BAB II
TINJAUAN PUSTAKA

2.1. Teori

2.1.1. Privasi

Privasi adalah hak untuk sendiri, hak seseorang untuk bebas dari keterbukaan publik. Aturan “hak privasi” merupakan aturan umum meliputi berbagai hak yang diakui dan melekat dalam konsep kebebasan untuk diperintah, dan beberapa hak yang mencegah pemerintah mencampuri urusan setiap individu untuk berhubungan dan berinteraksi, kebebasan individu, untuk membuat pilihan hidup yang diterapkan pada dirinya, keluarganya, dan hubungannya dengan individu lain (Garner, 1999 : 74-75). Definisi lain dari privasi adalah proses pengontrolan yang selektif terhadap akses kepada diri sendiri dan akses kepada orang lain (Altman, 1975:221). Sedangkan menurut Rapoport (1982), privasi merupakan suatu kemampuan untuk mengontrol interaksi, kemampuan untuk memperoleh pilihan-pilihan dan kemampuan untuk mencapai interaksi yang diinginkan. Privasi jangan dipandang hanya sebagai penarikan diri seseorang secara fisik terhadap pihak-pihak lain dalam rangka menyepi saja.

Berdasarkan pendapat para ahli tersebut maka dapat disimpulkan bahwa privasi merupakan hak yang dimiliki setiap individu untuk dapat mengontrol sesuatu yang bersifat rahasia dan pribadi yang terdapat pada dirinya yang tidak boleh diketahui oleh orang lain.

2.1.2. Data

Data adalah kumpulan dari fakta, konsep, atau instruksi pada penyimpanan yang digunakan untuk komunikasi, perbaikan dan diproses secara otomatis yang

mempresentasikan informasi yang dapat dimengerti oleh manusia (Inmon, 2005: 493). Definisi lain dari data adalah deskripsi dasar dari benda, peristiwa, aktivitas dan transaksi yang direkam, dikelompokkan, dan disimpan tetapi belum terorganisir untuk menyampaikan arti tertentu (Turban dkk, 2010:41). Sedangkan menurut Irmansyah (2003:14), secara umum, pengertian data dapat didefinisikan sebagai nilai (*value*) yang merepresentasikan deskripsi dari suatu objek atau peristiwa. Data dibentuk dari data mentah (*raw data*) yang berupa angka, karakter, gambar, atau bentuk lainnya. Data adalah bentuk jamak dari datum. Data merupakan keterangan-keterangan tentang suatu hal, dapat berupa sesuatu yang punya makna. Data dapat diartikan sebagai sesuatu yang diketahui atau yang dianggap atau anggapan.

Berdasarkan pendapat para ahli di atas, maka dapat disimpulkan bahwa data merupakan suatu informasi yang merepresentasikan deskripsi dari suatu objek atau peristiwa, baik berupa angka ataupun keterangan.

2.1.3. Model

Model adalah pola (contoh, acuan, ragam) dari sesuatu yang akan dibuat atau dihasilkan (Departemen P dan K, 1984:75). Definisi lain dari model adalah abstraksi dari sistem sebenarnya, dalam gambaran yang lebih sederhana serta mempunyai tingkat prosentase yang bersifat menyeluruh, atau model adalah abstraksi dari realitas dengan hanya memusatkan perhatian pada beberapa sifat dari kehidupan sebenarnya (Simamarta, 1983: ix – xii). Sedangkan menurut Mahmud Achmad (2008:1), model adalah representasi dari suatu objek, benda, atau ide-ide dalam bentuk yang disederhanakan dari kondisi atau fenomena alam. Model berisi informasi-informasi tentang suatu fenomena yang dibuat dengan

tujuan untuk mempelajari fenomena sistem yang sebenarnya. Model dapat merupakan tiruan dari suatu benda, sistem atau kejadian yang sesungguhnya yang hanya berisi informasi-informasi yang dianggap penting untuk ditelaah.

Berdasarkan tiga pendapat di atas maka dapat di ambil kesimpulan bahwa model adalah sebuah pola yang dapat merepresentasikan suatu benda, sistem, ataupun kejadian sesuai dengan tujuan yang ingin di capai. Dengan adanya model, maka orang-orang akan lebih paham dengan maksud yang akan disampaikan.

2.1.4. Anonim

Menurut *Oxford Advanced Learner's Dictionary* (2010) anonim merupakan (of a person) with a name that is not known or that is not made public, written, given, made, etc. by somebody who does not want their name to be known or made public, without any unusual or interesting features. Sedangkan menurut Kamus Besar Bahasa Indonesia (KBBI), anonim merupakan kata sifat dan memiliki makna sebagai berikut:

(1) tanpa nama; tidak beridentitas; awanama;

(2) [Sos]tidak ada penandatangannya

Sehinggan dapat disimpulkan bahwa anonim berarti tidak diketahui namanya. Biasanya pihak terkait sengaja merahasiakan namanya untuk menjaga privasi.

