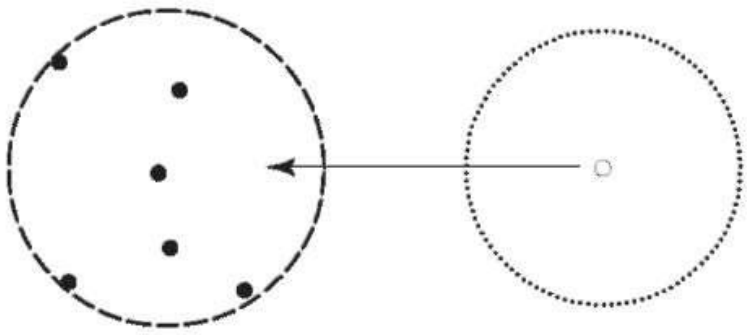
BAB II

TINJAUAN PUSTAKA

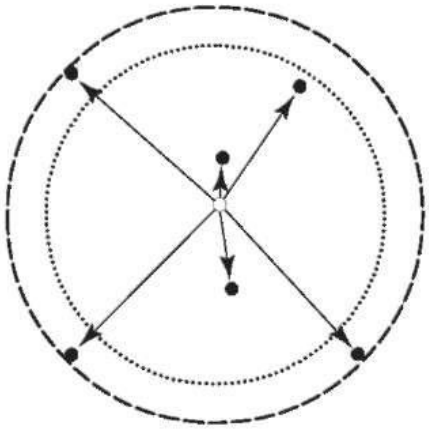
2.1 Penelitian Sebelumnya

2.1.1 *Predicting Serial Killers' Home Base Using A Decision Support System* [3]

Penelitian ini dilakukan untuk mengoptimisasi pencaharian pelaku Serial Killer menggunakan program pendukung keputusan yang disebut Dragnet. Dengan algoritme yang dipakai adalah *Circle Hypothesis*/*Theory*. Algoritme dari *Circle Hypothesis* sendiri memiliki dua model yakni *Commuter Model* (Gambar 2.1) dan *Marauder Model* (Gambar 2.2). Kedua model ini lah yang kemudian akan digunakan untuk memprediksi keberadaan Serial Killer agar pencaharian lebih efisien dan efektif. Algoritme ini terbukti cukup efektif untuk menangkap pembunuh berantai, Canter menyebutkan bahwa 87% dari pelaku pemerkosaan tinggal dalam lingkaran prediksi yang berada di daerah Inggris Selatan sementara tindakan kejahatan yang dilakukan oleh pelaku kriminal 91% berada dalam lingkaran prediksi. Namun, algoritme ini hanya memprediksi keberadaan dari pelaku kriminal berdasarkan *criminal consistency* dengan tanpa memerhatikan struktur permukaan bumi pada area yang didefinisikan. Algoritme ini pun mudah diterapkan dengan hanya menentukan kasus pertama yang muncul sebagai titik aman bagi pelaku kriminal dengan jarak rata-rata 19 km^2 dari titik pusat lingkaran.



Gambar 2.1. *Circle Hypothesis: Commuter Model* [3]



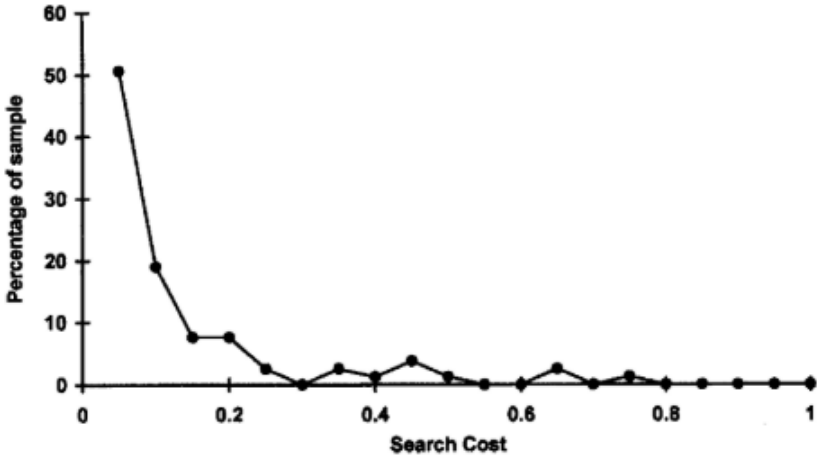
Gambar 2.2. *Circle Hypothesis: Marauder Model* [3]

Setelah titik kasus pertama ditentukan, ambil dua titik serangan terjauh yang dilakukan oleh pelaku kriminal. Gambar lingkaran berdasarkan kedua titik terjauh dengan memertimbangkan titik aman dari kasus pertama. Hasil gambar tersebut kemudian dimasukan ke dalam dua model yang tersedia. Esensinya, model ini digunakan untuk kasus yang berbeda seperti pada kasus *serial killer*. Model yang lebih cocok digunakan adalah *Commuter Model* karena pelaku kriminal cenderung membunuh target ditempat daripada membunuh target di rumah pelaku. Kasus pembunuhan diprioritaskan pada tempat pelaku kriminal membuang mayat korban, bukan tempat sebenarnya pelaku melakukan pembunuhan. Ilustrasi untuk model ini bisa dilihat seperti Gambar 2.1. Sedangkan *Marauder Model* lebih cocok untuk pelaku *serial rapist* yang cenderung melakukan penculikan korban pada titik tertentu dan diasumsikan pelaku selalu kembali kerumahnya setelah menculik satu target untuk kemudian mengeksekusi korban, hal ini dikarenakan pada kasus pemerkosaan pelaku kriminal merasa lebih aman untuk melakukan hal ini di lingkungan yang *familiar* baginya seperti rumah sendiri.

Ruang lingkup aktivitas normal pelaku pembunuh berantai bisa saja melebihi dari lingkaran prediksi, namun bukan berarti lingkungan tersebut cukup *familiar* dan membuat pelaku kriminal merasa “aman” untuk melakukan suatu aktivitas kriminal. Dengan menggunakan sampel dari 79 kasus *serial killer*, mereka memiliki jumlah penyerangan dari 2 sampai 24 (mean, 8; standar

deviasi, 4.53) dan berisikan jarak dari 0 sampai 845km (mean, 46.39; standar deviasi, 85.71km). Hal ini mengeliminasi efek bias penggunaan seri ganda dari peta geografi yang sama. Latar belakang penelitian ini ada pada penentuan titik masukan yang validitasnya belum dapat dipastikan, ditambah probabilitas berpindahnya korban oleh faktor eksternal. Salah satu kasus yang masih sulit diselesaikan dengan cara ini apabila jika pelaku kriminal mengubur korban dibawah tanah agar korban tidak ditemukan dari awal.

