BAB II

LANDASAN TEORI

2.1 DATA MINING

Data mining mempunyai fungsi yang penting untuk membantu mendapatkan informasi yang berguna serta meningkatkan pengetahuan bagi pengguna. Berikut ini merupakan beberapa pengertian data mining :

Menurut Kamber et al (2012 : 16) menyatakan: "*Mining* adalah istilah yang sangat jelas untuk menggambarkan proses yang menemukan sejumlah kecil nominal berharga dari banyaknya bahan mentah".

Menurut Retno Tri Vulandari (2017:1) menyatakan bahwa:

"*Data Mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstrasi dan mengenali pola yang penting atau menarik dari data yang terdapat pada basis data. Data mining terutama digunakan untuk mencari pengetahuan yang terdapat dalam basis data yang besar sehingga sering disebut *Knowledge Discovery Databases* (KDD)".

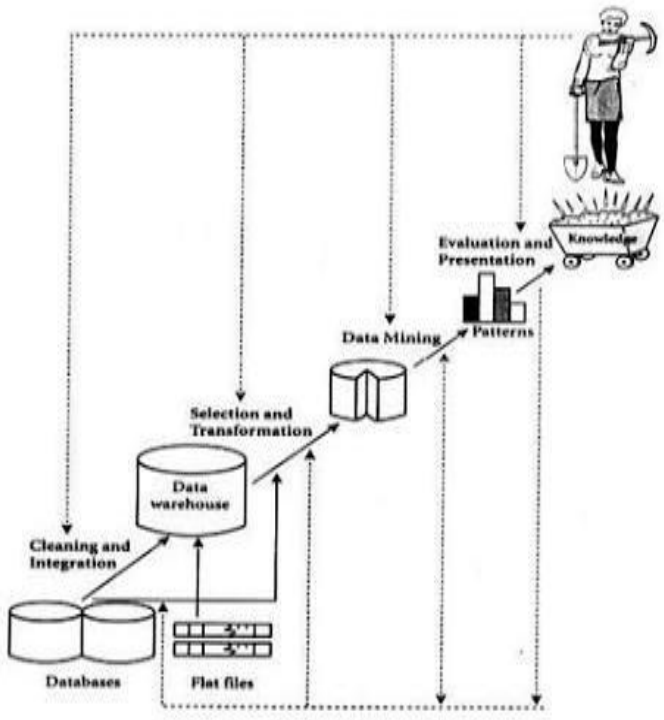
Menurut Wardhani dalam jurnal Mentari Tri Indah Rahmayan (2018:42) menyatakan bahwa:

"*Data mining* disebut juga dengan *pattern recognition* yang merupakan metode dalam pengolahan data untuk menemukan pola tersembunyi dari data yang diolah kemudian menghasilkan suatu pengetahuan baru yang bersumber dari data lama, hasil dari pengolahan data tersebut dapat digunakan dalam menentukan keputusan di masa yang akan datang".

Dari pengertian diatas dapat disimpulkan bahwa *data mining* merupakan proses mencari pola atau informasi menarik dalam data terpilih menggunakan teknik atau metode tertentu untuk mendapatkan pengetahuan yang tersembunyi dari kumpulan data yang berukuran besar.

2.2 PROSES TAHAPAN *DATA MINING*

Banyak orang memperlakukan penambangan data sebagai sinonim untuk istilah lain yang digunakan secara populer, penemuan pengetahuan dari data, atau KDD, sementara yang lainnya melihat penambangan data hanya sebagai langkah penting dalam proses penemuan pengetahuan. Proses penemuan pengetahuan ditunjukkan pada gambar 2.1 (Kamber et al 2012 : 17)



Gambar 2.1 *Tahapan Data Mining*(Han, dan Kamber 2006)

Sebagai urutan iteratif dari langkah beikut :

1. Data *Cleaning*

   Pembersihan data untuk menghapus kebisingan dan data yang tidak konsisten.

2. Data *Integration*

   Integrasi data di mana beberapa sumber data mungkin digabungkan sebuah tren populer di industri informasi adalah untuk melakukan pembersihan data dan integrasi data sebagai langkah preprocessing, di mana data yang dihasilkan disimpan dalam gudang data.

3. Data *Selection*

   Pemilihan data di mana data yang relevan dengan tugas analisis diambil dari database.

