BAB III
METODOLOGI PENELITIAN

3.1. Tempat dan Waktu Penelitian

Penelitian dilaksanakan di Laboratorium Komputer Program Studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta. Waktu penelitian dilakukan selama semester ganjil (105) tahun ajaran 2016/2017.

3.2. Alat dan Bahan Penelitian

Alat dan bahan yang digunakan untuk menunjang penelitian ini adalah:

1. Perangkat Keras

Laptop yang digunakan dalam penelitian ini adalah Acer One 14 Z1402-308T dengan spesifikasi sebagai berikut:

Tabel 3.1 Spesifikasi Laptop yang Digunakan

Processor Onboard	Intel Core i3 5005U-2.0Ghz
Memori Internal	2 GB
Tipe Grafis	Intel HD 5500
Ukuran Layar	14″
Resolusi Layar	1366 x768
Tipe Layar	HD Acer CineCrystal LED
Audio	Integrated
Hard Disk	500 GB
Optical Drive	8X DVD Super Multi Plus Drive (M-DISC Ready Drive)
Webcam	Acer Crystal Eye HD webcam, 1280 x 720 resolution
Koneksi	Gigabit Ethernet, Wake-on-LAN ready Wireless-AC, 802.11ac/a/b/g/n wireless LAN
Bluetooth	Bluetooth 4.0

Interfaces	SD card reader Two USB3.0 & One USB 2.0 ports HDMI™ port with HDCP support & VGA port
Sistem Operasi	DOS
Ragam Input Device	Multi-gesture touchpad, supporting two-finger scroll and pinch. Swipes access charms, application commands and previous applications
Baterai	37 Wh 2500 mAh 14.8 V 4-cell Li-ion battery pack
Dimensi	343 x 245 x 24.9 mm
Berat	2.1 Kg

2. Perangkat Lunak

- Windows 10 64 bit
- Netbeans IDE 8.1
- XAMPP versi 3.2.2
- Microsoft Office 2013

3. Internet

Interconnection networking atau yang lebih dikenal dengan internet merupakan jaringan komunikasi global yang terbuka dan menghubungkan jutaan bahkan milyaran komputer dengan menggunakan telepon, satelit, dll. Pada penelitian ini, internet digunakan untuk mencari data yang terdapat pada UCI *Machine Learning Repository*.