Hasil penelitian ini ialah untuk mengukur tingkat keefektivitasan dan keefisienan waktu dan biaya yang diukur menggunakan fungsi mentah QRange, memperlihatkan bahwa terdapat frekuensi tinggi dari pencaharian dengan biaya rendah diikuti angka kecil dengan biaya pencaharian yang lebih tinggi. Kesimpulannya rata-rata untuk melakukan sebuah pencaharian adalah 11% dari potensi area yang diprediksi. Untuk 51% (yang masuk kategori Commuter Model) pelaku penyerangan, kurang dari 5% area yang perlu dicari. Sementara untuk 87% (yang masuk kategori *Marauder* Model) area yang perlu dicari kurang dari 25%. Pada Gambar 2.3, terlihat bahwa persentase yang besar berbanding lurus dengan biaya pencarian yang kecil.



Gambar 2.3. Grafik Persentasi Sampel dari *QRange* [3]

2.1.2 *Predicting The Home Location of Serial Offenders: A Preliminary Comparison of The Accuracy of Human Judges with A Geographic Profiling System* [8]

Penelitian ini dilakukan untuk mengetes akurasi prediksi yang dilakukan manusia dan sistem pendukung keputusan Dragnet dalam menentukan *base* dari *serial offender*. Sistem yang digunakan adalah sistem yang telah berevolusi dari *Circle Theory* menjadi *Geographic Profiling* System. Algoritme yang akan dipelajari terlebih dahulu oleh sisi manusia adalah algoritme *Circle Theory* dan *Distance Decay*. Berdasarkan penelitian Levine, dkk. yang dikutip dalam jurnal ini, mereka membandingkan beberapa model prediksi untuk mengetahui metode yang paling cocok dengan data pembunuhan. Hasilnya, setiap model memiliki akurasi dan kegunaan yang sama bergantung pada utilitas penggunaan dan kecocokan terhadap kasusnya. Hal tersebut mendasari penelitian ini untuk membuktikan bahwa jika manusia dengan pengalaman yang minim di bidang ini diberikan perintah yang jelas, keakuratan prediksi manusia tidak akan jauh berbeda dengan prediksi yang dibuat oleh sistem.

Objek penelitian yang digunakan sebagai sampel dibagi menjadi dua grup yaitu grup kontrol yang berisikan 21 mahasiswa (empat belum lulus, 17 sudah lulus), tanpa pengetahuan samasekali mengenai *geographic profiling*, dari berbagai jurusan di suatu universitas di **Liverpool**. Mahasiswa terdiri dari sepuluh laki-laki dan 11 perempuan dengan asal negara yang berbeda. Umurnya berkisar dari 19 tahun hingga 40 tahun dengan rata-rata 25.90 tahun dan standar deviasi 5.23. Grup lainnya adalah grup eksperimental yang berisikan 21 mahasiswa (delapan belum lulus, 13 sudah lulus), tanpa pengetahuan samasekali mengenai *geographic profiling*¸ dari berbagai jurusan. Terdiri dari 13 laki-laki dan delapan perempuan dari berbagai negara dengan kisaran umur 21 hingga 61 tahun dan rata-rata umur 27.67 tahun (standar deviasi, 8.75).

Prosedur yang dilakukan pada tahap pertama, pada grup kontrol diberikan sebuah materi dan contoh kasus kemudian menebak hasil pada 10

studi kasus selanjutnya di tahap kedua, sedangkan pada grup eksperimental diberikan informasi yang esensial dengan sistem pada tahap kedua.

Spatial display	Control group		Experimental group		Dragnet
	Phase 1 ($n=21$)	Phase 2 ($n=21$)	Phase 1 ($n=21$)	Phase 2 ($n=21$)	
1	23.86 (20.17)*	26.10 (25.86)	21.30 (20.88)	15.76 (8.87)	19.00
2	37.33 (21.72)	35.07 (21.07)	35.76 (20.75)	26.00 (7.27)**	21.00
3	43.24 (17.99)	41.38 (17.55)	39.38 (17.89)	39.67 (13.52)	51.00
4	23.62 (17.00)	22.81 (16.19)	25.38 (20.38)	13.85 (7.07)**	20.00
5	37.57 (28.50)	35.05 (26.54)	33.48 (27.67)	19.24 (11.60)**	14.00
6	27.52 (22.12)	25.67 (15.70)	28.29 (19.75)	20.43 (6.11)	23.00
7	29.67 (25.88)	27.57 (25.89)	30.43 (32.64)	13.24 (14.88)**	2.000
8	32.05 (25.56)	33.00 (24.23)	33.43 (25.30)	19.95 (8.39)**	17.00
9	46.05 (12.75)	42.67 (9.23)	41.05 (17.15)	42.67 (8.77)	49.00
10	52.10 (14.31)	52.10 (12.32)	50.86 (17.61)	45.62 (7.31)	37.00
Mean	35.30	34.22	33.94	25.65	25.30

Gambar 2.4. Jarak Rata-Rata Galat dan Standar Deviasi dari 2 Grup dalam Dua Tahap dan Galat dari Dragnet [8]

Merujuk pada Gambar 2.4, nomor di dalam kurung sebagai standar deviasi dan label "**" yang menandakan perbedaan signifikan antar tahap pada satu kasus sebaris dan satu grup. Dapat disimpulkan, grup eksperimental memiliki jarak signifikan dari tahap pertama dan tahap kedua saat diberitahu informasi yang esensial. Sementara itu grup kontrol tidak menghasilkan perubahan yang signifikan antar tahap tanpa informasi tersebut. Dan pada akhirnya galat yang dimiliki oleh sistem tidak lebih besar daripada grup manusia, tapi dibuktikan bahwa prediksi manusia tidak jauh beda dengan sistem jika diberikan perintah yang jelas bahkan tanpa pelatihan khusus maupun pengalaman yang banyak dalam bidang ini.

2.1.3 *Extensions to The K-Means Algorithm for Clustering Large Data Sets with Categorical Values* [9]

Keterkaitan penelitian Zhexue Huang dengan penelitian ini terdapat pada potensi *heatmap* yang dihasilkan *geographic profiling* dengan pengelompokan kasus pembunuh berantai berdasarkan pola dan kemiripan antar kasus. Hal ini diperkuat oleh penelitian Brent Snook, dkk yang menyebutkan bahwa teknik yang digunakan pada *geographic profiling* dapat digunakan manusia dengan atau tanpa

latar belakang khusus pada bidang terkait. Dengan potensi tersebut peneliti melihat peluang untuk menggunakan algoritme *clustering* seperti *K-Prototype* digunakan karena algoritme ini dapat menangani data yang ada pada dunia nyata (*mixed attributes*) sehingga cocok digunakan pada data kasus pembunuhan di Amerika Serikat yang memiliki campuran data berjenis numerik dan *categorical* seperti yang dikatakan pada penelitian Zhexue Huang.

