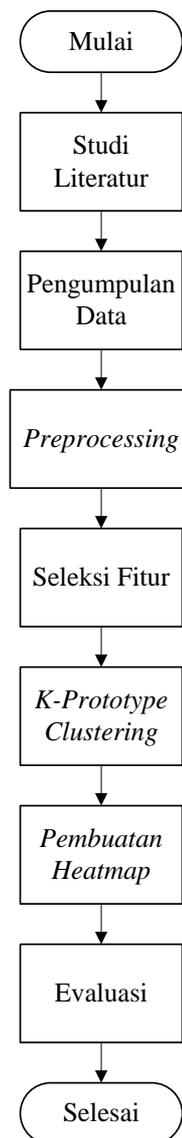


BAB III

METODE PENELITIAN

3.1 Tahap Penelitian

Pada bab ini dijelaskan tahapan-tahapan penelitian yang diimplementasikan dalam penelitian ini. Tahapan tersebut ditunjukkan oleh diagram alir pada Gambar 3.1:



Gambar 3.1. Diagram Alir Penelitian

3.2 Studi Literatur

Peneliti melakukan studi literatur terhadap dokumen penelitian sebelumnya yang berkaitan dengan masalah pembunuh berantai, *Geographic Profiling*, dan masalah yang dapat diselesaikan dengan *Clustering*. Rujukan utama dari penelitian ini adalah penelitian David Canter, dkk dengan detil mengenai *geographic profiling*, *circle theory*, dan *serial killer*; Penelitian Brent Snook, dkk dengan percobaan algoritme *Circle Theory* dan *Distance Decay* dalam mengidentifikasi pembunuh pada manusia tanpa latar belakang khusus; Penelitian Zhexue Huang dengan algoritme *K-Prototype Clustering* pada data *mixed attributes*; Penelitian Sandie Taylor, dkk dengan pendekatan *Bottom-up* dan analisis *cluster* pada data spesifik kasus *serial killer* di Amerika Serikat dan Eropa; dan penelitian R. Madhuri, dkk mengenai perbandingan performa metode *K-Means*, *K-Modes*, dan *K-Prototype* menggunakan data dengan atribut tertentu.

3.3 Pengumpulan Data

Pengumpulan Data dilakukan dengan cara mengambil data dari Murder Accountability Project (<http://www.murderdata.org/p/data-docs.html>) yang dibuat oleh Thomas K. Hargrove serta sebagai *Chairman* di proyek tersebut. *Dataset*-nya terdiri 638.455 baris data dengan 24 atribut berdasarkan kasus dari tahun 1980 hingga tahun 2014. *Dataset* diambil sebanyak mungkin agar *overview* jumlah kasus memberikan hasil yang lebih *universal*. Batas tahun 2014 diambil karena pada *dataset* terbaru atribut “city” dihilangkan, *dataset* akan diperlakukan sebagai “data” tanpa memperhatikan kasus yang sudah selesai atau belum. *Dataset* ini didasari oleh data kasus yang dikelola oleh FBI (Federal Bureau of Investigation). Atribut *dataset* yang digunakan dalam penelitian terdapat pada Tabel 3.1.

Tabel 3.1. Atribut pada *Dataset* Kasus Pembunuhan di Amerika Serikat Tahun 1980 - 2014

| Nama Atribut | Jenis Data |
|------------------------------|--------------------|
| <i>Record_ID</i> | <i>Categorical</i> |
| Nama Atribut | Jenis Data |
| <i>Agency_Code</i> | <i>Categorical</i> |
| <i>Agency_Name</i> | <i>Categorical</i> |
| <i>Agency_Type</i> | <i>Categorical</i> |
| <i>City</i> | <i>Categorical</i> |
| <i>State</i> | <i>Categorical</i> |
| <i>Year</i> | Numerik |
| <i>Month</i> | <i>Categorical</i> |
| <i>Incident</i> | Numerik |
| <i>Crime_Type</i> | <i>Categorical</i> |
| <i>Crime_Solved</i> | <i>Categorical</i> |
| <i>Victim_Sex</i> | <i>Categorical</i> |
| <i>Victim_Age</i> | Numerik |
| <i>Victim_Race</i> | <i>Categorical</i> |
| <i>Victim_Ethnicity</i> | <i>Categorical</i> |
| <i>Perpetrator_Sex</i> | <i>Categorical</i> |
| <i>Perpetrator_Age</i> | Numerik |
| <i>Perpetrator_Race</i> | <i>Categorical</i> |
| <i>Perpetrator_Ethnicity</i> | <i>Categorical</i> |
| <i>Relationship</i> | <i>Categorical</i> |
| <i>Weapon</i> | <i>Categorical</i> |
| <i>Victim_Count</i> | Numerik |
| <i>Perpetrator_Count</i> | Numerik |
| <i>Record_Source</i> | <i>Categorical</i> |