4. Data *Transformation*

   Transformasi data di mana data diubah dan dikonsolidasikan ke dalam formulir yang sesuai untuk penambangan dengan melakukan ringkasan atau operasi agregasi, terkadang transformasi data dan konsolidasi dilakukan sebelum proses pemilihan data, khususnya dalam hal pergudangan data. Pengurangan data juga dapat dilakukan untuk mendapatkan representasi yang lebih kecil dari data asli tanpa mengorbankan integritasnya.

5. Data Mining

   Data mining proses penting di mana metode cerdas diterapkan untuk mengekstrak pola data

6. *Pattern Evaluation*

   Evaluasi pola untuk mengidentifikasi pola yang sangat menarik yang mewakili pengetahuan berdasarkan tindakan ketertarikan.

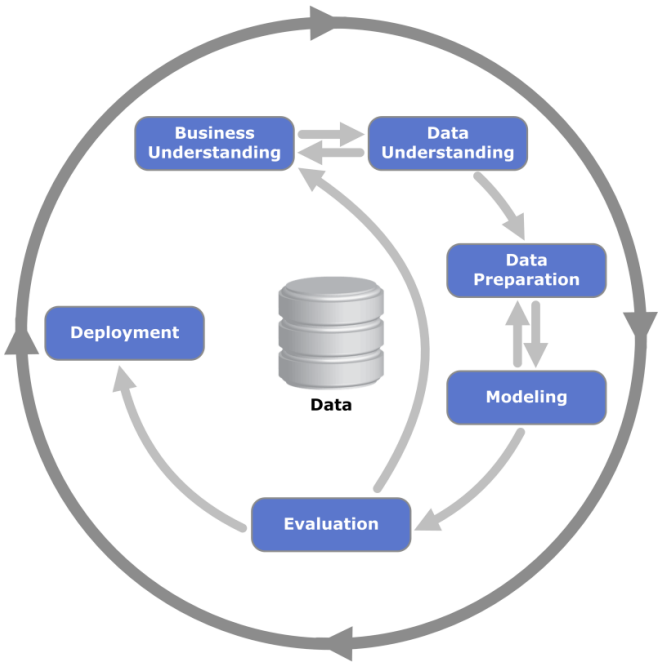
7. *Knowledge Presentation*

   Presentasi pengetahuan dimana visualisasi dan teknik representasi pengetahuan digunakan untuk mempresentasikan pengetahuan yang ditambang kepada pengguna.

Langkah 1 sampai 4 adalah berbagai bentuk preprocessing data, di mana data disiapkan untuk pertambangan. Langkah penambangan data dapat berinteraksi dengan pengguna atau basis pengetahuan. Pola yang menarik disajikan kepada pengguna dan dapat disimpan sebagai pengetahuan baru dalam basis pengetahuan.

2.3 MODEL CRIPS-DM

Proses *Data mining* berdasarkan CRIPS-DM terdiri dari 6 fase yaitu (Budiman, Prahasto, & Christyono, 2012)



Gambar 2.2 *Model Crips-dm*(Budiman, Prahasto, & Christyono, 2012)

1. *Business Understanding* adalah pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan, kebutuhan dari perpektif bisnis.
2. *Data Understanding* adalah fase mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai, mengidentifikasi masalah yang berkaitan dengan kwantitas data, mendeteksi subset yang menarik data untuk membuat hipotesa awal.
3. *Data Preparation* sering disebut sebagai fase yang padat karya aktivitas yang dilakukan antara lain memilih *table* dan *field* yang akan ditransformasikan ke dalam database baru untuk bahan *data mining.*

4. *Modeling* adalah fase menentukan teknik data mining yang digunakan, menentukan *tools data mining* , teknik *data mining* , algoritma *data mining*, menentukan parameter dengan nilai yang optimal.
5. *Evaluation* adalah fase interpretasi terhadap hasil data mining yang ditunjukan dalam proses pemodelan pada fase sebelumnya.
6. *Deployment* atau penyebaran adalah fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses *data mining.*

2.4 PENGELOMPOKKAN *DATA MINING*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Wicaksana et.al 2013:43):

1. *Classification* (Klasifikasi)

   Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Salah satu contoh yang mudah dan popular adalah dengan Decision tree yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

2. *Association* (Asosiasi)

   Digunakan untuk mengenali kelakuan dari kejadiankejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Salah satu contohnya adalah Market Basket Analysis, yaitu salah sati metode asosiasi yang menganalisa kemungkinan pelanggan untuk membeli beberapa item secara bersamaan.

