BAB III

METODOLOGI PENELITIAN

3.1. Tempat dan Waktu Penelitian

Tempat dan waktu penelitian merupakan hal yang perlu diperhatikan dalam penelitian, yaitu untuk mempermudah penulis dalam melakukan penelitian dan sesuai dengan harapan penulis. Penelitian dilaksanakan di Laboratorium Multimedia Gedung L2 (*ex.* Jurusan Teknik Elektro) lantai 3, Fakultas Teknik. Universitas Negeri Jakarta. Waktu Penelitian dimulai tanggal 10 Agustus sampai tanggal 16 Desember 2016.

3.2. Alat dan Bahan Penelitian

Alat dan bahan yang digunakan dalam penelitian antara lain:

3.2.1 *Alat*

Alat yang digunakan dalam penelitian yaitu Laptop Lenovo™ Ideapad™ 310-14iSK dengan spesifikasi yaitu:

3.2.1.1 Perangkat Keras (Hardware)

Perangkat keras yang digunakan antara lain:

1. Prosesor Intel® Core(TM) i5-6200U CPU @ 2.30 GHz 2.40 GHz.
2. Resolusi layar 14.0" *FHD LED Glare Wedge* 1920x1080
3. Sistem grafis NVIDIA GeForce 920MX 2GB
4. *Random Access Memory* (RAM) 4GB PC4-17000 DDR4-2133 MHz.
5. *Hard Drive* WDC WD10JPCX-24UE4T0 1 TB 5400-rpm

3.2.1.2 Perangkat Lunak (Software)

Perangkat lunak yang digunakan antara lain:

1. *Windows 10 Home Single Languange* 64-bit © 2016.

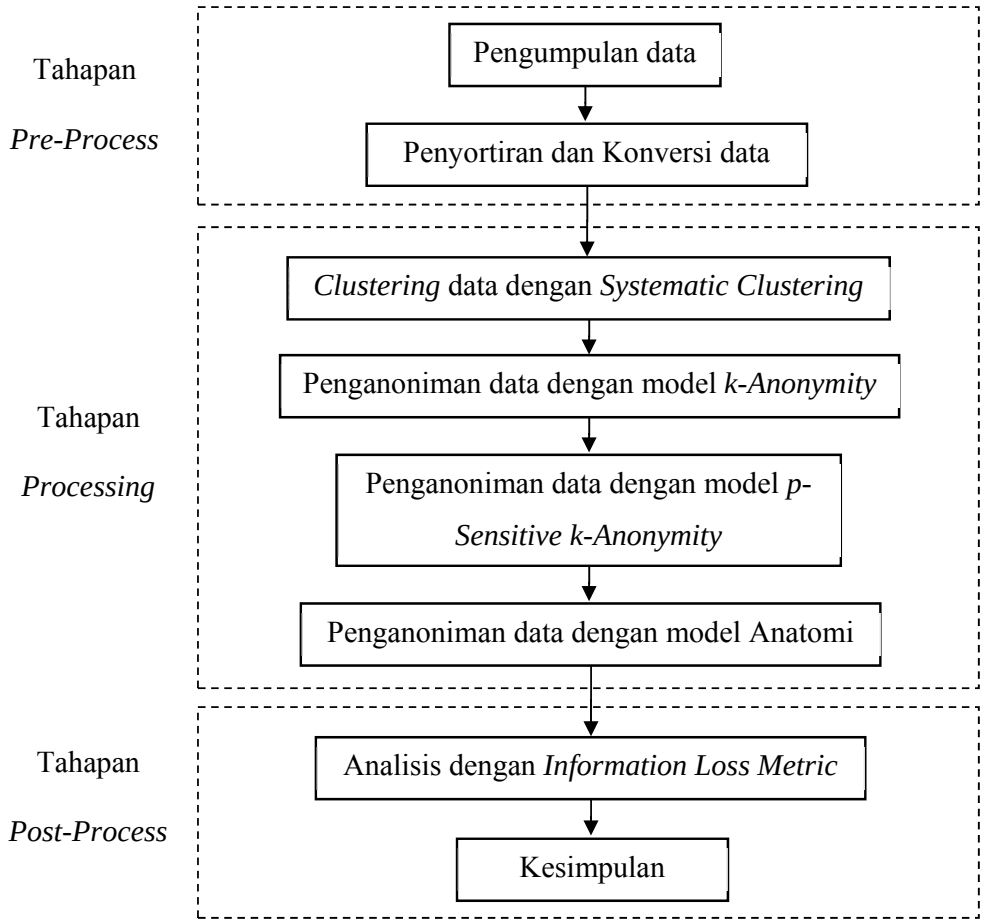
2. *Editor Java NetBeans IDE* 8.1 (*Build* 201510222201).
3. *Microsoft Office Professional Plus* 2016 versi *en-us.*
4. XAMPP versi 7.0.8-0 sebagai *web server* basis data.
5. *Mozilla Firefox* versi 50.1.0 sebagai antarmuka basis data.