2.1.5. Algoritma

Menurut Rinaldi Munir (2005 : 176), algoritma adalah urutan logis langkah-langkah penyeleseian masalah yang disusun secara sistematis. Sedangkan menurut Thomas H. Cormen (2009:5), algoritma adalah prosedur komputasi yang mengambil beberapa nilai atau kumpulan nilai sebagai *input* kemudian di proses

sebagai *output* sehingga algoritma merupakan urutan langkah komputasi yang mengubah *input* menjadi *output*.

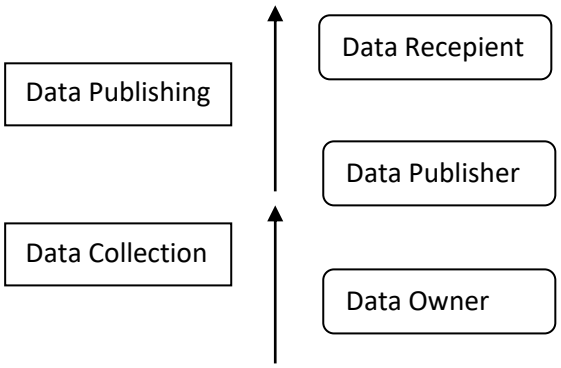
Jika dilihat dari asal katanya yaitu “algoritma”, kata ini tidak muncul dalam kamus Webster pada tahun 1957. Menurut Rinaldi Munir (2011:10), Para ahli bahasa menemukan kata algorism berasal dari nama cendikiawan muslim yang terkenal yaitu Abu Ja’far Muhammad Ibnu Musa Al-Khuwarijmi (Al-Khuwarijmi dibaca oleh orang Barat menjadi algorism) dalam bukunya yang berjudul Kitab Aljabar Wal-muqabala, yang artinya “Buku Pemugaran dan Pengurangan” (The book of restoration and reduction). Dari judul buku itu kita memperoleh kata “aljabar” (algebra). Perubahan dari kata algorism menjadi algorithm muncul karena kata algorism sering dikelirukan dengan arithmetic sehingga akhiran –sm berubah menjadi –thm.

Pada zaman di mana teknologi informasi sedang berkembang dengan pesatnya seperti sekarang, algoritma tidak dapat dipisah dengan program yang berjalan di berbagai *device*s, seperti komputer, *netbook*, dll. Hal ini dikarenakan algoritma itu sendiri merupakan sebuah pola kerja dari program yang ada. Dengan adanya algortima maka langkah-langkah yang ada pada suatu penyelesaian masalah dapat dilakukan seefisien mungkin karena langkah-langkahnya sudah tersusun secara rapi sehingga memungkinkan untuk mendapatkan tujuan yang maksimal.

Sehingga dapat disimpulkan bahwa algoritma merupakan langkah-langkah atau prosedur yang dibuat untuk menyelesaikan suatu masalah.

2.1.6. Privacy Preserving Data Publishing

Secara umum, *Privacy Preserving Data Publishing* memiliki 2 fase, yaitu *data collection* dan *data publishing.* Hal ini mengacu pada tiga jenis peranan yang terdapat selama prosesnya, yaitu *data owner, data publisher*, dan *data recepient.* Untuuk lebih jelas, dapat dilihat pada gambar dibawah:



Gambar 2.1 Hubungan Antara Fase dan Peranan di PPDP

Berdasarkan Gambar 2.1 dapat dilihat bahwa pada fase *data collection*, *data publisher* mengambil data dari *data owner*. Kemudian pada fase data *publishing*, *data publisher* mengirim data yang telah diproses kepada *data recepient*. Menurut Johannes Gehrke (2006:105), *data publisher* dapat dibagi menjadi dua kategori. Kategori pertama adalah *untrusted model* dimana *data publisher* yang lebih rumit memungkinkan untuk mendapatkan privasi dari dataset. Sedangkan untuk kategori kedua, yaitu *trusted model, data publisher* yang handal dengan data ditangan mereka akan menjadi lebih aman tanpa ada resiko.

Privacy Preserving Data Publishing berbeda dari *Privacy Preserving Data Collection,* karena diasumsikan bahwa semua catatan sudah tersedia kepada pihak yang terpercaya, yang mungkin memiliki data saat ini. Mereka kemudian ingin melepaskan (atau mempublikasikan) data ini untuk di analisis. Misalnya, rumah

sakit yang ingin melepaskan catatan anonim tentang pasien untuk mempelajari efektifitas dari berbagai alternatif pengobatan.