Menggunakan koefisien similaritas *Gower* pada penelitian Gower (1971), pengukuran ketidaksamaan lain seperti yang tercantum pada penelitian Gowda dan Diday (1991), dan *hierarchical clustering* dapat mengatasi tipe data numerik dan *categorical* menurut Anderberg (1973) dan Jain (1988) yang diacu dalam jurnal Zhexue Huang. Biaya komputasi membuatnya tidak bagus untuk dataset yang besar. Sementara menggunakaan *K-Means* dengan mengubah tipe data *categorical* ke numerik tidak selalu memproduksi nilai yang berarti kecuali data *categorical* tersebut memiliki urutan alias data *ordinal*. Penelitian Ralambondrainy (1995) yang dikutip pada penelitian ini pada dasarnya menggunakan *One-Hot Encoding* untuk input *K-Means*. Namun cara ini hanya akan menambah beban komputasi dan ruang lingkup algoritme, selain itu angka 0 dan 1 tidak mendefinisikan karakter suatu *cluster*. Berdasarkan masalah tersebut, dicetuskan algoritme *K-Prototype* yang merupakan integrasi antara algoritme *K-Means* dan *K-Modes*. Keseluruhan prosesnya sama dengan *K-Means* hanya saja data *categorical* dihitung menggunakan *K-Modes*, dengan ini model dapat mempertahankan efisiensi dari *K-Means Clustering*.

2.1.4 *Adopting the Bottom-up Approach and Cluster Analysis on North American and European Male Serial Killers: A Follow-up Study*[10]

Jurnal Sandie Taylor, dkk meneliti tentang klasifikasi *serial killer* yang tergorganisir dan yang tidak terorganisir menggunakan pendekatan *Bottom-up* dan analisis *cluster* pada *serial killer* di Amerika Selatan dan Eropa. Karakteristik lokasi kejahatan yang memiliki signifikansi personal yang memberikan gambaran terhadap pelaku meliputi jenis kelamin korban, usia, dan penampilan fisik (contoh: model rambut). Pada kasus kejahatan Ted Bundy, pelaku mengincar

korban berjenis kelamin wanita yang memiliki model rambut mirip dengan mantan kekasihnya. Tipe lain dari dari *victim specific data* pada lokasi kejahatan termasuk luka tusukan, penempatan mayat setelah penyerangan serta pertanda pemerkosaan. Diluar kekhawatiran mengenai reliabilitas dan akurasi dari pendekatan yang dilakukan FBI selama ini, mengelompokan *serial killer* berdasarkan pelaku yang terorganisir atau tidak masih umum diaplikasikan terutama di Amerika Selatan.

Berdasarkan 52 kasus kejahatan dari sumber sekunder, pendekatan *Bottom-up* diaplikasikan dengan membagi kriteria dengan ketentuan 1 = ada pada lokasi kejahatan dan 0 = absen dari lokasi kejatahan. Menghasilkan interpretasi yang ditampilkan pada Tabel 2.1. Analisis *cluster* yang digunakan adalah *aggloromerative hierarchical clustering* dengan pembagian *cluster* 1 merupakan kriteria lokasi kejahatan dengan nilai = 0 dan *cluster* 2 dengan nilai = 1. *Cluster* tersebut menghasilkan pengelompokan kasus terorganisir dan tidak terorganisi yang hampir terdistribusi sama rata, frekuensi kemunculan setidaknya enam kasus terorganisir melampaui kasus tidak terorganisir dengan pengecualian *vurnerable victim* pada 48%. Pada kedua *serial killer* baik di Amerika Selatan dan Eropa, kebanyakan *serial killer* memiliki *modus operandi* yang terorganisir, dalam jumlah terbatas, untuk memastikan keberhasilan eksekusi pembunuhan.

Tabel 2.1. Kriteria Lokasi Kejahatan Berdasarkan *Bottom-up*

Crime Scene Criteria	0 (absen) atau 1 (ada)
restraints	1
victim known	0
controlled scene	1
weapon planned	1
act focused	0

2.1.5 *Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms*[10]

Penelitian yang dilakukan pada *paper* ini adalah membandingkan penggunaaan tiga algoritma yaitu *K-Means, K-Modes,* dan *K-Prototype* dengan kombinasi *dataset* yang berbeda. Metode *K-Means* diuji dengan *dataset* numerik "Iris" terdiri dari empat atribut 155 individual, dan "Kolesterol" terdiri dari dua atribut 250 individual, *K-Modes* diuji dengan *dataset categorical* "Kontak Lensa" terdiri dari lima atribut 24 individual, dan "Pasca Operasi" terdiri dari tujuh atribut 190 individual, sementara *K-Prototype* diuji dengan *dataset* campuran "Informasi Darah" terdiri dari tiga atribut numerik, dua atribut *categorical*, 200 individual, dan "Cuaca" terdiri dari dua atribut *categorical*, dua atribut numerik, 350 individual. Dari skema pasangan antar *dataset* dan algoritme, algoritma *K-Means* dan *K-Modes* lebih efektif jika diterapkan pada *dataset* yang memang sudah didesain untuk algoritma tersebut, sementara *K-Prototype* lebih efektif digunakan pada data campuran dibandingkan data dengan atribut hanya numerik atau hanya *categorical*.

Tabel 2.2. Tabel Penelitian Sebelumnya (*State of The Art*)

No.	**Nama Penulis, Algoritma**	**Judul Penelitian**	**Tahun**	**Masalah**	**Hasil Penelitian**	**Perbedaan dengan Penelitian Sebelumnya**
1.	David Canter, dkk; Algoritme: *Circle Theory* [3].	Predicting Serial Killers' Home Base Using A Decision Support System	2000	Biaya dalam pencaharian pelaku kriminal tinggi.	Aplikasi **Dragnet** yang mengaplikasikan *geographic profiling* ditambah teori lingkaran dapat mengoptimalkan biaya pencaharian pelaku kriminal.	Penelitian ini hanya mengoptimasi dengan membandingkan akurasi pengelompokan data tanpa melakukan implementasi sebenarnya dari *Geographic Profiling*.