3.2.2 *Bahan*

Bahan yang digunakan dalam penelitian yaitu *dataset* bersumber dari *Adult dataset*, *UC Irvine Machine Learning Repository*.

3.3. Diagram Alir Penelitian

Gambar 3.1 merupakan tahapan untuk mengukur kinerja model Anatomi yang digunakan dalam penelitian:



Gambar 3.1. Diagram Alir Penelitian

3.3.1 *Tahapan Pre-Process*

Pada tahapan *Pre-Process* terdiri dari beberapa sub tahapan, yaitu:

3.3.1.1 Pengumpulan data

Pengumpulan data bertujuan untuk memperoleh informasi yang diperlukan untuk mencapai tujuan penelitian. Data yang digunakan dalam penelitian tidak diperoleh secara langsung melalui observasi atau mengisi kuisioner, namun data dalam penelitian bersumber dari sebuah data *repository* yaitu *Adult dataset*, *UC Irvine (UCI) Machine Learning Repository*. Alasan pengumpulan data dari *UCI* karena data tersebut sangat representatif sebagai *microdata*, telah didedikasian untuk penelitian yang berkaitan dengan *Data Mining* karena menyediakan 351 *dataset,* dan termasuk dalam kategori *Open Source Repository*.

Pemilihan *Adult dataset* sebagai sumber data penelitian karena bersumber dari *US Census Bureau* pada tahun 1996, dan telah banyak digunakan untuk penelitian yang berkaitan dengan *Data Mining*. Jika ditinjau dari segi *tupel* data tersebut memiliki banyak atribut yang dapat digolongkan menjadi atribut sensitif dan *Quasi-Identifier*, memiliki variasi data yang tergolong banyak yang dapat digeneralisasi, serta tingkat *error* yang rendah ketika diteliti. Beberapa penelitian yang menggunakan *dataset Adult*, yaitu:

1. Penelitian Bianca Zadrozny yang berjudul *Learning and Evaluating Classifiers under Sample Selection Bias*. (ICML: 2004)
2. Penelitian Bart Hamers dan J. A. K Suykens yang berjudul *Coupled Transductive Ensemble Learning of Kernel Models*. (Bart De Moor: 2003)
3. Penelitian Traian Marius Truta dan Bindu Vinay yang berjudul *Privacy Protection: p-Sensitive k-Anonymity Property*. (IEEE: 2006)

3.3.1.2 Penyortiran dan Konversi Data

Adult dataset merupakan data dengan tupel yang ditulis dalam satu baris yang dipisahkan dengan tanda koma dan diakhiri dengan tanda titik atau *Comma-Separated Value* (CSV) dengan populasi data sebanyak 48842 *field*. Permasalahan yang ditemukan dalam *Adult dataset* adalah banyaknya nilai dalam satu tupel untuk satu atribut tidak diketahui atau tanda tanya (?). Untuk mengatasi permasalahan tersebut, maka data disimpan dalam format *.csv* menggunakan *Microsoft Excel*. Langkah dalam proses penyortiran dan konversi data, yaitu:

1. Buat basis data dalam DBMS dengan struktur yang sama dengan data dalam *Microsoft Excel*, kemudian data diimpor ke DBMS.
2. Data dalam DBMS disortir melalui bahasa pemrograman *Java* untuk menghilangkan *record* yang tidak diketahui nilainya. Penggunaan *dataset Adult* dibatasi sebanyak 12.600 *record* data sebagai sampel data.
3. Pilih data yang akan dimasukkan ke dalam atribut *Quasi-Identifier* dan *Sensitive* dan simpan pada tabel yang berbeda. Atribut *Quasi-Identifier* penelitian: *Age*, *Workclass*, dan *Race*, sedangkan untuk atribut sensitifnya adalah *Marital-status*.

3.3.2 *Tahapan Processing*

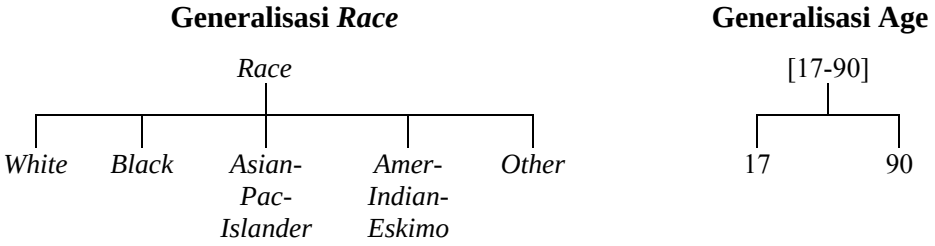
Pada tahapan *Processing* terdiri dari beberapa sub tahapan, yaitu:

3.3.2.1 *Clustering* data dengan *Systematic Clustering*

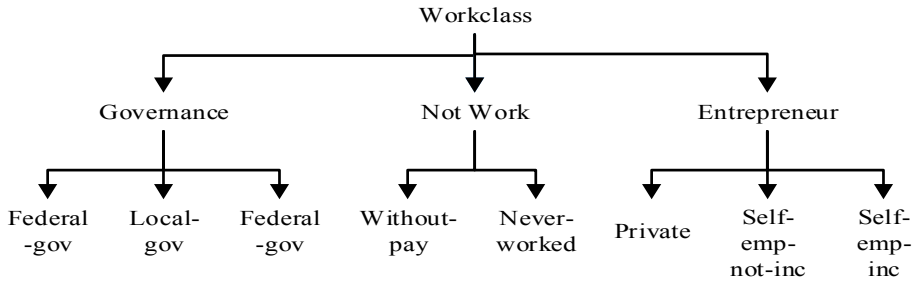
Pada tahapan *clustering* data dengan *Systematic Clustering*, data yang telah dikelompokkan menjadi atribut *Quasi-Identifier* dan *Sensitive* akan dilakukan *Clustering* data dengan *Systematic Clustering*. Langkah dalam proses *clustering* data dengan *Systematic Clustering* yaitu:

1. Tentukan banyaknya data dalam satu *cluster*. Ditetapkan sebanyak dua sampai sepuluh *tuple* dalam satu *cluster*.
2. Lakukan pengurutan data penelitian berdasarkan atribut *Age* dari data yang terkecil ke terbesar (*ascending*).
3. Tambahkan atribut pada data yaitu *Cluster* pada setiap *tuple* yang telah masuk ke dalam *cluster*. Hasil dari tahapan ini akan disimpan dalam tabel yang berbeda dengan tabel pada tahap sebelumnya.

3.3.2.2 Penganoniman data dengan model *k-Anonymity*



Gambar 3.2 Generalisasi *Race* dan *Age*



Gambar 3.3 Generalisasi *Workclass*

Tahapan penganoniman data dengan model *k-Anonymity* merupakan bagian dalam penganoniman data dengan langkah-langkah sebagai berikut:

1. Tentukan jumlah *k* atau *cluster*, yaitu $3 \leq k \leq 10$.
2. Tentukan *Taxonomy Tree* dari proses generalisasi dan supresi atribut *Quasi-Identifier*. Tingkatan (*node*) pada atribut *Age* adalah satu bahkan nol, karena telah diurutkan pada tahapan sebelumnya. *Node* pada atribut *Race* sama dengan

Age, yaitu satu, karena hanya terdapat lima kategori yang dapat digeneralisasi menjadi *Race*. Atribut *Workclass* terdapat delapan kategori, sehingga perlu digeneralisasi sebanyak dua kali atau dua *node*.