Beberapa teknik untuk melakukan privasi data adalah sebagai berikut:

2.1.4.1. *Pertubative*

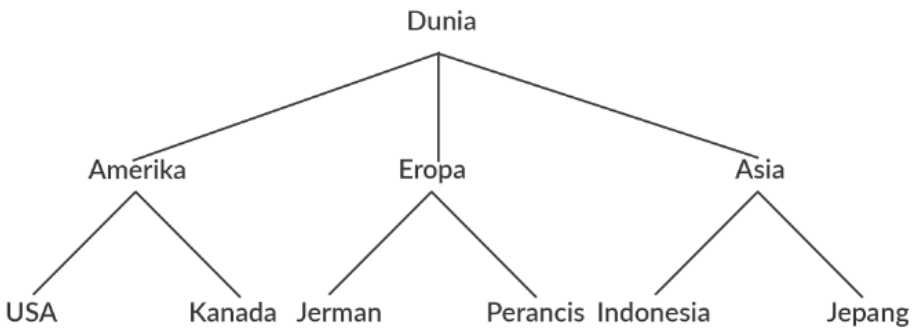
Pertubative merupakan penyembunyian data dengan cara mengubah data. Contohnya seperti Kriptografi dan memberikan *noise*. Tapi sayangnya cara ini tidak efektif karena dapat mengubah isi data sehingga data tidak dapat dikenali.

2.1.4.2. *Non-pertubative*

Non-Pertubative merupakan penyembunyian data tanpa mengubah datanya, pada cara ini yang dilakukan ialah penganoniman data. Contohnya adalah Generalisasi dan Suppression.

a. Generalisasi

Ide untuk menggeneralisasikan sebuah atribut merupakan konsep yang sederhana. Generalisai umumnya menggunakan *taxonomy tree*. Pada metode ini data akan diubah menjadi range yang lebih luas. Untuk lebih mudah memahami maksud dari generalisasi dapat dilihat pada Gambar 2.2:



Gambar 2.2 *Taxonomy Tree* Negara

Pada Gambar 2.2 dapat dilihat negara yang tadinya berisi USA, Kanada, Jerman, Perancis, Indonesia dan Jepang digeneralisasi menjadi Amerika {USA,

Kanada}, Eropa {Jerman, Perancis}, dan Asia {Indonesia, Jepang}. Selanjutnya Amerika, Eropa, dan Asia di perkecil lagi menjadi Dunia.

b. Suppression

Suppression berarti menutup data. Maksudnya adalah satu atau beberapa nilai yang ada pada atribut tertentu ditutup atau diganti dengan suatu lambang, misalnya * atau # sesuai dengan kondisinya. Jika satu huruf saja yang di tutup sudah bisa membuat suatu tabel tidak terhubung ke tabel lain, maka satu suppression saja sudah cukup. Akan tetapi, jika tabel-tabel yang ada masih bisa dihubungkan maka nilai sebelumnya harus ditutup sampai tabel-tabel yang ada tidak bisa terhubung.

Suppression sendiri terdiri dari dua jenis, yaitu *Global Recoding* dan *Local Recoding*.

1. *Global Recoding*

Global recoding merupakan teknik untuk memetakan seluruh nilai atribut kategori *quasi identifier* menjadi nilai yang lebih umum dalam hirarki domain generalisasi. Contohnya:

Tabel 2.1. Tabel Data Masyarakat yang Telah di Suppresi dengan *Global Recording*

Tuple	Age	Zipcode	Gender
r1	20-30	41***	Male
r2	20-30	41***	Male
r3	30-40	48***	Female
r4	30-40	48***	Female
r5	30-40	48***	Female

Sumber: K-anonymity: a model for protecting privacy hal 11 (Dengan Penyesuaian)

2. *Local Recoding*

Local Recording adalah teknik melakukan generalisasi untuk setiap grup *quasi identifier* yang berbeda. Contohnya:

Tabel 2.2. Tabel Data Masyarakat yang Telah di Suppresi dengan *Local Recording*

Tuple	Age	Zipcode	Gender
r1	20-30	410**	Male
r2	20-30	410**	Male
r3	30-40	*****	Female
r4	30-40	*****	Female
r5	30-40	*****	Female

Sumber: K-anonymity: a model for protecting privacy hal 11 (Dengan Penyesuaian)

2.1.7. Atribut

Atribut adalah ciri-ciri kualitatif yang dimiliki oleh suatu obyek, yang mencerminkan sifat-sifat dari obyek tersebut. Field menyatakan data terkecil yang memiliki makna. Istilah lain untuk field yaitu elemen data, kolom item. Jenis-jenis atribut tersebut adalah *Explicit Identifier, Quasi Identifier, Sensitive Attribute,* dan *Non-sensitive Attribute*.

2.1.7.1. *Explicit Identifier*

Explicit Identifier merupakan atribut yang secara eksplisit menyatakan identitas dari suatu individu, misalnya noreg, nama, NIP, NIK (Nomor Induk Kependudukan).