No.	Nama Penulis, Algoritme	Judul Penelitian	Tahun	Masalah	Hasil Penelitian	Perbedaan dengan Penelitian Sebelumnya
2.	Brent Snook, dkk; Algoritme: *Circle Theory* dan Distance Decay [8].	Predicting The Home Location of Serial Killers: A Preliminary Comparison of The Accuracy of Human Judges with A *Geographic Profiling* System	2002	Belum ada yang membandingkan akurasi dari sistem Dragnet dengan akurasi prediksi manusia.	Studi untuk melakukan teknik heuristik seperti *Circle Theory* tidak memerlukan pengalaman khusus dalam bidang investigasi kriminal.	Penulis menggunakan acuan ini untuk membandingkan hasil studi lapangan dengan hasil optimasi *K-Means*.
3.	Zhexue Huang; Algoritme: *K-Modes*, *K-Means*, dan *K-Prototype*	Extensions to The K-Means Algorithm for Clustering Large Data Sets with Categorical Values	1998	Algoritme *K-Means* hanya menerima data numerik.	Menghasilkan sebuah algoritme yang mengombinasikan *K-Means* dan *K-Modes*.	Penulis menggunakan *K-Prototype* sebagai metode utama dalam penelitian.
4.	Sandie Taylor, dkk; *Bottom-up* dan *Agglomerative Hierarchical Clustering*[10]	Adopting the Bottom-up Approach and Cluster Analysis on North American and European Male Serial Killers: A Follow-up Study	2017	Mengelompokan pelaku terorganisir dan tidak terorganisir dengan *data-driven analysis.*	Menghasilkan *dendogram* pembagian *serial killer* terorganisir dan yang tidak berdasarkan kriteria yang dihasilkan *bottom-up*.	Penulis menggunakan *partition-based clustering*, kasus *serial killer* belum dikelompokan sementara pada penelitian Sandie Taylor, data yang digunakan seluruhnya adalah data *serial killer*.
5.	R. Madhuri, dkk; *K-Means*, *K-Modes*, dan *K-Prototype*[10].	Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms	2014	Memastikan efisiensi masing-masing algoritme dalam mengelompokan berbagai *dataset* dengan atribut tertentu.	Interpretasi bahwa masing-masing algoritma memiliki efektivitas yang lebih tinggi pada atribut dengan satu jenis (numerik atau *categorical*) atribut.	Analisis pada penelitian R. Madhuri hanya diaplikasikan menggunakan *dataset* publik yang umum.

2.2 Dasar Teori

2.2.1 *Data mining*

Metode yang digunakan untuk mengolah suatu data yang besar untuk memperoleh informasi yang baru serta mengenali polanya adalah *data mining*. Dengan mengetahui polanya dan menyelaraskan hubungan antar data, solusi dapat ditemukan melalui analisis. *Data mining* adalah salah satu bidang yang mendasari beberapa cabang ilmu seperti *maching learning*, statisktik, dan sistem *database*. Tahap yang dilakukan dalam melakukan *data mining* bisa disebut juga *KDD* (*Knowledge Discovery from Data*) yang berisikan tahap-tahap keseluruhan dalam melakukan *data mining* seperti memilih target data, *preprocessing* suatu data, mentransformasinya jika perlu, melakukan *data mining* untuk ekstraksi pola dan hubungan antar data, dan menginterpretasikan serta menilai data berdasarkan struktur yang ditemukan [11]. Adapun pengertian *Data mining* menurut para ahli:

1. D. Hand dan kawan-kawan [11] menyatakan bahwa *Data mining* adalah analisis dari dataset observasi (biasanya besar) untuk menemukan suatu hubungan dan menyimpulkannya ke data ke dalam bentuk seperti novel yang keduanya dapat di mengerti dan berguna untuk pemilik data. Hubungan dan kesimpulan diambil *data training* yang sering disebut model atau pola. Contohnya persamaan linear, aturan, klaster, graf, struktur pohon, pola berulang di suatu waktu.
2. Ian H. Witten dan Frank Eibe [12] mendefinisikan *Data mining* sebagai proses untuk menemukan sebuah pola dalam data. Prosesnya bisa otomatis atau biasanya semiotomatis. Pola yang ditemukan haruslah berarti dan mengarah ke sesuatu yang menguntungkan, biasanya keuntungan ekonomi. Data tersebut selalu ada dalam kuantitas substansial.
3. Jiawei Han [13] juga menjelaskan *Data mining* dengan istilah populer lainnya yaitu *KDD* merupakan sekuen iterative yang terdiri dari:

a. *Data Cleaning*, yaitu membersihkan *noise* dan data yang tidak konsisten.
b. *Data Integration*¸ yaitu menentukan banyak sumber data yang kemungkinan bisa digabungkan.

Kedua proses di atas masuk ke dalam proses *preprocessing* dari *Data mining*.

c. *Data Selection*, yaitu data yang akan dianalisis diambil dari *database*.
d. *Data Transformation*, yaitu data di transformasi dan di konsolidasi ke dalam bentuk yang lebih pantas untung *mining* dengan cara melakukan operasi penyimpulan atau agregasi.
e. *Data mining*, proses yang esensial dalam *data mining* yaitu menggunakan suatu algoritme untuk mengekstrak pola data.
f. *Pattern Evaluation*, untuk mengidentifikasi pola yang menarik yang merepresentasikan pengetahuan berdasarkan *interestingness measures*.
g. *Knowledge Presentation*, yaitu teknik visualisasi dan representasi pengetahuan digunakan memberitahu data yang telah digali kepada pengguna.

2.2.2 *Clustering*

Clustering merupakan teknik yang bersifat *unsupervised*, tujuan dari teknik *clustering* yaitu agar dapat menggabungkan objek-objek tersebut ke dalam suatu kelompok dengan memiliki kesamaan antar objek di dalamnya atau memiliki hubungan satu sama lain. Menurut Johnson dan Wichern (2007) diacu dalam [14] *cluster analysis* atau analisis kelompok merupakan sebuah metode menganalisis untuk mengelompokan suatu objek menjadi beberapa kelompok sehingga akan diperoleh sebuah kelompok dimana objek-objek dalam satu kelompok tersebut memiliki banyak persamaan sedangkan pada anggota kelompok yang lain memiliki banyak perbedaan.

2.2.2.1 *K-Means Clustering*

Algoritme *K-Means Clustering* merupakan salah satu algoritme pengelompokan yang populer dan cukup sederhana. *Clustering* digunakan saat tidak ada kelas yang dispesifikasikan dalam data [12]. Hasil dari *clustering* adalah data yang sudah diberi label sesuai dengan penempatan *cluster*nya.

Metode *K-Means* di sisi lain menggunakan teknik yang disebut *Iterative Distance-based Clustering*. Disebut demikian karena untuk menggunakan teknik ini pertama perlu dispesifikasikan berapa jumlah *cluster* yang ingin dibuat, parameter ini disebut *k*. Lalu poin *k* dipilih secara acak sebagai *centroid cluster*. Seluruh intansi dikumpulkan kepada inti *cluster* terdekat berdasarkan dengan pengukuran jarak *Euclidean Distance*. Kemudian *centroid* atau rata-rata dari setiap *cluster* dihitung, ini lah bagian "means"nya.