Gambar 3.2 dan Gambar 3.3 merupakan *Taxonomy Tree* yang digunakan dalam penelitian. Langkah terakhir adalah melakukan generalisasi dan supresi pada atribut *Quasi-Identifier*. Teknik yang digunakan adalah *Local Recording* pada atribut *Workclass* dan *Age*, serta teknik *Global Recording* pada atribut *Race* yang akan disimpan pada tabel yang sama dengan tabel sebelumnya, karena pada tahapan ini hanya melakukan generalisasi dan supresi pada data.

3.3.2.3 Penganoniman data dengan model *p-Sensitive k-Anonymity*

Pada tahapan penganoniman data dengan model *p-Sensitive k-Anonymity* akan dilakukan pengecekan atribut sensitif dalam satu *cluster*, untuk mencegah terjadinya atribut sensitif bernilai sama dalam satu *cluster*. Jika ditemukan dalam satu *cluster* terdapat nilai atribut sensitif sama, maka salah satu *tuple* dalam *cluster* tersebut akan ditukar dengan *tuple* lain pada *cluster* berikutnya, sedangkan *tuple* pada *cluster* berikutnya akan berpindah ke *cluster* sebelumnya. Toleransi banyaknya nilai yang sama dalam satu *cluster* dinyatakan dengan *p*. Nilai *p* mengikuti nilai *k* yang telah didefinisikan sebelumnya, yaitu $3 \leq p \leq 10$. Hasil dari penganoniman data dengan model *p-Sensitive k-Anonymity* akan disimpan pada tabel yang berbeda. Nilai *ID* pada tabel sebelumnya diambil, kemudian dicocokkan dengan tabel awal *record* akan disimpan dalam tabel baru.

Hal tersebut bertujuan untuk mengecek kesesuaian isi tabel dengan ketentuan yang dibuat. Jika isi pada tabel baru sesuai dengan ketentuan tahapan ini, maka data tersebut akan dilakukan generalisasi dan supresi ulang dan disimpan pada tabel

yang berbeda, karena tabel yang diacu untuk melakukan langkah ini akan digunakan pada tahapan berikutnya.

3.3.2.4 Penganoniman data dengan model Anatomi

Pada tahapan penganoniman data dengan model Anatomi, tabel hasil penganoniman data dengan model *p-Sensitive k-Anonymity* yang belum dilakukan generalisasi dan supresi akan dipecah menjadi dua tabel, yaitu menjadi *Quasi-Identifier Table* (*QIT*) dan *Sensitive Table* (*ST*). Langkah dalam penganoniman data dengan model Anatomi, yaitu:

1. Lakukan pengecekan banyaknya atribut sensitif beserta dengan jumlah tiap atribut sensitif dalam satu *cluster*, kemudian simpan dalam *ST* dengan format atribut: *Cluster*, Atribut sensitif, dan Jumlah.
2. Hapus seluruh atribut sensitif dalam *QIT* dan diganti dengan *Cluster*. Hasil dari tahapan ini disimpan dalam dua tabel yang berbeda, yaitu QIT dan ST.

3.3.3 Tahapan *Post-Process*

Pada tahapan *Post-Process* terdiri dari beberapa sub tahapan, yaitu:

3.3.3.1 Analisis dengan *Information Loss Metric*

Pada tahapan analisis dengan *Information Loss Metric*, hasil yang diperoleh pada tahapan sebelumnya dilakukan analisis yang bertujuan untuk membandingkan efektivitas dua model yang digunakan dalam penelitian, yaitu model *p-Sensitive k-Anonymity* dengan model Anatomi. Tahapan ini dilakukan sebanyak delapan kali, yaitu kedua model diberikan perlakuan penetapan *cluster* yang sama, yaitu mengikuti nilai k yang telah ditentukan $3 \leq k \leq 10$. Parameter yang digunakan pada tahapan ini adalah mengukur banyaknya informasi yang hilang menggunakan *Information Loss Metric*.