2.1.7.2. *Quasi Identifier*

Quasi Identifier merupakan gabungan beberapa atribut yang bisa berpotensi untuk menunjuk identitas tertentu jika dilakukan join dengan data lain.

2.1.7.3. *Sensitive Attribute*

Atribut Sensitif merupakan atribut yang dianggap pribadi dan dianggap bersifat rahasia bagi seseorang, seperti data mengenai penyakit, data tindak kriminal seseorang, lama hukuman, status disabilitas seseorang, bahkan gaji juga bisa dianggap sebagai atribut sensitif bagi orang tertentu.

2.1.7.4. *Non-sensitive Attribute*

Non-Sensitive Attribute merupakan atribut yang tidak termasuk ke dalam tiga kategori sebelumnya.

2.1.8. K-Anonymity

Model k-anonimitas diusulkan oleh Samarati (2001) dan Sweeney (2002) merupakan pendekatan perlindungan privasi yang simpel dan praktis untuk melindungi data dari identifikasi individu. Model k-anonymity bekerja dengan memastikan setiap *record* pada tabel identik dengan setidaknya (k-1) *record* lainnya sehubungan dengan serangkaian fitur yang berhubungan dengan privasi, yang biasa disebut *quasi-identifier*, yang berpotensi digunakan untuk identifikasi individu dengan menghubugkan atribut yang ada dengan data set di luar (Lin & Wei, 2008, diacu dalam Kabir et al. 2010: 93).

Nilai k dalam model k-anonimitas ditentukan oleh pengguna sesuai dengan tujuan aplikasi mereka. Dengan menegakkan persyaratan k-anonimitas, dijamin bahwa meskipun lawan tahu bahwa tabel k-anonim berisi catatan individu tertentu dan juga tahu beberapa kuasi-identifier nilai atribut individu, ia tidak bisa menentukan yang mana catatan dalam tabel yang sesuai dengan individu dengan probabilitas lebih besar dari 1 / k (Byun et al. 2007 diacu dalam Kabir et al. 2010: 93). Hal ini menunjukkan bahwa semakin besar nilai-nilai k, lawan memiliki

sedikit kesempatan untuk menentukan informasi identitas pribadi dan data yang dilindungi. Di sisi lain, jika k-nilai yang terlalu besar itu menimbulkan informasi lebih banyak kerugian. Jadi, nilai k dari k-*anonymitas* tidak boleh terlalu kecil atau terlalu besar (Kabir et al. 2010: 93).

Menurut Kabir et al. (2010: 94) tidak diragukan lagi, penganoniman pasti disertai dengan hilangnya informasi (*information loss*). Agar berguna dalam prakteknya, dataset harus tetap informatif sebanyak mungkin. Oleh karena itu, perlu untuk mempertimbangkan secara mendalam antara privasi dan kehilangan informasi. Untuk meminimalkan *informatiom loss* karena k-*anonymity*, semua catatan dibagi menjadi beberapa kelompok sehingga masing-masing kelompok berisi setidaknya sebanyak k-catatan serupa yang berhubungan dengan *quasi-identifier* dan kemudian catatan dalam setiap kelompaok di generalisasi atau disuppresi sedemikian rupa sehingga nilai-nilai pada setiap *quasi-identifie*r adalah sama. Beberapa kelompok yang serupa disebut sebagai *cluster*.

Jadi, model k-*anonymity* dapat diatasi dari sudut pandang *clustering*. Namun model k-*anonymity* dapat menghasilkan informasi sensitif dengan dua serangan, yaitu serangan homogenitas (*Homogenity Attack*) dan serangan pengetahuan latar belakang (*Background Attack*) (Machanavajjhala et al. 2006, diacu dalam Kabir et al. 2010: 94). Misalnya, Jack dan Ron adalah dua tetangga yang tidak akrab. Jack tahu bahwa Ron pergi ke rumah sakit baru-baru ini dan mencoba untuk mencari tahu penyakit yang diderita Ron. Jack menemukan tabel 3-*anonymity*. Dia tahu bahwa Ron berusia 39 tahun dan tinggal di pinggiran kota dengan kode pos 4350. Ron harus menjadi *record* 4, 5 atau 6. Ketiga pasien yang menderita diabetes, sehingga Jack tahu pasti bahwa Ron menderita dari diabetes. Nilai-nilai homogen

dalam atribut sensitif dari kelompok k-*anonymity* dapat memberitahu informasi pribadi. Dengan demikian k-*anonymity* tidak dapat melindungi individu dari serangan *background attack.*

2.1.9. Systematic Clustering

Menurut Kabir et al. (2011: 58), permasalahan pada *systematic clustering* adalah menemukan satu set kelompok dari pemberian set n-*record* dimana setiap kelompok mengandung paling sedikit k ($k \leq n$) *record* (dimana *record* diseleksi secara sistematik dan termasuk ke dalam kelompok yang menyebabkan *information loss* paling sedikit) dan jumlah dari semua jarak intra-kelompok menjadi kecil. Secara spesifik, jika η merupakan set dari n-*record* dan k merupakan spesifikasi paramater penganoniman, solusi optimal dari permasalahan *systematic clustering* adalah pengelompokan setiap dataset.