4. Dataset *Adult* dari UCI *Machine Learning Repository*

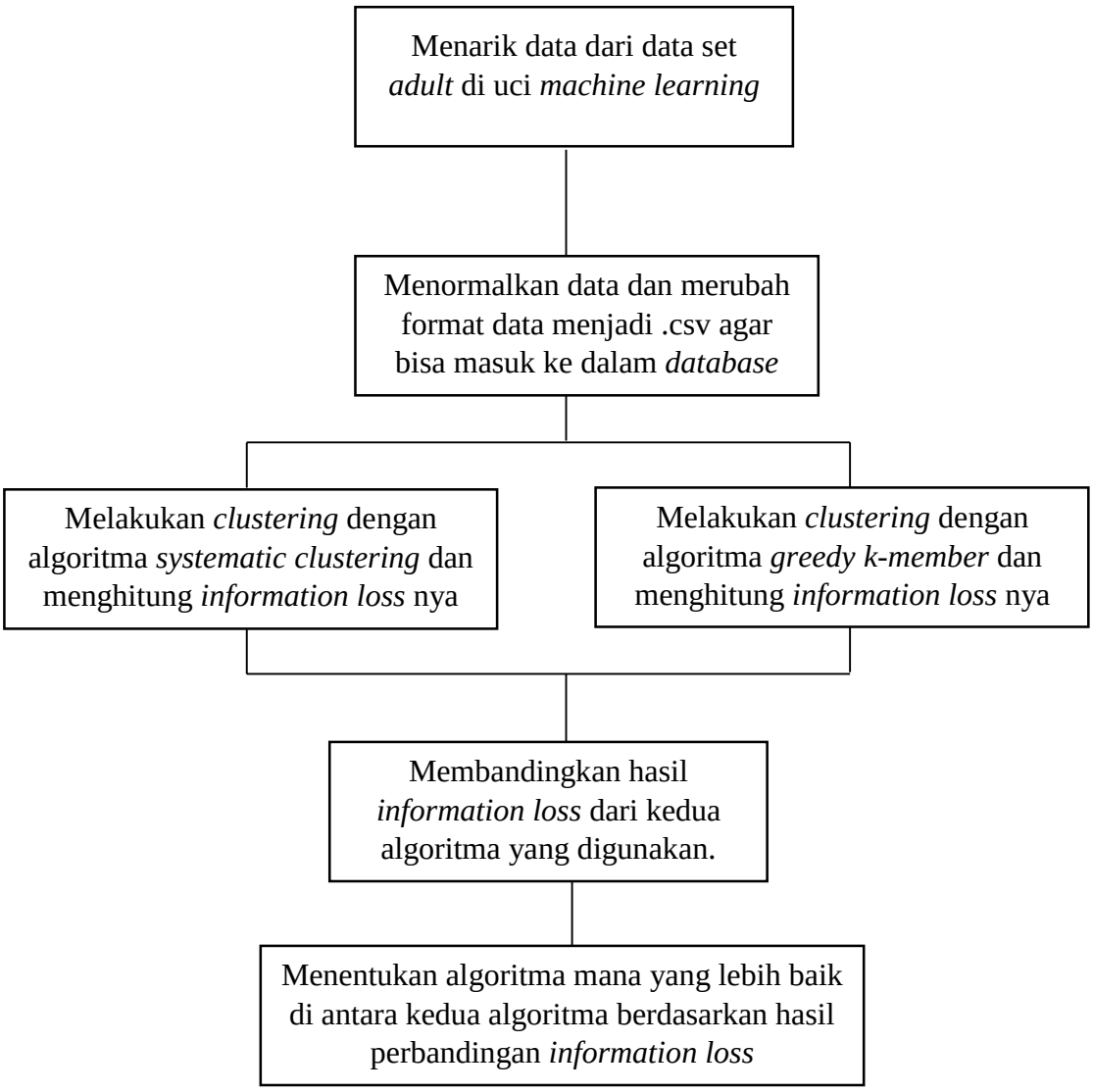
UCI Machine Learning Repository adalah kumpulan database, teori domain, dan data generator yang digunakan oleh komunitas *machine learning* untuk analisis empiris dari algoritma *machine learning*. Arsip ini diciptakan sebagai arsip ftp pada

tahun 1987 oleh David Aha dan rekannya mahasiswa pascasarjana di UC Irvine. Sejak saat itu, data di repository ini telah banyak digunakan oleh siswa, pendidik, dan peneliti di seluruh dunia sebagai sumber utama data set *machine learning*. Sebagai indikasi dari dampak arsip, arsip ini telah dikutip lebih dari 1000 kali sehingga menjadikannya salah satu dari 100 *papers* paling dikutip dalam semua ilmu komputer. Versi pada situs web yang ada saat ini dirancang pada tahun 2007 oleh Arthur Asuncion dan David Newman, dan proyek ini bekerjasama dengan Rexa.info di University of Massachusetts Amherst. Selain itu, repository ini juga mendapatkan dukungan dana dari National Science Foundation (A. Asuncion and D.J. Newman, 2007).

Oeh karena itu penulis memilih mengambil data dari UCI *Machine Learning Repository*. Data set yang di ambil ialah *Adult Data Set*. Orang yang membagikan data set ini adalah Ronny Kohavi dan Barry Becker. Data set ini juga telah digunakan di banyak penelitian-penelitian diantaranya adalah pada Prociding yang di lakukan oleh Ron Kohavi dengan judul Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid pada Proceedings of the Second International Conference on Knowledge Discovery and Data Mining tahun 1996. Untuk dapat mengetahui *Adult Data Set* lebih lengkap, maka dapat mengakses link berikut https://archive.ics.uci.edu/ml/datasets/Adult.