Secara bertahap, tahapan *K-Means Clustering* dapat dijabarkan sebagai berikut[15]:

1. Memasukan data yang akan di *cluster*.
2. Menentukan jumlah *cluster*.
3. Menentukan *centroid* secara acak yang berfungsi sebagai pusat di masing-masing *cluster*.
4. Menghitung jarak antara data dengan pusat *cluster* dengan menggunakan persamaan:

$$D(i,j) = \sqrt{(xl_i - xl_j)^2 + \cdots + (xk_i - xk_j)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Keterangan:

$D(i,j)$ = Jarak *Euclidean* data *i* ke *centroid j*;

xk_i = Data ke *i* pada atribut ke-*j*;

xk_j = Titik *centroid* ke-*j* pada atribut *k*;

5. Menghitung kembali *centroid cluster* dengan anggota *cluster* yang baru menggunakan *mean* dari individual numerik.
6. Ulangi proses 3 hingga 5, berhenti jika *centroid* tidak berubah lagi nilainya.

Kelemahan dari Algoritme *K-Means Clustering* adalah nilai akhir atau kualitas dari *cluster* sangat ditentukan oleh *centroid* awal yang ditentukan acak. Salah satu cara konvensional yang dapat dilakukan untuk mengatasi masalah ini adalah dengan mencoba setiap kemungkinan dimulai dari $k = 1$, namun cara ini mengurangi kecepatan proses *clustering*. Keseluruhan proses perhitungan yang dibutuhkan oleh *K-Means Clustering* dapat dikurangi dengan menggunakan informasi perhitungan jarak dari iterasi sebelumnya dan beberapa iterasi pertama [16].

2.2.2.2 *K-Modes Clustering*

K-Modes merupakan pengembangan dari *K-Means* yang pada dasarnya memiliki langkah kerja yang sama dari *K-Means* namun persamaan jarak yang digunakan adalah *Simple Disimilarity Matching*. *K-Means* tidak dapat digunakan pada jenis fitur data jenis *categorical* yang berupa nominal atau ordinal, untuk menyelesaikan permasalahan tersebut perlu digunakan algoritme yang sesuai, salah satunya *K-Modes Clustering*. Modifikasi yang dilakukan *K-Modes* pada *K-Means Clustering* adalah sebagai berikut [17]:

1. Menggunakan sebuah ukuran pencocokan ketidakmiripan sederhana pada fitur data jenis *categorical* yang berupa nominal atau ordinal.
2. Mengganti nilai *Mean Cluster* dengan nilai *Modus* (nilai yang paling sering tampil).
3. Menggunakan metode berbasis frekuensi dalam mencari modus dari sekumpulan nilai yang ada.

Sehingga langkah-langkah dalam *K-Modes* sama dengan *K-Means*, hanya berbeda pada pengukuran jarak yang menggunakan *Dissimilarity Matching*. Tahapan *K-Modes* dapat dijabarkan sebagai berikut:

1. Memasukan data yang akan di *cluster*.
2. Menentukan jumlah *cluster*.
3. Menentukan *centroid* secara acak yang berfungsi sebagai pusat di masing-masing *cluster*.
4. Menghitung jarak antara data dengan pusat *cluster* dengan menggunakan persamaan:

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad \text{.....(2)}$$

Dimana:

$$\delta(x_j, yj) = \begin{cases} 0(x_j = y_j) \\ 1(x_j \neq y_j) \end{cases} \quad \text{.....(3)}$$

Keterangan:

$d_1(X,Y)$ = Jarak *Dissimilarity Matching* data *i* ke *centroid j*;

xk_i = Data ke *i* pada atribut ke-*j*;

xk_j = Titik *centroid* ke-*j* pada atribut *k*;

5. Menghitung kembali *centroid cluster* dengan anggota *cluster* yang baru menggunakan modus kemunculan individual *categorical*.
6. Ulangi proses 3 hingga 5, berhenti jika *centroid* tidak berubah lagi nilainya.

2.2.2.3 *K-Prototype Clustering*

K-Prototype sendiri merupakan gabungan dari *K-Means* dan *K-Modes*. Pada dasarnya data numerik akan dihitung dengan persamaan *Euclidean Distance* sementara data *categorical* akan masuk ke dalam

persamaan *Simple Dissimilarity Matching* kemudian jaraknya dijumlahkan. *K-Prototype* termasuk ke dalam kategori *hard partitioning cluster* [18], hal tersebut cocok dengan kasus dalam penelitian ini. Dengan persamaan [6]:

$$\vartheta(d_i, C_j) = \sum_{t=1}^{m_r}(d_{it}^r - C_{jt}^r)^2 + \gamma_j \sum_{t=1}^{m_c} \delta(d_{it}^c, C_{jt}^c) \text{.....(4)}$$

Yang berarti:

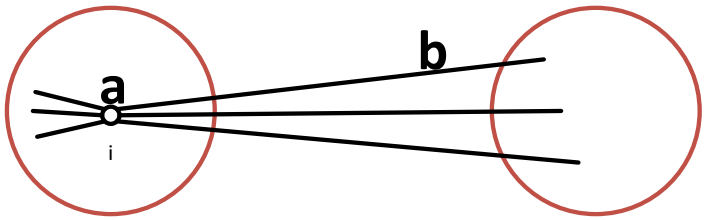
1. γ_j , frekuensi kemunculan individual *categorical*.
2. m_c, variabel *categorical*.
3. m_r, variabel numerik.
4. d_{it}^r, adalah individual numerik.
5. d_{it}^c, adalah individual *categorical* untuk objek data d_i.
6. C_j, merupakan *centroid cluster* untuk *cluster j*.
7. C_{jt}^c, merepresentasikan *most common value* atau mode untuk atribut kategorical *t* dan kelas *j*.
8. C_{jt}^r, merepresentasikan rata-rata dari atribut numerik untuk *cluster j*.

Berdasarkan rumus *K-Prototype*, dapat disimpulkan bahwa sisi kiri (sebelum tanda '+') rumus merupakan perhitungan data numerikal menggunakan *Euclidean Distance*, sedangkan pada sisi kanan (setelah tanda '+') merupakan perhitungan data *categorical* menggunakan *dissimilarity matching*.

2.2.3 *Silhouette Coefficient*

Cluster yang terbentuk pada penelitian ini kemudian akan divalidasi *cluster*nya melalui *Silhouette Coefficient*, cara ini dapat digunakan karena objek yang diciptakan oleh algoritme *K-Prototype Clustering* sama dengan *K-Means*. *Silhouette Coefficient* adalah sebuah metode penafsiran yang digunakan untuk validasi klaster pada objek-objek tertentu [19]. Teknik ini

memberikan sebuah representasi grafis singkat tentang seberapa baik setiap objek yang terletak dalam *cluster*-nya.



