

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 dikelompokkan 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 menunjukkan 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* membutuhkan 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 Erlina 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

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xxi
DAFTAR TABEL	xxv
DAFTAR KODE PROGRAM	xxix
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Batasan Masalah	4
1.5 Metodologi	4
1.6 Sistematika Pembahasan	5
DAFTAR NOTASI	1
2 LANDASAN TEORI	7
2.1 Privasi	7
2.2 <i>Data Mining</i>	8
2.2.1 Klasifikasi	9
2.2.2 <i>Clustering</i>	14
2.3 Tahapan Evaluasi <i>Data Mining</i>	19
2.3.1 Menghitung Tingkat Akurasi untuk Model Klasifikasi	19
2.3.2 Menghitung <i>Silhouette Score</i> untuk Model <i>Clustering</i>	19
2.4 <i>Privacy-Preserving Data Mining</i> (PPDM)	20
2.5 Metode Anonimisasi	20
2.6 <i>K-Anonymity</i>	21
2.7 <i>Domain Generalization Hierarchy</i> (DGH)	23
2.8 Metrik <i>Distance</i> dan <i>Information Loss</i>	24
2.8.1 <i>Distance</i>	24
2.8.2 <i>Information Loss</i>	25
2.9 <i>Greedy K-Member Clustering</i>	25
2.10 Sistem Terdistribusi	28
2.10.1 Penggunaan Sistem Terdistribusi	29
2.10.2 Komputasi Big Data Dengan Sistem Terdistribusi	29
2.11 Big Data	30
2.11.1 Karakteristik Big Data	30
2.11.2 Jenis Data Pada Big Data	30
2.12 Hadoop	31

2.12.1	HDFS	31
2.12.2	MapReduce	31
2.13	Spark	33
2.13.1	Arsitektur Spark	33
2.13.2	Jenis Instalasi pada Spark	34
2.13.3	Resilient Distibuted Datasets (RDD)	35
2.13.4	<i>DataFrame</i>	36
2.13.5	Komponen Spark	37
2.14	Spark MLlib	38
2.14.1	Tipe Data pada Spark MLlib	39
2.14.2	<i>Data Mining</i> pada Spark MLlib	39
2.15	Scala	42
2.16	Format Penyimpanan Data	43
2.16.1	CSV	43
2.16.2	JSON	43
3	ANALISIS	45
3.1	Analisis Masalah	45
3.2	Gambaran Umum Perangkat Lunak	46
3.2.1	Diagram	47
3.2.2	Diagram Kelas	48
3.2.3	Pengenalan Karakteristik Data	56
3.2.4	<i>Personally Identifiable Information</i>	56
3.2.5	Perhitungan <i>Distance</i> dan <i>Total Information Loss</i>	57
3.2.6	<i>Greedy K-Member Clustering</i>	58
3.2.7	<i>Domain Generalization Hierarchy</i>	60
3.2.8	<i>K-Anonymity</i>	61
4	EKSPLORASI SPARK	63
4.0.1	Instalasi Spark	63
4.0.2	Membuat <i>Project</i> Spark pada IntelliJ	67
4.1	Studi Kasus	69
4.1.1	Eksperimen Scala	69
4.1.2	Eksperimen Spark	72
4.1.3	Eksperimen Komponen Spark	75
4.1.4	Eksperimen Spark MLIB	79
5	PERANCANGAN	83
5.1	Spark Web UI	84
5.1.1	Menu Jobs	85
5.1.2	Menu Stages	86
5.1.3	Menu Storages	87
5.1.4	Menu Environment	88
5.1.5	Menu Executors	89
5.1.6	Menu SQL	90
5.2	Diagram Kelas Lengkap	91
5.2.1	Diagram Package	91
5.2.2	Diagram Kelas pada Package <i>ExploratoryModel</i>	92
5.2.3	Diagram Kelas pada Package <i>AnonymizationModel</i>	92
5.2.4	Diagram Kelas pada Package <i>ExaminationModel</i>	104
5.3	Masukan Perangkat Lunak	107
5.3.1	Masukan Perangkat Lunak Eksplorasi	108

