# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang Masalah

Dewasa ini perkembangan teknologi berkembang cukup pesat. Salah satu yang termasuk ke dalam perkembangan teknologi adalah *database*. *Database* merupakan hal yang sangat penting karena dapat menyimpan data yang sangat banyak secara digital. Dengan adanya *database* maka akan mempermudah dalam proses pencarian data. Perusahaan atau organisasi pasti memiliki data yang banyak sehingga pengunaan database pada perusahaan atau organisasi sangat diperlukan.

Data-data yang terdapat pada perusahaan atau instansi terkadang harus dipublikasikan ke pihak ketiga untuk suatu keperluan. Di antara data-data tersebut, terdapat data yang bersifat sensitif sehingga harus dijaga privasinya. Data sensitif ini tidak bisa dipublikasikan ke sembarang pihak karena biasanya berupa rahasia. Sebagai contoh sebuah rumah sakit yang menyimpan data diri pasien. Terkadang hal seperti ini harus dilakukan untuk mengedukasi masyarakat. Sebagai contoh dapat dilihat pada Tabel 1.1:

**Tabel 1.1 Data Diri Pasien di Rumah Sakit**

| Rec.No | Name | Age | Gender | Zip Code | Disease | Cost |
|---|---|---|---|---|---|---|
| 1 | George | 35 | M | 302023 | Flu | 150.000 |
| 2 | Barbara | 31 | F | 302025 | Stomach Cancer | 10.00.000 |
| 3 | Charles | 29 | M | 302020 | Bronchitis | 450.000 |
| 4 | Esra | 33 | F | 302022 | Pneumonia | 700.000 |
| 5 | Febi | 24 | M | 302018 | Stomach Cancer | 20.000.000 |
| 6 | Mike | 30 | M | 302020 | Flu | 300.000 |
| 7 | Peter | 27 | M | 302020 | Pneumonia | 750.000 |
| 8 | Polat | 25 | M | 302018 | Stomach Cancer | 25.000.000 |
| 9 | Jessica | 21 | F | 302018 | Stomach Cancer | 30.000.000 |
| 10 | Jack | 26 | M | 302019 | Gastiris | 5.000.000 |

Sumber: K-anonymity: a model for protecting privacy hal 9 (Dengan Penyesuaian)

Sementara itu, di masyarakat pun terdapat informasi pribadi dari pasien itu sendiri seperti yang ditampilkan pada Tabel 1.2:

**Tabel 1.2 Data Diri Pasien di Masyarakat**

| Rec.No | Name | Age | Gender | Zip code |
|---|---|---|---|---|
| 1 | George | 35 | M | 302023 |
| 2 | Barbara | 31 | F | 302025 |
| 3 | Charles | 29 | M | 302020 |
| 4 | Esra | 33 | F | 302022 |
| 5 | Febi | 24 | M | 302018 |
| 6 | Mike | 30 | M | 302020 |
| 7 | Peter | 27 | M | 302020 |
| 8 | Polat | 25 | M | 302018 |
| 9 | Jessica | 21 | F | 302018 |
| 10 | Jack | 26 | M | 302019 |

Sumber: K-anonymity: a model for protecting privacy hal 9 (Dengan Penyesuaian)

Dengan melihat Tabel 1.1 dan Tabel 1.2 jika dicocokan maka akan didapatkan hubungan antar-tabel, misalnya ketika kita melihat nama Barbara pada Tabel 1.2, lalu mencari nama Barbara di Tabel 1.1 maka akan diketahui bahwa Barbara ternyata terkena penyakit kanker perut dengan biaya pengobata sebesar Rp. 10.000.000.

Padahal penyakit yang diderita dan biaya pengobatan merupakan data yang bersifat sensitif karena tidak bisa di beritahukan ke sembarang orang. Oleh karena itu, data sensitif ini perlu disembunyikan. Hal ini disebut juga dengan *Privacy Preserving Data Mining.* Ada dua cara yang dapat dilakukan untuk menjaga privasi data, yaitu:

1. *Pertubative*

*Pertubative* merupakan penyembunyian data dengan cara mengubah data. Contohnya seperti Kriptografi dan memberikan *noise*.

2. *Non-Pertubative*

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

Pada penelitian kali ini, penulis akan menggunakan cara kedua, yaitu *Non-Pertubative*. Hal ini dikarenakan jika menggunakan cara pertama, misalnya Kriptografi maka semua data akan diacak dan diubah sehingga tidak memiliki makna lagi karena sudah tidak bisa dimengerti, selain itu nilai dari *information loss* nya juga akan semakin besar. Dengan demikian privasi data dengan menggunakan Kriptografi menjadi kurang efisien. Maka penelitian ini akan lebih difokuskan pada penganoniman data. Dengan adanya penganoniman data ini, maka akan didapatkan tabel seperti Tabel 1.3:

**Tabel 1.3. Data Diri Pasien yang Telah di Generalisasi**

| Gr. No | Age | Gender | Zip Code | Disease | Cost |
|---|---|---|---|---|---|
| 1 | 21 – 25 | Person | 302018 | Stomach Cancer | 20.000.000-30.000.000 |
| | 21 – 25 | Person | 302018 | Stomach Cancer | 20.000.000-30.000.000 |
| | 21 – 25 | Person | 302018 | Stomach Cancer | 20.000.000-30.000.000 |
| 2 | 26 – 30 | Person | 302019 – 302020 | Gastritis | 300.000-5.000.000 |
| | 26 – 30 | Person | 302019 – 302020 | Flu | 300.000-5.000.000 |
| | 26 – 30 | Person | 302019 – 302020 | Pneumonia | 300.000-5.000.000 |

| | | | | | |
|---|---|---|---|---|---|
| | 26 – 30 | Person | 302019 – 302020 | Btonchitis | 300.000-5.000.000 |
| 3 | 31 – 35 | Person | 302022 - 302025 | Flu | 150.000-10.000.000 |
| | 31 – 35 | Person | 302022 - 302025 | Stomach Cncer | 150.000-10.000.000 |
| | 31 – 35 | Person | 302022 - 302025 | Pneumonia | 150.000-10.000.000 |

Sumber: K-anonymity: a model for protecting privacy hal 9 (Dengan Penyesuaian)

Dapat dilihat pada Tabel 1.3, ketika sudah dilakukan penganoniman data maka kita sudah tidak bisa mencocokan tabel yang ada sehingga informasi yang diinginkan tidak didapatkan.

Berdasarkan penjabaran diatas, maka perlu ditentukan suatu model untuk bisa mengimplementasikan *Privacy Preserving Data Mining* tersebut, di mana dalam penelitian ini yang digunakan adalah Model *K-Anonymity*. Menurut Kabir et al. (2011: 52), *K-Anonimyty* dikatakan respek terhadap atribut *quasi identifier* jika terdapat setidaknya transaksi k dalam database yang memiliki nilai yang sama sesuai dengan atribut *quasi identifier*. Maksudnya di sini adalah k merupakan jumlah minimal *record* dalam satu kelompok. Akan tetapi, untuk membedakan masalah yang terdapat pada tulisan ini, penulis menggunakan penganoniman data dengan menggunakan dua atribut sensitif dikarenakan masih sangat jarangnya penggunaan atribut dua sensitif itu sendiri. Selain itu, masih adanya kekurangan pada penggunaan satu atribut sensitif, misalnya *homogenity attack.* Dalam implementasi, selain menentukan model, algoritma yang ingin digunakan juga harus ditentukan, di mana pada penelitian kali ini yang digunakan adalah Algoritma *Systematic Clusteing* dan *Gready K-Member*.

Pada penelitian yang telah lebih dulu dilakukan oleh Md. Enamul Kabir, Hua Wang, dan Elisa Berto pada jurnalnya yang berjudul *Efficient Systematic Clustering*

*Method for K-Anonymization*, di mana mereka membandingkan algoritma *Systematic Clustering* dengan *Greedy K-Member* tapi masih dengan menggunkan satu atribut sensitif. Hasil yang didapat adalah *Systematic Clustering* lebih baik daripada *Greedy K-Member* baik dari segi *information loss* nya maupun *running time* nya. Pada penelitian kali ini, penulis akan mencoba menggunakan dua atribut sensitif. Penulis menggunakan dua atribut sensitif dikarenakan jika menggunakan satu atribut sensitif, terkadang masih ada informasi yang bisa di ketahui. Oleh karena itu penulis membuat penelitian Perbandingan Kinerja Algoritma *Systematic Clustering* dan *Greedy K-Member* pada Model *K-Anonymity* yang Menggunakan Dua Atribut Sensitif.

### 1.2. Identifikasi Masalah

Berdasarkan latar belakang yang telah di paparkan, masalah yang dapat diidentifikasikan adalah sebagai berikut:

1. Adanya data-data yang bersifat *sensitive* sehingga tidak boleh di pubilkasikan ke khalayak ramai.
2. Tidak efektif jika menyembunyikan data dengan menggunakan Teknik Kriptografi sehingga perlu ada cara lain.
3. Masih terdapat kekurangan jika melakukan penganoniman data dengan satu atribut sensitive.

### 1.3. Pembatasan Masalah

Melihat luasnya ruang lingkup pada permasalahan yang ada, maka sangat penting untuk menetapkan pembatasan masalah. Maka penelitian ini dibatasi pada:

1. *Privacy* data yang dilakukan hanya menggunakan Model *K-Anonymity*.
2. Menggunakan 2 atribut sensitive pada data.

3. Algoritma yang akan digunakan adalah *Systematic Clustering* dan *Gready K-Member.*

### 1.4. Perumusan Masalah

Berdasarkan latar belakang, identifikasi, dan pembatasan masalah, maka perumusan masalahnya adalah:

**Bagaimana perbandingan kinerja Algoritma *Systematic Clustering* dan *Greedy K-Member* pada Model K-Anonymity yang menggunakan dua Atribut Sensitif?**

### 1.5. Tujuan Umum Penelitian

Tujuan umum dari penelitian ini adalah membandingkan kinerja Algoritma *Systematic Clustering* dan *Greedy K-Member* pada Model K-Anonymity yang menggunakan dua Atribut Sensitif.

### 1.6. Manfaat Penelitian

Dengan adanya penelitian ini, diharapkan data-data yang sensitif bisa disembunyikan dengan baik sehingga privasi dapat terjaga. Penelitian ini dibuat dengan Model *K-Anonymity*, dengan demikian data yang ada tidak perlu diacak seperti jika menggunakan Kriptografi, tapi data yang ada hanya di sembunyikan dengan menggunakan *Local Recording* atau *Global Recording*, sehingga lebih memudahkan dalam pencarian data. Selain itu, penghitungan *Loss Matrix* juga akan semkin mudah.