Gambar 2.5. *Silhouette Coefficient* jarak *a*(*i*) dan (*b*)

Merujuk pada Gambar 2.5, nilai *silhouette coefficient* dari sebuah objek (misalkan objek *a*(*i*)) berada pada posisi rentang antara nilai -1 sampai dengan 1. Semakin dekat nilai *silhouette* objek *a*(*i*) ke 1, akan berdampak dengan semakin tingginya derajat kepemilikan objek *a*(*i*) di dalam *cluster* tersebut. Untuk mencapai nilai *silhouette coefficient* mendekati satu, jarak *centroid a* terhadap *data point* intra-*cluster* harus kecil, semantara jarak *centroid a* dengan *data point* terdekat di dalam inter-*cluster b* harus besar. Berikut ini merupakan penggambaran objek untuk menentukan *silhouette coeficient* dari sebuah cluster. Angka *silhouette s*(*i*) yang tercipta dari *a*(*i*) dan *b*(*i*) didapatkan menggunakan [20]:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, jika\ a(i) < b(i); \\ 0, jika\ a(i) = b(i); \\ \frac{b(i)}{a(i)} - 1, jika\ a(i) > b(i); \end{cases} \quad \text{.....(5)}$$

Dengan *s*(*i*) antara -1 ≤ *s*(*i*) ≤ 1, dihitung menggunakan rumus:

$$s(i) = \frac{b_i - a_i}{Max(a_i, b_i)} \text{.....(6)}$$

Yang berarti:

b_i = *Data point* ke-*i cluster b* yang terdekat dengan *centroid cluster a*.

a_i = *Centroid i* dari *cluster a*.

2.2.4 *Dimensionality Reduction*

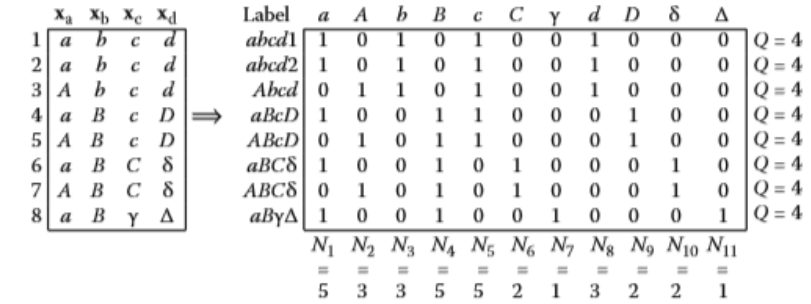
Dalam *dimensionality reduction* dalam menganalisis data dimensi tinggi seringkali mengalami permasalahan dikarenakan tingginya biaya komputasi dan penggunaan memori. *Dimensionality reduction* adalah transformasi data dimensi tinggi menjadi representasi bermakna dari pengurangan dimensi [21]. Idealnya, representasi tereduksi harus memiliki dimensi yang sesuai dengan dimensi intrinsik data. Dengan mengurangi fitur yang tidak relevan akan mempercepat proses komputasi dan dalam skema terbaik dapat memperbaiki akurasi.

2.2.5.1 *Principal Component Analysis*

Principal Component Analysis (*PCA*) merupakan salah satu teknik bersifat *unsupervised*, dan termasuk ke dalam teknik linear dalam ranah *dimensionality reduction*. Menurut H. Hotelling (1933) yang diacu dalam [21] *Principal Components Analysis* (PCA) bekerja dengan membuat representasi data berdimensi rendah yang menggambarkan sebanyak mungkin varians dalam data. *PCA* bertujuan membuat representasi data dengan dimensional rendah yang *meaningful* yang artinya memertahankan atribut dengan *variance* terbesar [22]. Standarisasi data dapat dilakukan sebelum melakukan *PCA* agar data dengan satuan ukur berbeda dapat terdistribusi dengan baik.

2.2.5.2 *Multiple Correspondance Analysis*

Salah satu metode yang digunakan dalam penelitian ini adalah *Multiple Correspondance Analysis* (*MCA*) sebagai sarana dalam pemilihan fitur pada variabel *categorical*. *MCA* merupakan metode dari *analyse des données* yang digunakan untuk menjelaskan, mengeksplor, menyimpulkan, dan memvisualisasi informasi yang terdapat pada data *table* berisi ‘N’ *individual* dijelaskan oleh ‘Q’ variabel *categorical* [23]. Pada *MCA*, data *categorical* yang digunakan akan dikonversi menjadi *binary* (*dummy variable*) melalui *one-hot encoding* (Gambar 2.6).



Gambar 2.6. Konversi Data Mentah ke Bentuk Matriks Indikator (*dummy variable*) [23]

2.2.5.3 Data *Correlation*

Analisis korelasi merupakan sebuah metode dalam ilmu statistik yang biasanya digunakan untuk menentukan suatu besaran yang akan menyatakan bagaimana hubungan satu variabel dengan variabel lain dengan tidak mempersoalkan apakah variabel tersebut akan bergantung pada variabel lain. Dua jenis teknik korelasi yang masih popular dan masih digunakan sampai saat ini, yaitu: Korelasi *Pearson Product Moment* dan Korelasi *Rank Spearman*. Korelasi *person* merupakan korelasi sederhana yang hanya melibatkan satu variabel terikat (*dependent*) dan satu variabel bebas (*independent*) [24].

Menurut Fidaus (2009) yang diacu dalam [24], penemu koefisien korelasi ini ditemukan oleh Karl Pearson pada tahun 1990. Koefisien korelasi merupakan ukuran yang biasa dipakai untuk mengetahui derajat hubungan antar variabel satu dengan variabel lainnya. Korelasi *pearson* ini menghasilkan sebuah koefisien korelasi yang bertujuan untuk mengukur kekuatan hubungan linier diantara dua variabel. Jika hubungan antar dua variabel tersebut tidak linier, maka dapat disimpulkan bahwa koefisien tersebut tidak mencerminkan kekuatan hubungan antar dua variabel yang sedang diuji meskipun kedua variabel ini mempunyai hubungan yang kuat.

Dalam korelasi *pearson* terdapat dua simbol yang digunakan yaitu: *p* dan r, simbol *p* digunakan jika diukur dalam bentuk populasi dan simbol r digunakan jika diukur dalam bentuk sampel. Nilai koefisien korelasi bernilai antara -1 < r < 1 yaitu:

1. Apabila r =-1 bernilai korelasi negatif sempurna, yang berarti taraf signifikan dari pengaruh variabel X terhadap variabel Y sangat kuat.
2. Apabila r = 1 bernilai positif sempurna, yang berarti taraf signifikan dari pengaruh variabel X terhadap variabel Y sangat lemah.
3. Apabila nilai koefisien korelasi menunjukkan angka 0, maka dapat disimpulkan bahwa tidak terdapat hubungan antara dua variabel yang diuji.