5.3.2	Masukan Perangkat Lunak Anonimisasi	108
5.3.3	Masukan Perangkat Lunak Pengujian	110
6	IMPLEMENTASI DAN PENGUJIAN	113
6.1	Implementasi Antarmuka	113
6.1.1	Komputer Lokal dengan IntelliJ	113
6.1.2	Hadoop Cluster dengan Terminal Ubuntu	121
6.2	Pengujian	129
6.2.1	Pengujian Fungsional	129
6.2.2	Pengujian Eksperimental	135
7	KESIMPULAN DAN SARAN	149
7.1	Kesimpulan	149
7.2	Saran	150
	DAFTAR REFERENSI	151
A	KONFIGURASI LIBRARY SPARK	153
A.1	Perangkat Lunak Eksplorasi	153
A.2	Perangkat Lunak Anonimisasi	153
A.3	Perangkat Lunak Pengujian	153
B	KODE PROGRAM PERANGKAT LUNAK EKSPLORASI	155
B.1	Kelas MainExploratory	155
C	KODE PROGRAM PERANGKAT LUNAK ANONIMISASI	157
C.1	Kelas Node	157
C.2	Kelas BinaryTree	157
C.3	Kelas GreedyKMemberClustering	158
C.4	Kelas KAnonymity	162
C.5	Kelas LowestCommonAncestor	165
C.6	Kelas MainAnonymization	165
C.7	Kelas MainClusterization	167
C.8	Kelas MainLCATesting	169
D	KODE PROGRAM PERANGKAT LUNAK PENGUJIAN	171
D.1	Kelas KMeansModel	171
D.2	Kelas MainExamination	172
D.3	Kelas NaiveBayesModel	179
E	MASUKAN FILE JSON SELURUH PERANGKAT LUNAK	181
E.1	Masukan JSON Perangkat Lunak Eksplorasi	181
E.2	Masukan JSON Perangkat Lunak Anonimisasi	181
E.3	Masukan JSON Perangkat Lunak Pengujian	188
F	HASIL PENGUJIAN FUNGSIONAL	191
F.1	Hasil Pengelompokan Greedy K-Member Clustering	191
F.2	Hasil Anonimisasi K-Anonymity	192
F.3	Hasil Clustering K-Means (Sebelum Anonimisasi)	193
F.4	Hasil Clustering K-Means (Setelah Anonimisasi)	194
F.5	Hasil Klasifikasi Naive Bayes (Sebelum Anonimisasi)	195
F.6	Hasil Klasifikasi Naive Bayes (Setelah Anonimisasi)	196

DAFTAR GAMBAR

2.2	Ilustrasi Supervised Learning [1]	9
2.3	Tahapan Pelatihan Model Klasifikasi [2]	9
2.4	Tahapan Prediksi Model Klasifikasi [2]	10
2.6	Contoh <i>Hierarchical Clustering</i> [1]	14
2.7	Contoh <i>Hierarchical Clustering</i> [1]	15
2.8	Contoh <i>Partitional Clustering</i> [1]	15
2.9	Contoh <i>Partitional Clustering</i>	16
2.10	<i>Privacy Preserving Data Mining</i> (PPDM) [3]	20
2.11	Pohon Generalisasi Data Numerik	21
2.12	Pohon Generalisasi Data Kategorikal	22
2.13	Contoh Implementas DGH,VGH (ZIP) [4]	23
2.15	Sistem Terdistribusi	28
2.16	Pemanfaatan Sistem Terdistribusi	29
2.17	Hadoop	31
2.18	Arsitektur HDFS [5]	31
2.20	Spark	33
2.21	Arsitektur Spark [6]	33
2.22	Jenis Instalasi Spark [6]	34
2.23	Beberapa Jenis Komponen Spark	37
2.24	Jenis Pemodelan Data Mining Spark MLib	38
2.25	Contoh Vektor Dense dan Sparse	39
2.26	Scala dan Java JVM	42
3.1	Flow Chart Penggunaan Perangkat Lunak	47
3.2	Diagram Kelas Perangkat Lunak Anonimisasi	49
3.3	Diagram Kelas Perangkat Lunak Pengujian	52
3.4	Diagram Aktifitas Perangkat Lunak Ekplorasi	54
3.5	Diagram Aktifitas Perangkat Lunak Anonimisasi	55
3.6	Diagram Aktifitas Perangkat Lunak Pengujian	55
3.7	Pohon DGH (NAME_INCOME_TYPE)	57
3.8	DGH dan VGH pada atribut ZIP	60
4.1	Environment Variables	64
4.2	Penambahan Variable Value	64
4.3	Perintah java -version	64
4.4	Environment Variable	65
4.5	Penambahan Variable Value	65
4.6	Spark 2.4.5	65
4.7	Instalasi IntelliJ	66
4.8	Plugins Scala	66
4.9	Memilih Bahasa Scala Berbasis sbt	67
4.10	Melakukan Konfigurasi Project Spark	67
4.11	Menambahkan Scala Class pada Project Spark	68