3.4 Preprocessing Data

Tahap *preprocessing* dalam penelitian ini, dilakukan dengan cara *lowercase* semua kolom untuk mempermudah *slicing* pada *dataframe*. Menjalankan fungsi *python* “*numpy.unique*” pada atribut untuk mendeteksi *outlier*, *outlier* distandarisasi jika sedikit, dihapus jika atribut memiliki banyak *outlier*. Pencilan

pada grafik yang tidak sesuai dengan hasil survei pada *serial killer database* juga dihapus, serta memperbaiki *typo* pada *item data*.

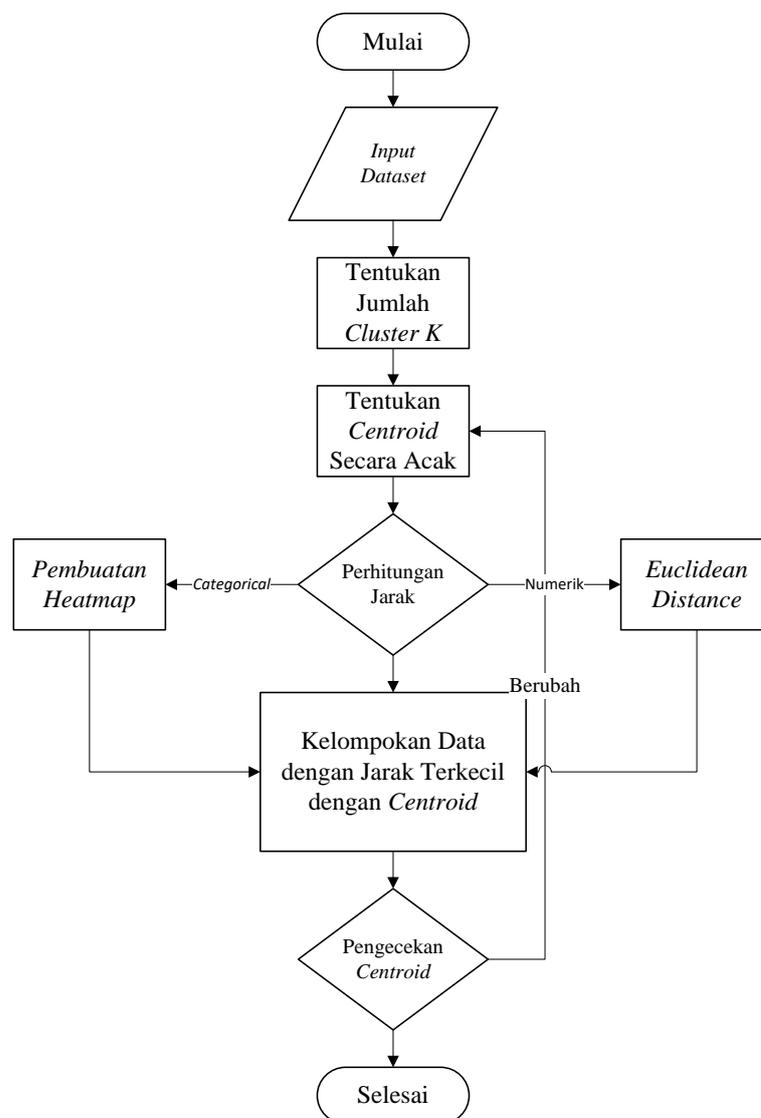
3.5 Seleksi Fitur

Pemilihan fitur pada data kasus pembunuhan di Amerika Serikat akan dipilih berdasarkan dua metode, metode pertama memilih atribut yang berkaitan dengan ciri-ciri korban karena secara heuristik ditemukannya korban adalah awal ditemukannya suatu kasus. Oleh karena itu atribut yang akan diambil dari data kasus pembunuhan di Amerika Serikat adalah *victim sex*, *victim age*, *victim race*, *victim ethnicity*, *relationship*, *weapon*, dan *state*. *Victim sex* hingga *victim ethnicity* dipilih sebagai atribut yang menyimpan ciri-ciri dari korban pembunuhan. Atribut *relationship* tetap dipertahankan untuk mengetahui apakah korban memiliki relasi dengan pembunuh atau tidak. Pelaku pembunuh berantai, dengan kemampuan kognitif di atas rata-rata akan menghindari pembunuhan terhadap orang yang memiliki relasi dekat dengan mereka. Sehingga *instance* yang dipertahankan dalam atribut *relationship* adalah *stranger*, *unknown*, *acquittance*, *neighbor*. Kolom *weapon* akan dipakai untuk mengelompokan pelaku berdasarkan *modus operandi*. Sementara untuk membatasi pengelompokan agar tidak terlalu luas, area akan dibatasi dengan atribut *state*.

