MSc, PDEng	**Dr. Ir. Veronica Sri Moertini, MT**
Ketua Tim Penguji	**Anggota Tim Penguji**
Husnul Hakim, S.Kom, M.T.	**Pascal A. Nugroho, S.Kom, M.Comp**

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

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

PENERAPAN ALGORITMA GREEDY K-MEMBER CLUSTERING UNTUK ANONIMISASI DATA PADA LINGKUNGAN BIG 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 29 Januari 2021

Stephen Jordan
NPM: 2016730018

ABSTRAK

Data mining umumnya digunakan untuk menganalisis pola-pola dari data yang dikumpulkan. Untuk mendapatkan hasil yang valid, data yang dianalisis harus sangat banyak. Oleh karena itu, teknologi *big data* muncul untuk menangani masalah tersebut. Sayangnya, proses *data mining* dapat menimbulkan masalah privasi. Privasi adalah hak seseorang untuk memiliki kendali atas bagaimana informasi pribadi dikumpulkan dan digunakan. *Privacy-Preserving Data Mining* (PPDM) digunakan untuk melindungi privasi individu sebelum dilakukan proses *data mining*. Contoh dari metode PPDM adalah *k-anonymity*. *K-anonymity* adalah metode anonimisasi data dari PPDM untuk menjaga agar sebuah data tidak dapat dibedakan dengan $k-1$ data lainnya. Karena metode anonimisasi menderita kehilangan informasi yang besar, maka data akan dikelompokan terlebih dahulu menggunakan algoritma *greedy k-member clustering*. Tujuan dari penelitian ini adalah melakukan implementasi algoritma *greedy k-member clustering* dan *k-anonymity* pada lingkungan *big data* dan menguji model *data mining* klasifikasi dan *clustering* sebelum dan setelah data dilakukan anonimisasi data.

Pada penelitian ini, telah dibangun tiga buah perangkat lunak dengan *framework* Spark. Perangkat lunak eksplorasi yang bertujuan mencari nilai unik sebuah kolom untuk dipakai dalam membuat pohon generalisasi. Perangkat lunak anonimisasi yang berisi implementasi algoritma *greedy k-member clustering* dan *k-anonymity*. Perangkat lunak pengujian untuk mengamati hasil pemodelan *data mining* sebelum dan setelah dilakukan anonimisasi data. Hasil perangkat lunak anonimisasi dipakai untuk tahap analisis. Analisis dilakukan dengan pengujian fungsional dan eksperimental. Pengujian fungsional bertujuan untuk memeriksa apakah perangkat lunak sudah berfungsi dengan seharusnya. Pengujian eksperimental bertujuan mendapatkan waktu komputasi algoritma *greedy k-member clustering* dan *k-anonymity*, waktu komputasi model *data mining* klasifikasi dan *clustering*, menghitung *total information loss*, melakukan evaluasi hasil *data mining*, dan mencari perbedaan hasil prediksi terbaik.

Hasil pengujian kualitas informasi menunjukan bahwa *total information loss* terendah dicapai menggunakan kolom campuran dengan bobot *total information loss* yang diperoleh yaitu 523541 untuk $k = 25$ dan 1000 data dengan 1 kolom kategorikal dan 1 kolom numerik, lalu memilih jumlah *quasi-identifier* yang tidak terlalu banyak (3-5 atribut) dengan bobot *total information loss* yang diperoleh yaitu 148091 untuk $k = 100$ dan 1000 data dengan 2 kolom kategorikal dan 1 kolom numerik, terakhir menggunakan ukuran data yang relatif kecil (kurang dari 10.000 data) dengan bobot *total information loss* yang diperoleh yaitu 3.10×10^7 untuk $k = 75$ dan 10.000 data dengan 2 kolom kategorikal dan 1 kolom numerik. Untuk waktu komputasinya, algoritma *greedy k-member clustering* membutuhan waktu sangat lama dalam melakukan pengelompokan data, yaitu lebih dari 3 jam untuk 10.000 data, sedangkan algoritma *k-anonymity* dapat dilakukan komputasi dengan cepat, yaitu kurang dari 15 menit untuk 10.000 data. Berdasarkan hasil pengujian, diketahui persentase perbedaan hasil *clustering* sebelum dan setelah anonimisasi data yang cukup jauh sekitar $0.70 - 0.85\%$, sedangkan persentase perbedaan hasil klasifikasi sebelum dan setelah anonimisasi data yang cukup dekat yaitu $0.30 - 0.55\%$. Sehingga, model *data mining* yang lebih tepat dipakai untuk anonimisasi data adalah klasifikasi.