4.12	Memilih Tipe Object pada Scala Class	68
4.13	Hasil Naive Bayes Spark MLlib	80
4.14	Hasil K-Means Spark MLlib	81
5.1	Spark Web UI	83
5.2	Tampilan Utama Cluster Web UI	84
5.3	Tampilan Utama Spark Web UI	84
5.4	Menu Jobs	85
5.5	Number of jobs per status	85
5.6	Event timeline	86
5.7	Details of Jobs Grouped By Status	86
5.8	Menu Stages	86
5.9	Number of stages per status	87
5.10	Details of stages grouped by status	87
5.11	Summary	87
5.12	RDD Detail	88
5.13	Menu Environment	88
5.14	Hadoop properties	89
5.15	System properties	89
5.16	Classpath Entries	89
5.17	Menu Executors	90
5.18	Menu SQL	90
5.19	Diagram Kelas pada Package	91
5.20	Diagram Kelas pada Package ExploratoryModel	92
5.21	Diagram Kelas pada Package AnonymizationModel	93
5.22	Diagram Kelas pada ExaminationModel	104
6.1	Tombol Selector MainExploratory	113
6.2	Konfigurasi Parameter Perangkat Lunak Ekplorasi	114
6.3	Cara Menjalankan Perangkat Lunak Ekplorasi	114
6.4	Log Perangkat Lunak Ekplorasi	115
6.5	Folder Output Perangkat Lunak Ekplorasi	115
6.6	Flowchart Penggunaan Perangkat Lunak	116
6.7	Tombol Selector MainAnonymization	116
6.8	Konfigurasi Parameter Perangkat Lunak Anonimisasi	116
6.9	Menjalankan Perangkat Lunak Anonimisasi	117
6.10	Log Perangkat Lunak Anonimisasi	117
6.11	Folder Output Greedy K-Member Clustering	118
6.12	Folder Output K-Anonymity	118
6.13	Flowchart Penggunaan Perangkat Lunak (Lanjutan)	118
6.14	Tombol Selector MainExamination	119
6.15	Konfigurasi Parameter Perangkat Lunak Ekplorasi	119
6.16	Menjalankan Perangkat Lunak Ekplorasi	120
6.17	Log Perangkat Lunak Ekplorasi	120
6.18	Folder Output Perangkat Lunak Pengujian	121
6.19	Folder Output Perangkat Lunak Pengujian	121
6.20	Spesifikasi Slaves Node pada Hadoop Cluster	122
6.21	File Input Eksplorasi HDFS	123
6.22	Log Perangkat Lunak Ekplorasi	123
6.23	Folder HDFS Hasil Ekplorasi	124
6.24	Folder HDFS Hasil Ekplorasi Atribut Race	124
6.25	File Input Anonimisasi HDFS	125