Metode kedua dalam seleksi fitur adalah dengan membuat *correlation plot* lalu menghapus salah satu variabel dengan nilai korelasi positif yang tinggi yang dalam penelitian ini lebih tinggi dari 0,9. Namun *correlation plot* hanya dipertimbangkan jika hasil *dimensionality reduction* dengan metode lain tidak sesuai ekspektasi. Kemudian seleksi variabel yang akan digunakan juga mempertimbangkan hasil *Multi Correspondance Analysis (MCA)* pada atribut *categorical*, dengan mengambil nilai korelasi paling rendah dari dimensi pertama sesuai dengan saran Husson Francois sebagai pencetus algoritma *MCA* pada video (<https://www.youtube.com/watch?v=S11-UYD6iac>). Sementara data numerik yang diambil menggunakan *Principal Component Analysis (PCA)* berdasarkan *variance* tertinggi yang didapatkan dari program dengan menggunakan *python library* yaitu “`sklearn.decomposition.PCA`”.

3.6 K-Prototype Clustering

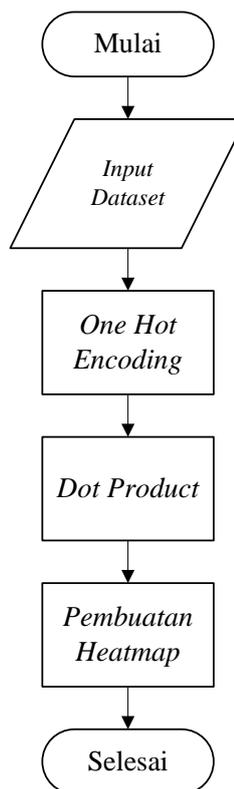
Tahapan dari *K-Prototype Clustering* dengan memasukan data hasil *preprocessing*, dan menentukan jumlah kluster k . Algoritme dijalankan berdasarkan *hit and trial* yang hingga 10 kluster untuk membatasi ruang lingkup penelitian. Hasil percobaan kluster selanjutnya dibandingkan pada langkah evaluasi dan juga perbedaan *heatmap* yang terbentuk. Hasil dari kluster masuk sebagai atribut baru “cluster” pada *Data Frame* yang sebelumnya dimasukan sebagai sampel ke dalam program. Langkah *K-Prototype* digambarkan dengan diagram alir pada Gambar 3.2.



Gambar 3.2. Diagram Alir *K-Prototype Clustering*

3.7 Pembuatan Heatmap

Pembuatan *heatmap* adalah sebagai pengganti simulasi dari sistem pendukung keputusan Dragnet, karena keterbatasan data untuk memproyeksikan *heatmap* pada peta. Sebagai alternative, *heatmap* diproyeksikan ke dalam bidang dua dimensi menggunakan *python library* seaborn. *Heatmap* dibuat berdasarkan kesamaan setiap *record* pada *Data Frame* menggunakan *dot product* dari hasil *One-Hot Encoding* yang sebelum proses ini data numerik dibagi menjadi data *categorical* untuk memenuhi syarat *One-Hot Encoding* yang hanya menerima data *categorical*. Semakin banyak baris yang memiliki tingkat *similarity* tinggi maka semakin terkonsentrasi juga intensitas *heatmap* yang tergambar. Diagram alir pembuatan *heatmap* ditunjukkan oleh Gambar 3.3.



Gambar 3.3. Diagram Alir Pembuatan *Heatmap*

3.8 Evaluasi

Evaluasi dari penelitian ini adalah membandingkan hasil pembuatan *heatmap* pada data pembunuhan baik tanpa atau dengan proses *K-Prototype Clustering*.

Hasil *cluster* dievaluasi dengan *Silhouette Coefficient*. Hipotesis dalam penelitian ini, adalah jika pengelompokan data yang dihasilkan oleh *K-Prototype* memberikan pola *heatmap* yang berbeda dan lebih terkelompok. Jika hipotesis benar, dapat disimpulkan bahwa algoritme *K-Prototype Clustering* dapat menunjang performansi dari *Geographic Profiling* dengan mengklaster data *input* yang lebih valid dan sudah terkelompok.