Kata-kata kunci: *Data Mining*, *Big Data*, Privasi, *Privacy Preserving Data Mining (PPDM)*, K-Anonymity, Greedy K-Member Clustering

ABSTRACT

Data mining is used to analyze patterns of collected data. To get a valid result, the analyzed data must be very large. Hence, big data technology emerged to address this problem. Unfortunately, data mining process can create privacy concerns. Privacy is a person's right to have control over how personal information is collected and used. Privacy-Preserving Data Mining (PPDM) is used to protect the privacy of each individual before the data mining process is carried out. An example of the PPDM method is the k-anonymity. K-anonymity is data anonymization method from PPDM to keep data indistinguishable from other $k-1$ data. Since the anonymization method suffers from a large loss of information, the data will be grouped first musing the greedy k-member clustering algorithm. The purpose of this research is to implement the greedy k-member clustering and k-anonymity algorithms in the big data and test the data mining classification and clustering models before and after the data is anonymized.

In this research, there are three pieces of software created with framework Spark. Exploration software aims to find the unique values of a column to use in creating a generalization tree. Anonymization software contains the implementation greedy k-member clustering and k-anonymity algorithms. Testing software for observing the results of data mining modeling before and after data anonymization. The results of the anonymization software are used for the analysis stage. The analysis was carried out by functional and experimental testing. Functional testing aims to check whether the software is functioning properly. Experimental testing aims to obtain the computation time of the greedy k-member clustering and k-anonymity algorithms, the classification time data mining and clustering, calculating total information loss, evaluate data mining results, and look for differences in the best prediction results.

The results of the information quality test show that the lowest total information loss is achieved by using a mixed column with the weight total information loss obtained is 523541 for $k = 25$ and 1000 data with 1 categorical column and 1 numeric column, then choose the number of quasi-identifier that is not too many (3-5 attributes) with the weight total information loss obtained is 148091 for $k = 100$ and 1000 data with 2 categorical columns and 1 numeric column, Finally, using a relatively small data size (less than 10,000 data) with a weight of total information loss that is 3.10×10^7 for $k = 75$ and 10,000 data with 2 categorical columns and 1 numeric column. For computation time, the greedy k-member clustering algorithm takes a very long time to group data, which is more than 3 hours for 10,000 data, while the k-anonymity algorithm can be computed quickly, which is less than 15 minutes for 10,000 data. Based on the test results, the percentage difference in the results of clustering before and after data anonymization is quite far around $0.70 - 0.85\%$, while the percentage difference in classification results before and after data anonymization is quite close, namely $0.30 - 0.55\%$. Thus, the data mining model that is more appropriate for data anonymization is classification.

Keywords: Data Mining, Big Data, Privacy, Privacy Preserving Data Mining, K-Anonymity, Greedy K-Member Clustering

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 karuniaNya, penulis dapat menyelesaikan penyusunan buku skripsi. Penulisan buku skripsi ini bertujuan untuk memberikan pemahaman kepada pembaca mengenai penerapan konsep *Privacy-Preserving Data Mining* di lingkungan *big data* menggunakan metode *k-anonymity* dan algoritma *greedy k-member clustering* untuk menghasilkan kualitas *data mining* yang lebih baik. 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 utama dan Dr. Ir. Veronica Sri Moertini, MT selaku dosen pembimbing pendamping yang telah membimbing penulis selama proses penyusunan buku skripsi ini.

2. Bapak Husnul Hakim, S.Kom, M.T., dan Pascal A. Nugroho, S.Kom, M.Comp selaku dosen penguji yang telah memberikan kritik dan saran yang membangun.

3. Paman penulis yaitu Djunaedi Tatang dan kakak kandung penulis yaitu Erline yang telah bekerja keras untuk membiayai kuliah penulis sampai akhir.

4. Ibu kandung penulis yaitu Mariana Tatang dan seluruh kerabat penulis yang selalu memberikan doa dan dukungan yang terbaik bagi penulis.

5. Sahabat baik penulis yang sudah dianggap seperti keluarga sendiri, yaitu Regen Renaldo, Aldo Verrell mengizinkan penulis untuk tinggal sementara di kost seperti di rumah sendiri.

6. Sahabat baik penulis untuk berkeluh kesah dan berbagi sudut pandang penulis terhadap kejadian yang menggemparkan penulis, yaitu Chris Eldon dan Amabel Levint.