6.26	Log Perangkat Lunak Anonimisasi	126
6.27	Folder HDFS Hasil Pengelompokan Data	126
6.28	Folder HDFS Hasil Anonimisasi Data	126
6.29	File Input Pengujian HDFS	127
6.30	Log Perangkat Lunak Pengujian	128
6.31	Folder HDFS Hasil Pengelompokan K-Means	129
6.32	Folder HDFS Hasil Klasifikasi Naive Bayes	129
6.33	Waktu Pengelompokan Data (Jenis Kolom Bervariasi)	139
6.34	Waktu Anonimisasi Data (Jenis Kolom Bervariasi)	139
6.35	Total Information Loss (Jenis Kolom Bervariasi)	139
6.36	Waktu Pengelompokan Data (Jumlah QID bervariasi)	141
6.37	Waktu Anonimisasi Data (Jumlah QID bervariasi)	141
6.38	Total Information Loss (Jumlah QID bervariasi)	141
6.39	Waktu Pengelompokan Data (Ukuran data bervariasi)	143
6.40	Waktu Anonimisasi Data (Ukuran data bervariasi)	143
6.41	Total Information Loss (Ukuran data bervariasi)	143
6.42	Silhouette Score (K-means)	144
6.43	Waktu Komputasi (K-means)	144
6.44	Perbandingan Hasil Clustering Terhadap K-Value (K-means)	145
6.45	Perbandingan Hasil Clustering Terhadap Jumlah Data (K-means)	145
6.46	Perbandingan Tingkat Akurasi (Naive bayes)	146
6.47	Perbandingan Waktu Komputasi (Naive bayes)	146
6.48	Perbandingan Hasil Klasifikasi Terhadap Jumlah Data (Naive bayes)	146

DAFTAR TABEL

2.1	Contoh Kasus <i>PlayGolf</i>	11
2.2	Tabel Probabilitas pada Atribut <i>Outlook</i>	12
2.3	Tabel Probabilitas dari Atribut <i>Temperature</i>	12
2.4	Tabel Probabilitas dari Atribut <i>Humidity</i>	13
2.5	Tabel Probabilitas dari Atribut <i>Wind</i>	13
2.6	Tabel Dataset Mata Pelajaran	16
2.7	Hasil Pengelompokan Awal	17
2.8	Mencari Centroid Kelompok	17
2.9	Hasil Cluster Baru	18
2.10	Euclidean Distance Cluster 1, Cluster 2	18
2.11	Hasil Pengelompokan Akhir	18
2.12	Perbandingan Konsep Supresi, Generalisasi	23
2.13	Penjelasan Fungsi <i>Transformation</i> pada Spark	35
2.14	Penjelasan Fungsi <i>Action</i> pada Spark	35
2.15	Penjelasan Operasi Select, Filter pada Spark	36
2.16	Penjelasan Operasi Inner-Join, Cross-Join pada Spark	36
2.17	Penjelasan Operasi Agregat pada Spark	36
3.1	Tabel Hasil Clustering Data pada Cluster 1	58
3.2	Data Credit Score	58
3.3	Hasil Pengelompokan Akhir (Greedy k-member clustering)	59
3.4	Hasil Pengelompokan Data (Greedy k-member clustering)	61
3.5	Hasil Anonimisasi Data (K-Anonymity)	61
6.1	Pengujian Kualitas Informasi	135
6.2	Pengujian Hasil Data Mining	136
6.3	Konfigurasi Pengujian	136
6.4	Konfigurasi Pengujian	137
6.5	Sampel Data Credit Score(Numerik)	137
6.6	Sampel Data Credit Score(Kategorikal)	137
6.7	Sampel Data Credit Score(Campuran)	138
6.8	Sampel Data Credit score (QID =2)	140
6.9	Sampel Data Credit score (QID =3)	140
6.10	Sampel Data Credit score (QID =4)	140
6.11	Sampel Data Credit score (10k dan 30k)	142
6.12	Kesimpulan Pengujian Eksperimental	147