Setiap set pengelompokkan dibangun sedemikian rupa dimana pengelompokan satu sama lain saling eksklusif, jumlah *record* dari semua catatan ialah sama dengan total bilangan di *record*, dan besar dari setiap kelompok paling sedikit adalah k yang memenuhi kriteria k-*anonymity*. Permasalahannya mencoba untuk meminimalkan jumlah dari jarak intra-kelompok dimana jarak dari intra-kelompok pada *cluster* menetapkannya sebagai jarak maksimum antara dua *record* di kelompok (Kabir et al., 2011: 58).

Menurut Kabir et al. (2011: 59), algoritma dari *systematic clustering* ini sendiri adalah pertama-tama kita harus mengecualikan semua catatan individual yang tidak berhubungan dengan privasi. Kemudian urutkan semua catatan berdasarkan *quasi-identifier*nya dan identifikasi persamaan derajat kelasnya. Lalu pilih *record* secara acak untuk *record* k pertama sebagai umpan dari kelompok

pertama. Pilih *record* lain dari *record* k pertama dan tambahkan ke kelompok yang *infromation loss*nya sedikit. Jika terdapat ukuran *cluster* yang persis sebesar k, maka berhenti menambahkan *record* untuk kelompok tersebut dan lanjutkan proses sampai semua *record* dari *record* k pertama selesai. Hanya *record* pertama yang dipilih secara random dan *record* selanjutnya dari jalur sistematik yang memiliki waktu eksekusi sedikit. Untuk lebih jelasnya berikut dapat dilihat pada Lampiran 1 halaman 49.

2.1.10. Greedy K-Member

Menurut Byun et al. (2006: 194), pemecahan masalah dari *clustering k-member* adalah memberikan catatan apakah ada skema pengelompokan, seperti:

a. $|e_i| \geq k, 1 < k \leq n$: ukuran setiap cluster lebih besar dari atau sama dengan bilangan bulat k positif, dan
b. $\sum_{i=1,...,l} IL(ei) < c$, $c > 0$: total-IL dari skema pengelompokan kurang dari konstanta c positif.

Byun et al (2016: 195) juga memaparkan beberapa teorema menganai Algoritma *Greedy K-Member yaitu:*

1. Pemecahan masalah *clustering k-member* adalah NP-*complete*.
2. n merupakan total bilangan dari input *record* dan k merupakan parameter penganoniman. Setiap *cluster* pada *greedy k-member clustering* menemukan paling sedikit k-*record*, tapi tidak lebih dari 2k-1 *record*.
3. n merupakan total bilangan dari input *record* dan k merupakan parameter penganoniman. Kompleksitas waktu dari algoritma *greedy k-member* adalah $O(n^2)$.

2.1.11. Information Loss

Menurut Kabir et al. (2010: 95), penganoniman berdasarkan generalisasi dan suppresi biasanya menyebabakan *information loss*. *Information loss* ini sendiri maksudnya adalah infromasi yang hilang atau infromasi yang datanya sudah tidak memiliki makna. Pertanyaannya adalah seberapa banyak informasi yang hilang? Hal ini dapt dihitung dengan menggunakan rumus *Information Loss (ILoss).*

Rumus *information loss* di sini berdasarkan pada teknik penghitungan yang dilakukan oleh Byun et al (2007). Dimana lambang η menunjukan kumpulan *record* dengan *r* merupakan bilangan *quasi-identifier* N_1, N_2,…, N_r dan *s* merupakan kategori *quasi-identifier* C_1, C_2,…, C_s. Untuk menggeneralisasi nilai dari kategori atribut C_i (i= 1, 2, …, *s*) maka τC_i merupakan pohon taxonomy yang akan menetapkan domain C_i.