2.2.5 *One-Hot Encoding*

Merupakan proses mengubah suatu variabel *categorical* (*nominal*) menjadi sebuah bentuk yang tersedia untuk sebuah algoritme *machine learning* dengan mengubah setiap kategori menjadi sebuah indikator biner sesuai dengan ilustrasi pada Gambar 2.5. *One-hot encoding* digunakan sebelum melakukan *MCA*. *One hot encoding* mengubah variabel *categorical* menjadi sebuah bentuk *biner* agar fitur tersebut dapat diperhitungkan dalam klasifikasi. Masalah klasik yang ada pada analisis data statistik ada pada data *categorical* baik *nominal* maupun *ordinal* yang tidak standar [25], karena tanpa supervisi eksternal oleh ahli yang memiliki pemahaman atas bidang data tersebut akan sulit bagi statistikawan untuk menganalisis data *categorical*.

Penggunaan metode tersebut tidak disarankan untuk *input cluster* seperti *K-Means* karena pengukuran jarak akan memiliki masukan biner (0 dan 1) yang mana membuat hasil *mean* dari *individual* kurang berarti, cara ini juga akan menambah dimensi data yang memiliki banyak *categories* secara drastis. Namun, metode ini bisa digunakan untuk mengonversi data *categorical* yang tidak standar (biasanya dari sumber data).

2.2.6 *Geographic Profiling*

Salah satu metode yang digunakan untuk melacak pembunuh berantai adalah *geographic profiling*, berbeda dengan metode yang biasa digunakan pihak kepolisian atau teknik *criminal profiling* yang digunakan FBI yang berpusat pada keseluruhan data yang bisa diambil dari lokasi kejadian dan mengidentifikasi motif pelaku. *Geographic profiling* menganalisis ciri-ciri dari kejadian dari sisi psikologi, dan menekankan pada apa dan siapa pada kasus terkait. Hal ini membuat *geographic profiling* dapat digunakan pada *dataset* yang lebih besar. Sejak awal tahun studi Canter dan Larkin di tahun 1993 dan studi Canter dan Gregory (1994) menunjukan validitas empiris dalam menentukan area dimana kemungkinan pembunuh berantai tinggal berdasarkan detail dari lokasi kejadian secara utuh disebut dengan *Geographic Profiling* [26]. Disebutkan juga bahwa tingkat efektivitas daripada orang maupun komputer bergantung pada validitas data, batasan efektivitas dari sistem dalam penggunaannya adalah *noise* dan *source error* di dalam data. Satu isu lagi pada *geographic profiling* terkait dengan jumlah minimum kasus sebelum bisa dilakukan *profiling*. Analisis 'Monte Carlo' terkadang mengarah pada klaim bahwa jumlah kasus minimum yang dibutuhkan adalah lima .

2.2.7 *Heatmap*

Heatmap adalah suatu grafis yang dapat memvisualisasikan data berdasarkan intensitas dan biasa tergambar dalam bentuk temperatur. Semakin tinggi intensitas pada suatu data, maka semakin pekat warna dari *heatmap* tersebut. Pada *geographic profiling*, *heatmap* yang dihasilkan dari sistem bisa digunakan bersamaan dengan algoritme *circle theory*. *Cluster heatmap* adalah lantai segi empat dari matriks data dengan pohon *cluster* yang ditambahkan ke dalam marjin [27]. *Cluster heatmap* sering digunakan dalam bidang *biological sciences* dengan warna merah untuk ekspresi tinggi dan warna biru atau hijau untuk ekspresi rendah.

Heatmap mempermutasi baris dan kolom dari matriks untuk memperjelas struktur data. Salah satu metode permutasi matriks yang

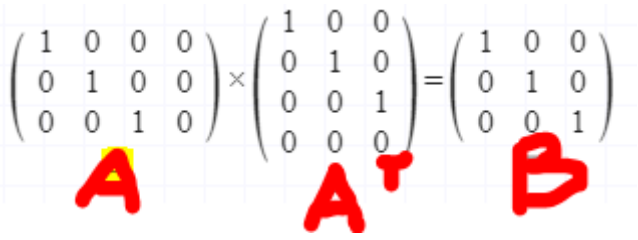
digunakan adalah *seriation*, pertama kali ditemukan oleh seorang Antropologis bernama **Petrie**. Pada sumber [28], *heatmap* dibuat dengan cara *transpose* matriks pada data (numerik dan *non*-numerik) kemudian langsung menggambarkannya ke dalam *heatmap* menggunakan data pada Gambar 2.7. Frekuensi data setiap tahun menggambarkan *density* dari *heatmap* yang dibentuk.

Total Crime Rate

State	*1965*	*1966*	*1967*	*1968*	*1969*	*1970*	*1971*	*1972*	*1973*	*1974*	*1975*
Alaska	2603.6	2785.7	2884.6	3320.9	3880.1	3801.8	4164.5	4478.5	4943.3	5239.8	6196.6
Alabama	1592.5	1758.3	1851	1999	2126.6	2479.5	2498.4	2394.5	2582.3	3000.1	3472.5
Arkansas	1274.2	1382.9	1628.5	1958.5	2188.5	2421.2	2328.2	2352.4	2756.5	3300.7	3540.1
Arizona	3547.8	4135.8	4837.9	4874.4	5224.6	5914.2	5941.5	5933.3	6703.9	8221.7	8341.5
California	4319.4	4549.4	5055.1	5721.1	6099.7	6339.1	6690.1	6413.1	6304.9	6846.8	7204.6
Colorado	2704.5	3009.6	3309.1	3862.6	4498.2	5318.2	5517	5593.6	5495.8	6165.8	6675.5
Connecticut	1834.4	1982.3	2281.2	2890.4	3225.5	3489.4	3646.2	3403.1	3664.4	4407	4957

Gambar 2.7. Sampel Data Yang Digunakan Pada Sumber [28]

Walaupun dijelaskan bahwa data dapat berupa numerik maupun *non*-numerik, data yang digunakan sebagai intensitas adalah numerik. Sementara data yang digunakan penelitian ini adalah campuran dari kedua jenis tersebut. Untuk mengatasi hal ini peneliti akan menggunakan *One-Hot Encoding* untuk menserialisasi data ke dalam dimensi yang lebih besar dalam bentuk biner. Walaupun jumlah dimensi lebih besar, pembuatan *heatmap* tetap dapat dilakukan karena *heatmap* sendiri secara relatif dapat ditampilkan melalui nilai biner [27].