DAFTAR KODE PROGRAM

2.1	Import Library Spark Core	37
2.2	Import Library Spark SQL	37
2.3	Import Library Spark MLlib	37
2.4	Memberi Nama Perangkat Lunak Spark (Naive Bayes)	40
2.5	Memisahkan Data Training	40
2.6	Melatih Model Naive Bayes	40
2.7	Prediksi Model Naive Bayes	40
2.8	Menghitung Tingkat Akurasi Naive Bayes	40
2.9	Menyimpan Model Naive Bayes	40
2.10	Memberi Nama Perangkat Lunak Spark (K-Means)	41
2.11	Melatih Model K-Means	41
2.12	Prediksi Model K-Means	41
2.13	Menghitung Silhouette Score K-Means	41
2.14	Menyimpan Model K-Means	41
2.15	Baris Kode Java	42
2.16	Baris Kode Scala	42
2.17	Format Penyimpanan CSV	43
2.18	Format Penyimpanan JSON	43
3.1	Dataset Credit Score	56
4.1	Melakukan Import Library Spark	68
4.2	Menambahkan Main method pada Scala Class	68
4.3	Menentukan Jenis Variabel pada Scala	69
4.4	Menentukan Jenis Tipe Data pada Scala	69
4.5	Membuat immutable collection pada Scala	70
4.6	Membuat mutable collection pada Scala	70
4.7	Membuat Kelas Object pada Scala	70
4.8	Membuat Kelas Object pada Scala	71
4.9	Membuat Fungsi Sedehana pada Scala	71
4.10	Membuat Fungsi Percabangan pada Scala	71
4.11	Membuat Fungsi Perulangan pada Scala	72
4.12	Konfigurasi Spark	72
4.13	Cara Pembuatan RDD	73
4.14	Cara Pembuatan Dataframe	73
4.15	Contoh Fungsi Transformation	73
4.16	Contoh Fungsi Action	74
4.17	Contoh Fungsi RDD	75
4.18	Membuat Variabel Global	75
4.19	Membuat SparkSession	75
4.20	Melihat dan Mengatur Partisi RDD	75
4.21	Membuat Fungsi Transformation	75

4.22	Membuat Fungsi Action	76
4.23	Membuat Perintah SparkSession	76
4.24	Membuat Dataframe	77
4.25	Membuat Tabel Sementara	77
4.26	Mencari Nilai Statistik	77
4.27	Mencari Nilai Median	77
4.28	Mencari Nilai Modus	77
4.29	Membuat Perintah SparkSession	78
4.30	Membuat Skema Dataframe	78
4.31	Mengubah CSV Menjadi Dataframe	78
4.32	Membuat Kolom Index	78
4.33	Membuat Kolom Vektor	79
4.34	Memilih Jenis Vektor	79
4.35	Membuat Vektor Fitur	79
4.36	Eksperimen Naive Bayes Spark MLlib	80
4.37	Eksperimen K-Means Spark MLlib	81
5.1	Data Credit Score	107
5.2	Input JSON untuk Eksplorasi Data	108
5.3	Input JSON untuk Anonimisasi Data	109
5.4	Input JSON untuk Pengujian Data	111
6.1	Perintah Spark untuk Perangkat Lunak Eksplorasi	122
6.2	Perintah Eksekusi Spark	123
6.3	Perintah Spark untuk Perangkat Lunak Anonimisasi	124
6.4	Perintah Eksekusi Spark	125
6.5	Perintah Spark untuk Perangkat Lunak Pengujian	127
6.6	Perintah Eksekusi Spark	128
6.7	Sampel Data Pengujian Fungsional	130
6.8	Sampel Data Credit Score	130
6.9	Hasil Pengelompokan Greedy K-Member Clustering	130
6.10	Sampel Pengelompokan Data	131
6.11	Hasil Anonimisasi K-Anonymity	132
6.12	Hasil Pengelompokan K-Means Sebelum Anonimisasi	132
6.13	Hasil Pengelompokan K-Means Setelah Anonimisasi	133
6.14	Perbedaan Silhouette Score	133
6.15	Perbedaan Persentase Hasil Pengelompokan (%)	133
6.16	Hasil Klasifikasi Naive Bayes Sebelum Anonimisasi	134
6.17	Hasil Klasifikasi Naive Bayes Setelah Anonimisasi	134
6.18	Perbedaan Tingkat Akurasi	134
6.19	Perbedaan Persentase Hasil Klasifikasi (%)	134
A.1	build.sbt (Eksplorasi)	153
A.2	build.sbt (Anonimisasi)	153
A.3	build.sbt (Pengujian)	153
B.1	MainExploratory.scala	155
C.1	Node.scala	157
C.2	BinaryTree.scala	157
C.3	GreedyKMemberClustering.scala	158
C.4	KAnonymity.scala	162
C.5	LowestCommonAncestor.scala	165