3.3. Diagram Alir Penelitian

Berikut adalah prosedur penelitian yang memperlihatkan langkah-langkah untuk model K-*Anonymity* dengan algoritma *Systematic Clustering* dan *Greedy K-Member*:



Gambar 3.1. Diagram Alir Penelitian

Hal pertama yang dilakukan pada penelitian ini adalah menarik data dari UCI *Machine Learning*. Data yang diambil adalah data set *adult*. Data set *adult* dari UCI awalnya akan disimpan dengan format .txt. Selanjutnya data dipindahkan ke *Microsoft Excel* dan disimpan dengan format .csv. Hal ini dilakukan agar data bisa di *import* ke *database*.

Setelah data di *import*, selanjutnya adalah melakukan *cluster* pada kedua algoritma, baik itu *systematic clustering* dan *greedy k-member*. Akan tetapi, pada *systematic clustering* sebelum dilakukan *cluster* akan dilakukan pengurutan terlebih dahulu, misalnya berdasarkan umur. Lalu dicari *information loss* dari kedua algoritma tersebut.

Proses penghitungan *information loss* ini tidak dilakukan sekali, tapi beberapa kali untuk memastikan akurat atau tidaknya hasil tersebut sebelum disimpulkan. Setelah itu, nilai *information loss* nya dibandingkan dan dilihat algoritma mana yang lebih baik.

3.4. Teknik dan Prosedur Pengumpulan Data

Data yang terdapat pada penelitian ini diambil dari UCI *Machine Learning Repository*. Di dalam web, data tersimpan dalam bentuk seperti gambar di bawah ini:

```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical,
Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse,
Exec-managerial, Husband, White, Male, 0, 0, 13, United-States,
<=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners,
Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-
cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-
specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-
managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
```

Gambar 3.2. Data Set *Adult* pada UCI *Machine Learning*

Data dalam bentuk seperti di atas tidak dapat dibaca oleh DBMS. Salah satu format yang bisa dibaca oleh DBMS adalah csv sehingga perlu untuk mengubah data di atas ke dalam bentuk csv. Setelah data format data di ubah ke csv, selanjutnya adalah

menghapus *missing value* yang terdapat pada atribut data. Hal ini dilakukan supaya memudahkan dalam proses *clustering*.

Langkah selanjutnya adalah mengurutkan data berdasarkan umur untuk memudahkan algoritma clustering. Setelah itu, data dalam bentuk csv tersebut akan diimport ke dalam DBMS MySQL sehingga formatnya akan berubah menjadi .sql. Data seperti inilah yang akan dikenal oleh JDBC di Java, sehingga nantinya data dengan format .sql inilah yang akan digunakan untuk *clustering* dengan menggunakan *Systematic Clustering* dan *Greedy K-Member*.

3.5. Teknik Analisis Data

Pada penelitian ini, analisis data dilakukan dengan mencari *information loss* dan *running time* pada setiap algoritma. Mencari *information loss* berguna untuk melihat seberapa banyak data yang sudah tidak memiliki makna. *Information loss* nantinya akan dihitung nilainya sebanyak tujuh kali dan dibandingkan antara *information loss* algoritma *systematic clustering* dengan *greedy k-member*. Nilai dari *information loss* ini nantinya akan di masukan ke dalam grafik sehingga mudah untuk dibaca. Semakin kecil *information loss*nya maka semakin bagus pula algoritmanya.

Sedangkan mencari *running time* bertujuan unutk melihat seberapa lama data tersebut dapat diolah dengan algoritma tertentu. Nantinya nilai dari *running time* juga akan dibandingkan dan dibuatkan grafiknya. Sama seperti *information* loss, semakin kecil waktu yang dibutuhkan untuk mengolah data pada suatu program, maka semakin bagus algoritma yang digunakan.