7. Sahabat baik penulis dari SMA yaitu Raynaldo William, William Reynaldo, Joshua Johanes, dan Michael Christian yang selalu mendukung dan menyemangati penulis.

8. Sahabat baik penulis selama di UNPAR, yaitu Louis Genio, David Widjaya, Antonius Susanto, Frengki Ang sebagai teman berbagi cerita, teman bermain, menyemangati satu sama lain.

Penulis menyadari bahwa penelitian dan penulisan buku skripsi 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 berguna sebagai bahan studi untuk penelitian selanjutnya.

Bandung, Januari 2021

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

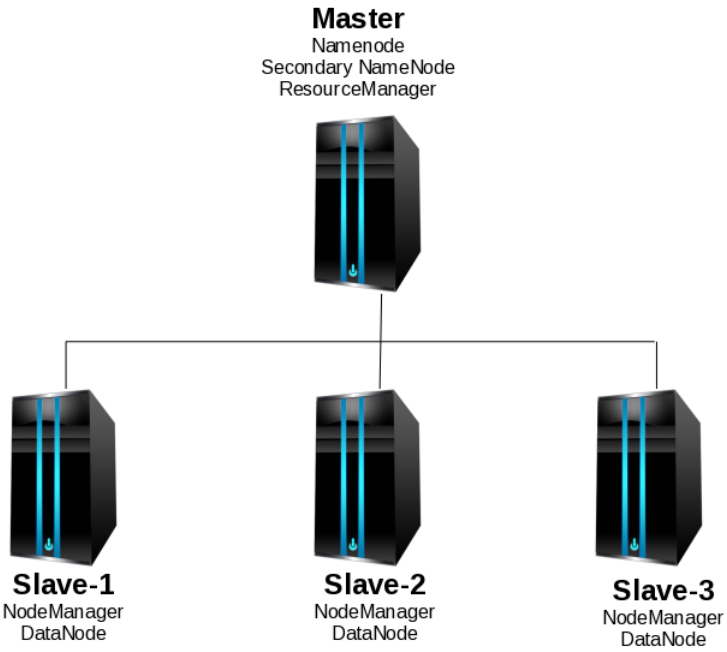
DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Data mining umumnya digunakan untuk menganalisis pola-pola dari data yang dikumpulkan. Untuk mendapatkan hasil yang valid, data yang dianalisis harus sangat banyak. Oleh karena itu, teknologi *big data* muncul untuk menangani masalah tersebut. Menurut Gartner, salah satu perusahaan riset teknologi informasi di Amerika Serikat, *big data* adalah adalah aset informasi bervolume tinggi, cepat, dan beragam yang menuntut bentuk pemrosesan informasi yang hemat biaya dan inovatif untuk meningkatkan wawasan dan pengambilan keputusan[1]. Teknologi *big data* dapat membagi pekerjaan pengolahan data ke beberapa komputer menggunakan konsep sistem terdistribusi. Sistem terdistribusi adalah solusi pengolahan *big data* karena terbukti dapat mengurangi biaya penyimpanan dan komputasi data dari pemrosesan data secara paralel. Contoh dari sistem terdistribusi *big data* adalah Hadoop. Gambar 1.1 menunjukan contoh dari sistem terdistribusi Hadoop, dimana pengguna hanya perlu menjalankan perangkat lunak dan memberikan input pada komputer *master* saja. Nantinya, komputer *master* akan memecah pekerjaan perangkat lunak terhadap masing-masing komputer *slave* sehingga pekerjaan dapat dilakukan secara paralel.



Gambar 1.1: Sistem Terdistribusi[2]

[1]https://www.gartner.com/en/information-technology/glossary/big-data
[2]http://kippel.net/blog/hadoop-cluster

Privasi adalah hak seseorang untuk memiliki kendali atas bagaimana informasi pribadi dikumpulkan dan digunakan. Privasi dapat terlanggar saat dilakukan proses *data mining*. *Privacy Preserving Data Mining* (PPDM) berperan penting untuk memberi perlindungan privasi dalam proses *data mining*. Konsep PPDM dapat dicapai dengan metode enkripsi dan anonimisasi. Menurut Gartner, enkripsi adalah proses pengkodean aliran bit secara sistematis sebelum dilakukan transmisi sehingga pihak yang tidak berwenang tidak dapat mengetahui arti pesan sebenarnya[3]. Anonimisasi adalah metode yang menyamarkan satu atau lebih nilai kolom data agar sebuah data tidak dapat saling dibedakan dengan data lainnya. Metode anonimisasi lebih unggul karena tidak perlu membuat kunci untuk menjaga privasi data. Metode anonimisasi dapat diterapkan pada data *credit score* yang menyimpan informasi pribadi dari pendaftar kartu kredit.