3.3.3.2 Kesimpulan

Tahapan kesimpulan merupakan penarikan kesimpulan hasil analisis pada tahapan *Processing*. Kesimpulan yang dibuat mengacu pada tujuan penelitian yang telah didefinisikan di awal penelitian serta hasil yang didapat dalam penelitian. Penulis juga memberikan deskripsi singkat mengenai tahapan dalam membentuk model *p-Sensitive k-Anonymity* dan model Anatomi serta kendala yang dialami dalam mengimplementasikan kedua model tersebut.

3.4. Teknik dan Prosedur Pengumpulan Data

Teknik yang digunakan dalam pengumpulan data penelitian adalah *Non-Participant Observation* atau disebut sebagai Studi Dokumenter. Teknik tersebut merupakan kegiatan observasi dimana penulis tidak secara langsung mengikuti atau terlibat dalam kegiatan yang sedang diamatinya, namun pengumpulan data dalam penelitian didapatkan melalui catatan-catatan pribadi atau hasil karya seseorang yang dibagikan dalam pangkalan data penelitian. Berdasarkan Gambar 3.1, penelitian diawali dengan pengumpulan data bersumber dari *Adult dataset*, melalui *website UC Irvine (UCI) Machine Learning Repository.* Data yang diperoleh dari *website UCI* memiliki format .data, yaitu data tersebut berbentuk baris kalimat yang dipisahkan dengan tanda koma (,) dan diakhiri dengan tanda titik (.). Format tersebut merupakan format baku dalam *website UCI*, namun format tersebut tidak dapat dibaca oleh DBMS jika tidak dilakukan konversi format data menjadi *CSV*. Setelah data dikonversi, maka data diimpor ke DBMS dan dinormalisasi dengan menghapus *record* yang memiliki *missing value* (?) pada atributnya.

Normalisasi data dilakukan untuk mempermudah proses anonim data, karena data yang akan dianonimkan harus memiliki nilai yang berarti dan dapat

digeneralisasi atau disupresi. Langkah selanjutnya ialah membatasi pengambilan data sebagai data penelitian yaitu dengan mengambil data pada *record* pertama sampai *record* ke-12.600. Langkah terakhir adalah menggolongkan atribut data penelitian menjadi atribut *Quasi-Identifier* dan *Sensitive*, merujuk pada penelitian *Traian Marius*, dkk. yang berjudul *Privacy Protection: p-Sensitive k-Anonymity Property* dan penelitian Pawan R. Bhaladhare dan Devesh C. Jinwala yang berjudul *Novel Approaches for Privacy Preserving Data Mining*.

3.5. Teknik Analisis Data

Teknik analisis data yang digunakan adalah *Information Loss Metric* melalui perbandingan dua model, yaitu model *p-Sensitive k-Anonymity* dengan model Anatomi. Teknik tersebut bertujuan untuk menganalisis hasil data berdasarkan banyaknya informasi yang hilang dalam proses anonim data. Perhitungan *Information Loss Metric* mengacu pada penelitian Ji-Won Byun, dkk. yang berjudul *Efficient k-Anonymization Using Clustering Techniques* atau menggunakan formula pada persamaan (2.1) dan (2.2).

Perhitungan dilakukan sebanyak delapan kali, sesuai dengan ketentuan pada tahapan proses, yaitu mengikuti nilai $k = 3 \leq k \leq 10$. Hasil perhitungan *Information Loss Metric* akan diubah menjadi bentuk grafik dan tabel untuk mempermudah analisis hasil penelitian. Langkah berikutnya adalah menginterpretasi hasil penelitian, yaitu menyimpulkan model mana yang menghasilkan nilai *Information Loss* terendah. Langkah terakhir adalah menjelaskan secara singkat mengenai tahapan dalam membangun privasi data menggunakan perpaduan model Anatomi dan *Systematic Clustering* serta kendala yang dialami penulis dalam penelitian.