Adapula $N_{i\,max}$ dan $N_{i\,min}$ yang merupakan nilai tertinggi dan terendah dari *record* Ω serta $\eta_{N_{imax}}$ dan $\eta_{N_{imin}}$ yang merupakan nilai tertinggi dan terendah dari *record* η sehubungan dengan atribut numerik N_i = (i = 1, 2, …, r) dan $\cup C_i$ menjadi nilai set satuan pada Ω terhadap atribut kategori C_i (i= 1, 2, …, *s*). Selanjutnya maka akan didapatkan rumus *Information Loss*(Ω) sebagai berikut:

$$IL(\Omega) = \mid \Omega \mid . \left(\sum_{i=1}^{r} \frac{N_{i_{max}} - N_{i_{min}}}{\eta N_{i_{max}} - \eta N_{i_{min}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup C_j))}{H(\tau C_j)} \right) \text{.....(1)}$$

Di mana Ω adalah banyaknya data dalam suatu cluster, N adalah quasi identifier numerical. Sedangkan C adalah quasi identifier kategorial. Sehingga $N_{i\,max}$ adalah Nilai numerik maksimal yang ada dalam cluster. Sedangkan $\eta_{N_{imax}}$

adalah nilai numerik terbesar dari suatu data. τ(∪Cj) adalah nilai taksonomi terkecil dari suatu akar pohon kategorial. ∪Cj and H(τ) adalah taksonomi tertinggi dari pohon τ. Tujuan utama dari teknik clustering ialah untuk membangun cluster sedemikian rupa sehingga total *information loss* dari η akan minimum.

2.2. Metode dan Proses Penelitian

Berbagai jenis penelitian yang berkaitan dengan penelitian ini adalah:

2.2.1. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression (Latanya Sweeney)

Jurnal ini membahas bagaimana mencapai perlindungan privasi K-A*nonymity* dengan menggunakan generalisasi dan suppresi. Untuk dapat merepresentasikan jurnal ini, penulis menggunakan Metode *Preferred Minimal Generalization Algorithm* (*MinGen*). Metode tersebut menggabungkan kedua teknik, yaitu teknik generalisasi untuk mengganti (merekam) nilai dengan tingkat spesifikasi yang rendah, dan teknik supresi untuk tidak mengubah nilai informasi secara keseluruhan. Metode tersebut digunakan untuk menemukan generalisasi minimal dari sebuah tabel menggunakan model *k-Anonymity* dengan distorsi yang minimum. Fokus penelitian ini adalah membandingkan Algoritma Datafly dan µ-Argus. Disini algoritma memutuskan bagaimana mengubah data supaya memiliki pengaruh yang sedikit pada kemampuan data untuk suatu tugas khusus.

Data dalam penelitian tersebut bersumber dari *Laboratory for International Data Privacy, Carnegie Mellon University*. Hasil dari penelitian ini adalah algoritma *Datafly* dapat mencapai tingkat generalisasi sesuai syarat *k-Anonymity* namun tidak memenuhi syarat *k-minimal distortions*, sedangkan pada algoritma *µ-*

Argus tidak dapat memenuh kedua syarat tersebut karena kehilangan efisiensi proses komputasinya.

Algoritma *Datafly* dan *µ-Argus* dapat melakukan generalisasi jika tidak terdapat syarat *k-minimal distortions* karena keduanya melakukan proses generalisasi pada level atributnya. Hal tersebut menyebabkan penilaian tingkat presisi keduanya rendah. Diperlukan kerja ekstra untuk memperbaiki pendekatan berbasis heuristik pada kedua algoritma tersebut.

2.2.2. Efficient Systematic Clustering Method For K-Anonymization (Md. Enamul Kabir, Hua Wang, dan Elisa Bertino)

Jurnal ini membahas penggunaan Metode *Systematic Clustering* secara efisien untuk K-*Anonymity*. Jurnal lebih menekankan bagaimana penggunaan *clustering* untuk mendapatkan *information loss* yang sedikit dengan tetap menjaga privasi data. Pada teknik ini data yang mirip akan dikelompokan bersama dan kemudian menganonimkan setiap data yang ada.

Pada algoritma ini, pertama data harus diurutkan terlebuh dahulu baru kemudian di *cluster* menggunakan generalisasi maupun suppresi. Selanjutnya dihitung *information loss*nya untuk melihat apakah metode ini akan menghasilkan kehilangan data yang sedikit atau tidak. Pada jurnal ini, penullis membandingkan *systematic clustering* dengan algortima *k-member* yang mana pada jurnal yang lain diketahui bahwa algoritma *k-member* memiliki *information loss* yang lebih sedikit dibandingkan dengan algoritma Mondrian.

Kedua eksperimen tersebut di implementasikan dengan menggunakan bahasa Excel VB Programming. Sistem operasi yang digunakan adalah Microsoft Windows XP Professional Versi 2002 dengan 3.20GHz Pentium (R) D CPU

dengan RAM 2 GB. Dataset yang digunakan pada jurnal ialah Adult dataset yang di ambil dari UCI *Machine Learning Repository*.

Hasil yang diperoleh dari jurnal ini adalah *information loss* untuk setiap algoritma meningkat seiring dengan meningkatnya nilai k. logika dibelakang ini ialah ketika nilai k meningkat maka *cluster* membutuhkan generalisasi yang maksimum dan menyebabkan hilangnya informasi. Hal terpenting dalam memilih metode *clustering* terbaik dalah yang memiliki *information loss* paling sedikit.