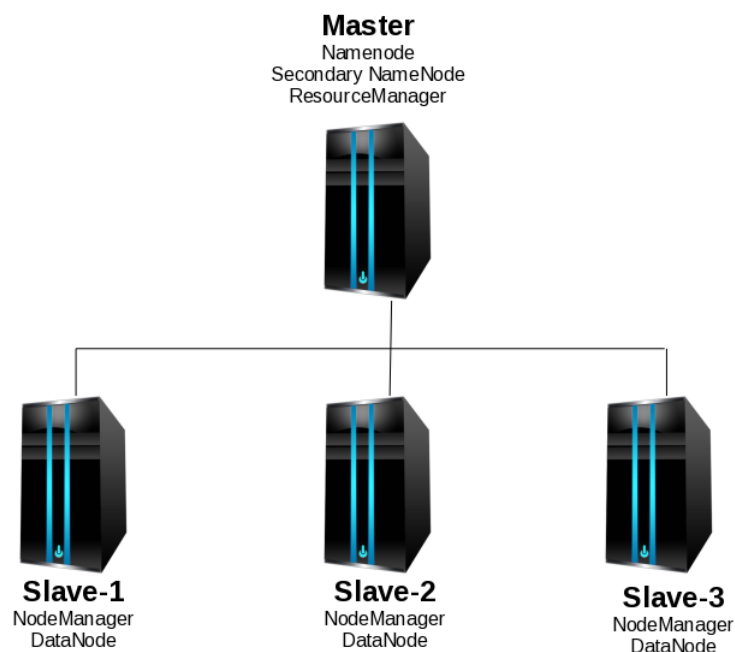
C.6	MainAnonymization.scala	166
C.7	MainClusterization.scala	167
C.8	MainLCATesting.scala	169
D.1	KMeansModel.scala	171
D.2	MainExamination.scala	172
D.3	NaiveBayesModel.scala	179
E.1	Data JSON untuk Perangkat Lunak Eksplorasi	181
E.2	Data JSON untuk Perangkat Lunak Anonimisasi	181
E.3	Data JSON untuk Perangkat Lunak Pengujian	188
F.1	Hasil Pengelompokan Greedy K-Member Clustering	191
F.2	Hasil Anonimisasi K-Anonymity	192
F.3	Hasil Clustering K-Means (Sebelum Anonimisasi)	193
F.4	Hasil Clustering K-Means (Setelah Anonimisasi)	194
F.5	Hasil Klasifikasi Naive Bayes (Sebelum Anonimisasi)	195
F.6	Hasil Klasifikasi Naive Bayes (Setelah Anonimisasi)	196

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 aset informasi bervolume tinggi, cepat, dan beragam yang menuntut bentuk pemrosesan informasi yang hemat biaya dan inovatif untuk meningkatkan wawasan dan pengambilan keputusan¹. 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 menunjukkan 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 Terdistribusi²

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 sebenarnya³. Anonimisasi adalah metode yang menyamakan 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-anonymity*. 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 dikelompokkan 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 dikelompokkan 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-anonymity* 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.