Umumnya data yang disamarkan dengan metode anonimisasi mengalami kehilangan informasi yang cukup besar. Istilah ini lebih sering dikenal dengan nama *total information loss*. Hal ini mengakibatkan data yang telah dianonimisasi memiliki hasil prediksi atau pengelompokan data yang buruk, jika mengguna model *data mining* klasifikasi/*clustering*. Salah satu cara untuk mencegah hal tersebut adalah dengan melakukan pengelompokan data terlebih dahulu sebelum dilakukan anonimisasi. Contoh algoritma pengelompokan data yang ingin diuji adalah *greedy k-member clustering*. Algoritma ini dipilih karena dapat mencari solusi paling optimal untuk mendapatkan kelompok data dengan nilai *information loss* yang rendah. Metode anonimisasi PPDM yang digunakan adalah *k-anonymity*. Metode ini menjaga sebuah data tidak dapat dibedakan dengan $k-1$ data lainnya. Karena penelitian berkaitan pada lingkungan *big data* dan membutuhkan implementasi algoritma secara iteratif, maka diperlukan teknologi untuk menangani komputasi secara paralel untuk kinerja yang lebih efisien. Spark merupakan pilihan yang tepat untuk pemrosesan *big data*, karena dapat memanfaatkan komputasi memori untuk komputasi yang lebih cepat dalam implementasi algoritma iteratif seperti *greedy k-member clustering*.

Pada skripsi ini, dibuat tiga jenis perangkat lunak yaitu perangkat lunak eksplorasi, perangkat lunak anonimisasi, dan perangkat lunak pengujian. Perangkat lunak eksplorasi bertujuan mencari nilai unik untuk pembuatan pohon generalisasi yang digunakan ketika menghitung jarak terdekat antar data kategorikal. Perangkat lunak anonimisasi bertujuan mengimplementasikan proses pengelompokan data dengan algoritma *greedy k-member clustering* dan proses anonimisasi data dengan metode *k-anonimity*. Perangkat lunak pengujian bertujuan membandingkan kualitas hasil data mining sebelum dan setelah data dianonimisasi. Contohnya *silhouette score*untuk menguji seberapa baik kualitas pengelompokan dari model *data mining*. Ketiga perangkat lunak ini dibuat dengan bahasa Scala, berjalan di atas Spark, dan menggunakan Hadoop untuk menyimpan hasil komputasi sementara dari algoritma *greedy k-member clustering*. Algoritma *greedy k-member clustering* menghasilkan anonimisasi terbaik karena memiliki *total information loss* paling rendah, jika memilih nilai k terkecil. Hal ini telah dijelaskan pada buku skripsi Edrick, mengenai analisis algoritma datafly dan greedy k-member clustering dalam menjaga privasi data. Penelitian ini memiliki tujuan utama yaitu membandingkan *silhouette score* untuk model *clustering k-means*, tingkat akurasi untuk model klasifikasi *naive bayes*, mencari parameter terbaik pada model *clustering k-means* dan klasifikasi *naive bayes*, dan terakhir mencari perbedaan hasil *clustering k-means* dan klasifikasi *naive bayes* sebelum dan setelah dilakukan anonimisasi.

[3]https://www.gartner.com/en/information-technology/glossary/encryption

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah pada skripsi ini adalah sebagai berikut:

1. Bagaimana cara kerja algoritma *greedy k-member clustering* untuk pengelompokan data?
2. Bagaimana cara kerja algoritma *k-anonymity* untuk anonimisasi data?
3. Bagaimana implementasi algoritma *greedy k-member clustering* pada Spark?
4. Bagaimana performa algoritma *greedy k-member clustering* dan *k-anonymity* untuk lingkungan *big data* pada *hadoop cluster*?
5. Bagaimana performa pemodelan *data mining clustering* (k-means) dan klasifikasi (naive bayes) untuk lingkungan *big data* pada *hadoop cluster*?
6. Bagaimana kualitas informasi pada data yang telah dikelompokan dengan algoritma *greedy k-member clustering* berdasarkan total information loss?
7. Bagaimana analisis kualitas hasil *data mining* terhadap pemodelan *clustering* sebelum dan setelah dilakukan anonimisasi?
8. Bagaimana analisis kualitas hasil *data mining* terhadap pemodelan klasifikasi sebelum dan setelah dilakukan anonimisasi?