Gambar 2.8 Matriks A sebagai Data, Matriks A^T sebagai Data Tranpose, dan B Hasil *Dot Product*

Setelah *One-Hot*, untuk visualisasi *heatmap* digunakan *dot product* data *categorical* dengan matriks transpose dari data tersebut, menghasilkan matriks

dengan similaritas tinggi = 1, dan similaritas rendah = 0. Detail mengenai *dot product* yang digunakan untuk *2D Heatmap* (ditunjukan pada Gambar 2.8).

2.2.8 Bahasa Pemrograman Python

Bahasa pemrograman *Python* merupakan bahasa yang sering digunakan dalam bidang *data science,* karena bahasanya sudah masuk ke dalam kategori *high-end language* atau bahasa tingkat tinggi yang sudah hampir menyerupai bahasa sehari-hari manusia. Selain itu, bahasa pemrograman *Python* juga merupakan bahasa yang ditujukan untuk tujuan umum. Sehingga selain untuk analisis data, bahasa pemrograman ini dapat digunakan dalam banyak bidang termasuk industri. Salah satu contoh aplikasi terkenal yang dibuat dengan bahasa pemrograman *Python* adalah Instagram, lebih spesifiknya menggunakan *framework* Django. *Python* dapat dideksripsikan dengan berbagai frasa, menurut salah satu *Developer Python* di *website* resmi *Python*, Tim Peters [29] menggambarkan *python* di bait pertama dengan kata:

> "Beautiful is better than ugly.
>
> Explicit is better than implicit.
>
> Simple is better than complex.
>
> Complex is better than complicated."

Python adalah bahasa pemrograman yang terinterpretasi, bahasa pemrograman yang berorientasi objek seperti PERL, yang mendapatkan popularitasnya berdasarkan sintaksnya yang jelas dan mudah dibaca [16]. *Python* juga dapat dikatakan relative mudah untuk dipelajari dan *portable*, yang artinya dapat digunakan di sistem berbasis UNIX, Mac OS, MS-DOS, OS/2, dan banyak versi dari Windows 98.

2.2.9 Bahasa Pemrograman R

Bahasa Pemrograman R adalah salah satu bahasa pemrograman yang biasa digunakan pada bidang statistik dan pada saat ini, *data science*. Bahasa pemrograman R juga *open source* sehingga dapat digunakan secara bebas dan

gratis untuk semua orang. Bahasa pemrograman ini memiliki kelebihan dalam menganalisis statistik karena sudah memiliki *built-in keyword* untuk perhitungan rumus matematika, bahasa ini juga disertai dengan visualisasi data tanpa harus meng*install* ekstensi. R adalah sistem yang diperuntukan untuk analisis statistik dan grafik dan dibuat oleh Ross Ihaka dan Robert Gentleman [30]. R merupakan perangkat lunak sekaligus bahasa pemrograman yang mengambil dialek dari bahasa pemrograman 'S' yang dibuat oleh AT&T Bell Laboratories. Bahasa pemrograman R terlihat kompleks bagi non-spesialis pada awalnya.

2.2.10 Data *Sampling*

Analisis data dapat dilakukan secara menyeluruh, maupun dengan mengambil sampel dari keseluruhan data. Pada kenyataannya bisa saja untuk menganalisis keseluruhan data seperti sensus. Dalam konteks penelitian, mustahil untuk mengumpulkan dan menganalisis semua data yang ada karena adanya batasan waktu, biaya, dan akses. Penulis akan mengambil sampel acak dari *dataset* yang ada sehingga sampel tipe ini masuk ke dalam kategori *probability sampling*.

Sampling frame untuk sampel probabilitas merupakan daftar lengkap dari keseluruhan kasus pada populasi yang mana sampel akan diambil [31]. Jika pertanyaan penelitian menyangkut anggota dari *club* golf lokal, maka *sampling frame*nya adalah daftar lengkap dari keanggotaan dari *club* golf tersebut. Disebutkan juga oleh penulis yang mengutip dari penelitian Edwards *et al*. (2007), bahwa peneliti harus sadar akan masalah yang mungkin terjadi jika menggunakan *database* yang sudah ada yaitu:

1. Database individu sering tidak lengkap.
2. Informasi terkandung mengenai organisasi di dalam *database* kadang tidak akurat.
3. Informasi terkandung dalam *database* akan ketinggalan zaman.

Menyikapi parameter di atas, penulis memastikan *dataset* yang digunakan dapat dipercaya, dan tidak ketinggalan zaman. Berdasarkan sumber data, dipastikan bahwa data yang tercatat adalah resmi dari Federal Bureau of Investigation (FBI). Dan berdasarkan *dataset* tersedia, *record* diperbaharui setiap tahunnya.

Dalam penelitian ini rumus yang akan digunakan adalah formula *Cochran*, formula ini cocok digunakan untuk populasi yang besar (pada penelitian ini total populasi berjumlah 638.455). Rumus untuk menghitung sampel berdasarkan proporsi[32]:

$$n_0 = \frac{Z^2 pq}{e^2} \quad \text{.....(7)}$$

Dimana n_0 adalah ukuran sampel, Z^2 sama dengan nilai *confidence* standar = 95%, *e* adalah tingkat presisi yang diinginkan, *p* adalah proporsi estimasi dari atribut yang terdapat di populasi, dan *q* adalah 1-*p*. Dengan menambah total jumlah populasi, rumus untuk menghitung sampel sebagai berikut:

$$sample\ size = \frac{\frac{Z^2.\ p\ (1-p)}{e^2}}{1+(\frac{Z^2.\ p\ (1-p)}{e^2 N})} \quad \text{.....(8)}$$

Keterangan:

1. *Z*, hasil invers distribusi normal dari rumus alpha dengan *confidence level* standar = 0,95:

$$\alpha = \frac{1-Confidence\ Level}{2} \quad \text{.....(9)}$$

2. *p*, proporsi data dengan nilai *default* = 0,5 jika nilai tersebut tidak diketahui.
3. *e*, parameter *error* dari rumus yang jika semakin besar, jumlah sampel semakin kecil.

4. *N*, total keseluruhan populasi. Perbandingan populasi terhadap jumlah sampel dengan parameter *margin of error* dapat dilihat pada Gambar 2.9.

	Margin of error			
Population	**5%**	**3%**	**2%**	**1%**
50	44	48	49	50
100	79	91	96	99
150	108	132	141	148
200	132	168	185	196
250	151	203	226	244
300	168	234	267	291
400	196	291	343	384
500	217	340	414	475
750	254	440	571	696
1 000	278	516	706	906
2 000	322	696	1091	1655
5 000	357	879	1622	3288
10 000	370	964	1936	4899
100 000	383	1056	2345	8762
1 000 000	384	1066	2395	9513
10 000 000	384	1067	2400	9595

Gambar 2.9. Ukuran Sampel Berbanding Total Populasi [31]