Selanjutnya untuk waktu eksekusi menunjukkan bahwa waktu yang dibutuhkan untuk algoritma *systematic clustering* lebih sedikit dari algortima *k-member*. Algortima *Greedy K-Member* membutuhkan waktu yang cukup banyak untuk menseleksi *record* dari input dataset. Dengan demikian dapat disimpulkan bahwa metode yang diusulkan, yaitu *systematic clustering* lebih unggu dibandingkan *greedy k-member* baik dari segi *information loss* dan waktu eksekusi.

2.2.3. Efficient k-Anonymization Using Clustering Techniques (Ji-Won Byun, Ashish Kamra, Elisa Bertino, dan Ninghui Li)

Jurnal ini ditulis karena sudah banyaknya matriks kualitas yang diusulkan untuk hirarki berbasis generalisasi. Sebuah metrik yang dapat mengukur secara tepat *information loss* yang diperkenalkan oleh hirarki generalisasi belum ada yang memperkenalkan. Untuk alasan ini, penulis menetapkan metrik kualitas data untuk hirarki generalisasi, yang disebut *loss metric*. Penulis juga menunjukkan bahwa dengan sedikit modifikasi, algortima tersebut bisa mengurangi kesalahan klasifikasi secara efektif.

Hal-hal yang dibahas pada jurnal ini adalah me*review* konsep dasar dari model k-*anonymity* serta mensurvey teknik yang sudah ada, menyatakan permasalahan dari k-*anonymity* sebagai permasalahan *clustering* dan memperkenalkan pendekatan yang akan dibahas, kemuidan mengevaluasi pendekatan tersebut berdasarkan hasil eksperimen, dan terakhir menyimpulkan hasil diskusi.

Tujuan utama dari percobaan ini adalah untuk menyelidiki kinerja pendekatan yang dilakukan dalam hal kualitas data, efisiensi, dan skalabilitas. Untuk secara akurat mengevaluasi pendekatan ini, penulis juga membandingkan implementasi kami dengan algoritma lain, yaitu Algortima Partioning.

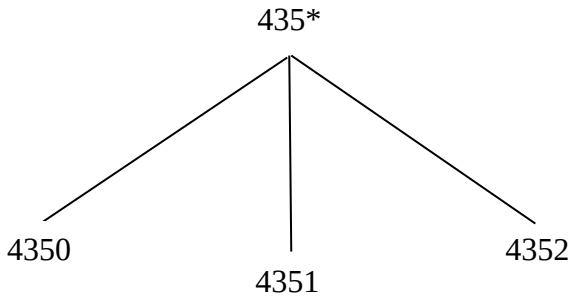
Dalam tulisan ini, penulis mengusulkan sebuah efisiensi algoritma k-*anonymity* dengan mengubah masalah k-anonimitas menjadi masalah pengelompokan k-*member*. Penulis juga mengusulkan dua elemen penting dari *clustering*, yaitu jarak dan biaya fungsi yang secara khusus dirancang untuk masalah k-*anonymity*. Penulis menekankan bahwa *cost metric*, ILmetric, secara natural menangkap distorsi data yang diperkenalkan oleh proses generalisasi dan cukup umum untuk digunakan sebagai metrik kualitas data untuk dataset k-anonim.

2.3. Konsep dan Prosedur Penelitian

Privasi data berarti menyembunyikan data yang bersifat sensitif guna menjaga kerahasiaan suatu perusahaan ataupun organisasi. Data yang bersifat sensitif ini atau disebut juga dengan atribut sensitif biasanya berupa aib ataupun informasi yang bersifat rahasia sehingga bisa saja disalahgunakan oleh pihak lain. Untuk menjaga privasi data suatu perusahaan pada kesempatan kali ini digunakan

model K-*Anonymity*. Terdapat dua macam teknik yang dapat digunakan untuk dapat menjaga privasi data dengan model ini, yaitu Generalisasi dan Suppresi.

Pada teknik generalisasi, data yang ada akan diubah menjadi *range* yang lebih luas dengan menggunakan *taxonomy tree*, sebagai contoh dapat dilihat pada Gambar 2.3:

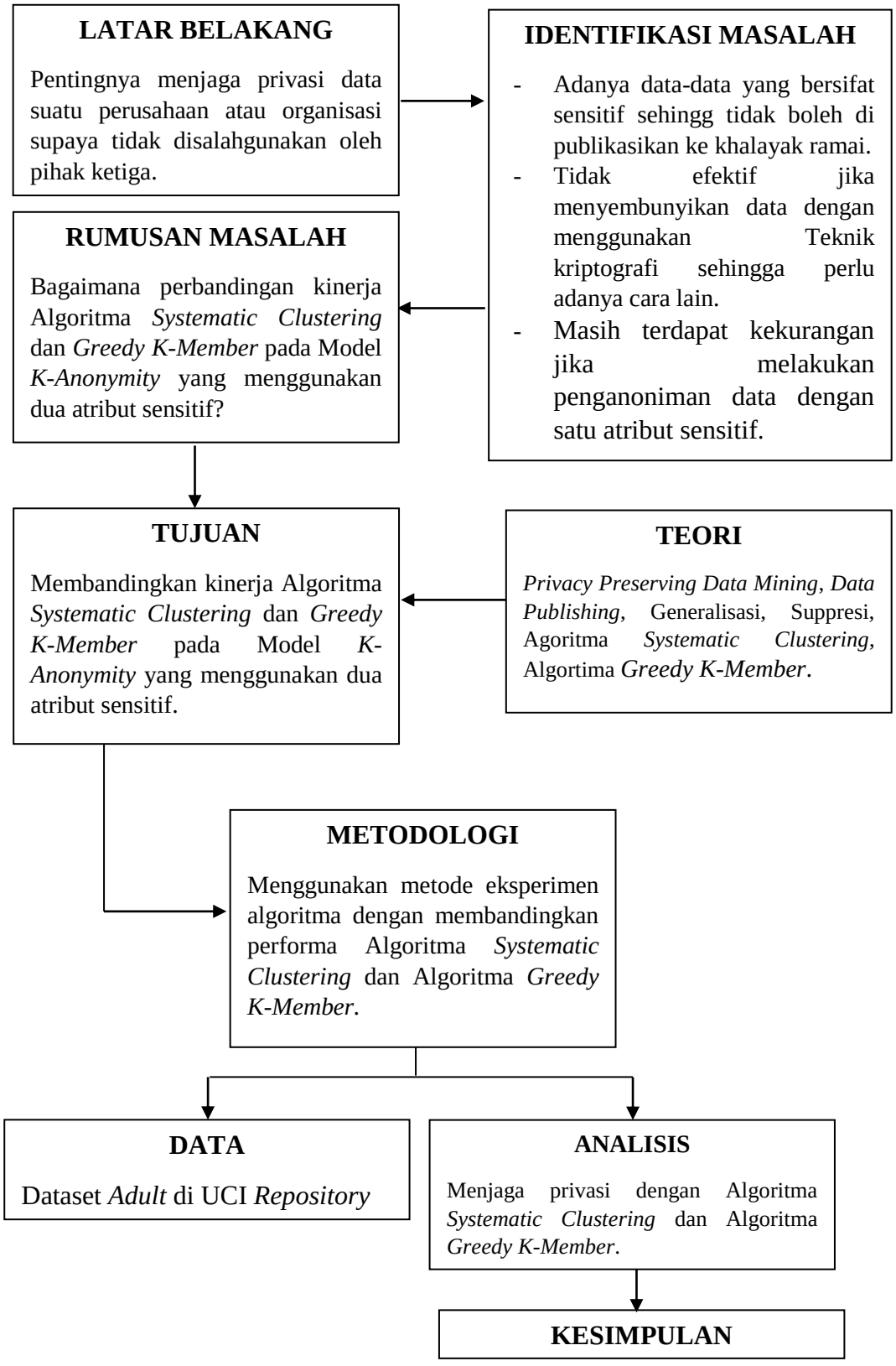


Gambar 2.3 *Taxonomy Tree Zip Code*

Teknik kedua adalah suppresi. Pada suppresi dibagi lagi menjadi dua bagian, yaitu *local recording* dan *global recording*. Pada intinya generalisasi maupun suppresi sama saja. Akan tetapi ketika menggunakan generalisasi maupun suppresi makan nantinya akan ada data yang tidak lagi memiliki makna yang mengakibatkan tidak dapat diketahui informasi apa yang terkandung di dalamnya. Kejadian seperti ini disebut juga dengan *Information Loss* atau hilangnya informasi.

Mencoba mengatasi permasalahan yang ada maka diterapkanlah algoritma yang ada untuk mencari *information loss* terkecil. Algoritma yang digunakan adalah *Systematic Clustering* dan *Greeedy K-Member*. Kedua algoritma tersebut memiliki tahapan yang berbeda, jika pada *Systematic Clustering* data yang ada harus disusun terlebih dahulu baru kemudian di generalisasi dan di hitung *Information Loss*nya, maka pada *Greedy K-Member* data yang ada langsung *cluster* lalu di cari *Information Loss*nya. Hasilnya algoritma yang memiliki

Information Loss paling sedikit merupakan algoritma yang lebih baik untuk diterapakan pada *Privacy Preserving Data Publishing*. Untuk lebih jelasnya dapat dilihat pada gambar di bawah:



Gambar 2.4. Bagan Kerangka Berpikir