1.3 Tujuan

Berdasarkan rumusan masalah di atas, tujuan dari skripsi ini adalah sebagai berikut:

1. Mempelajari cara kerja algoritma *greedy k-member clustering* untuk pengelompokan data.
2. Mempelajari cara kerja algoritma *k-anonymity* untuk anonimisasi data.
3. Mengimplementasi algoritma *greedy k-member clustering* dan *k-anonymity* pada Spark.
4. Menganalisis performa dari algoritma *greedy k-member clustering* dan *k-anonymity* untuk lingkungan *big data* pada *hadoop cluster*.
5. Menganalisis performa dari pemodelan data mining *clustering* (k-means) dan klasifikasi (naive bayes) untuk lingkungan *big data* pada *hadoop cluster*.
6. Menganalisis kualitas informasi pada data yang telah dikelompokan dengan algoritma *greedy k-member clustering* berdasarkan *total information loss*.
7. Menganalisis kualitas hasil metode *data mining clustering* berdasarkan *silhouette score*, mencari parameter *clustering* terbaik, dan mencari persentase perbedaan hasil *clustering* terbaik sebelum dan setelah dilakukan anonimisasi.
8. Menganalisis kualitas hasil metode *data mining* klasifikasi berdasarkan tingkat akurasi, mencari parameter klasifikasi terbaik, dan mencari persentase perbedaan hasil klasifikasi terbaik sebelum dan setelah dilakukan anonimisasi.

1.4 Batasan Masalah

Batasan masalah pada pengerjaan skripsi ini adalah sebagai berikut:

1. Masing-masing perangkat lunak menerima parameter masukan dalam format JSON dan data input dalam format CSV.

2. Perangkat lunak hanya dapat mengeluarkan output dalam format CSV. Folder penyimpanan output perangkat lunak dapat diubah pada data JSON.

3. Perangkat lunak anonimisasi dapat melakukan pengelompokan dan anonimisasi data hingga 100.000 data karena masalah waktu pengujian.

4. Perangkat lunak pengujian diuji dengan komputer lokal, karena listrik di UNPAR dipadamkan menjelang hari libur, sehingga *hadoop cluster* tidak dapat diakses secara *remote*.

1.5 Metodologi

Bagian-bagian pengerjaan skripsi ini adalah sebagai berikut:

1. Mempelajari dasar-dasar privasi data

2. Mempelajari konsep *k-anonimity* pada algoritma *greedy k-member clustering*.

3. Mempelajari teknik-teknik dasar *data mining*.

4. Mempelajari konsep Hadoop, Spark, dan Spark MLlib.

5. Mempelajari bahasa pemrograman Scala pada Spark.

6. Melakukan analisis masalah dan mengumpulkan data studi kasus.

7. Mengimplementasikan algoritma *greedy k-member clustering* pada Spark.

8. Mengimplementasikan teknik *data mining* menggunakan *library* Spark MLlib.

9. Melakukan pengujian fungsional dan experimental.

10. Melakukan analisis hasil *data mining* sebelum dan setelah dilakukan anonimisasi.

11. Menarik kesimpulan berdasarkan hasil eksperimen yang telah dilakukan.

1.6 Sistematika Pembahasan

Pengerjaan skripsi ini tersusun atas enam bab 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 konsep privasi, teknik *data mining*, *privacy-preserving data mining*, *k-anonymity*, algoritma *greedy k-member clustering*, metrik *distance* dan *information loss*, teknologi *big data*, pemrograman scala, dan format penyimpanan data.

- Bab 3 Analisis
  Berisi analisis penelitian mengenai analisis masalah (dataset eksperimen, *personally identifiable information*, perhitungan *distance* dan *information loss*, algoritma *greedy k-member clustering*, *k-anonymity*, *domain generalization hierarchy*), eksplorasi spark (instalasi spark, pembuatan *project* spark, menjalankan program spark), studi kasus (eksperimen scala, eksperimen spark), dan gambaran umum perangkat lunak (diagram kelas dan diagram aktivitas).

- Bab 4 Perancangan
  Berisi perancangan antarmuka perangkat lunak anonimisasi data dan analisis data, diagram kelas lengkap, masukan perangkat lunak anonimisasi data dan analisis data.

- Bab 5 Implementasi dan Pengujian
  Berisi implementasi perangkat lunak anonimisasi data dan analisis data, pengujian fungsional, pengujian eksperimental, dan melakukan analisis terhadap hasil pengujian.

- Bab 6 Kesimpulan dan Saran
  Berisi kesimpulan penelitian dan saran untuk penelitian selanjutnya.