3. *Clustering* (Pengklusteran)

   Digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya tool data mining. Biasanya menggunkan metode neural network atau statistik. Clustering membagi item menjadi kelompok-kelompok berdasarkan yang ditemukan tool data mining.

2.5 KELOMPOK KLASTERISASI (*CLUSTERING)*

Berdasarkan pengelompokkan *data mining* di atas, penulis mengambil metode klasterisasi dalam penelitian ini, karena pada dasarnya merupakan metode segmentasi data yang bertujuan menemukan kelompok (*cluster*) objek, yang akan digunakan sesuai dengan tujuan analisis data.

Menurut Kamber et al (2012 :362)) menyatakan bahwa:

“Analisis cluster atau hanya *clustering* adalah proses memartisi sekumpulan objek data (pengamatan) menjadi himpunan bagian. Setiap subset adalah sebuah *cluster*, sehingga objek dalam cluster mirip satu sama lain, namun berbeda dengan objek di *cluster* lain”.

Menurut Tan dalam buku Eko Prasetyo (2012 : 173) “Analisis kelompok (cluster analysis) adalah pekerjaan mengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya”.

Menurut Retno Tri Vulandari (2017: 54) “*Clustering* adalah metode *data mining* yang *unsupervised*, karena tidak ada satu atributpun yang digunakan untuk memandu proses pembelajaran, jadi seluruh atribut input diperlakukan sama”.

Dari pengertian tersebut dapat disimpulkan *clustering* adalah metode yang digunakan untuk membagi rangkaian data menjadi group berdasarkan kesamaan yang telah ditentukan sebelumnya. Dengan menggunakan klasterisasi, kita dapat mengidentifikasi daerah yang padat, menentukan pola-pola distribusi secara keseluruhan, dan menemukan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut-tribut data.

Metode klasterisasi yang dikelompokkan ke dalam empat kategori: metode berbasis partisi (*partitioning methods*), metode berbasis hirarki (*hierarchical methods*), metode berbasis kepadatan (*density-based methods*), dan metode berbasis kisi (*grid-based methods*).

2.5.1 Metode Berbasis Partisi

Sesuai dengan namanya, metode ini bekerja dengan cara membagi atau mempartisi data ke dalam sejumlah kelompok. Metode ini dikenal juga dengan metode berbasis pusat atau metode berbasis representatif (Zaki et al dalam Suyanto 2017: 260) karena bekerja dengan menentukan pusat-pusat klaster, di mana pusat klaster bisa berupa rata-rata, modus, atau sebuah objek representatif

dari semua objek dalam suatu klaster berdasarkan suatu ukuran tertentu. Algoritma-algoritma yang termasuk ke dalam metode berbasis partisi adalah:

1. *K-Means*

   Metode *k-means* merupakan algoritma klasterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya. Ide dasar algoritma *k-means* sangatlah sederhana, yaitu meminimalkan *Sum of Squared Error* (SSE) antara objek objek data dengan sejumlah *k centroid*. Algoritma *k-means* bekerja dengan empat langkah: Pertama, dari himpunan data yang akan diklasterisasi, dipilih sejumlah *k* objek secara acak sebagai *centroid* awal. Kedua, setiap objek yang bukan *centroid* dimasukkan ke klaster terdekat berdasarkan ukuran jarak tertentu. Ketiga, setiap *centroid* diperbarui berdasarkan rata-rata dari objek yang ada di dalam setiap klaster. Keempat, langkah kedua dan ketiga tersebut diulang-ulang (diiterasi) sampai semua *centroid* stabil atau konvergen, dalam arti semua *centroid* yang dihasilkan dalam iterasi saat ini sama dengan semua *centroid* yang dihasilkan pada iterasi sebelumnya.

2. *K-Modes*

   Algoritma *k-means* menggunakan ukuran rata-rata. Oleh karena itu, *k-means* hanya dapat diaplikasikan pada himpunan objek dengan atribut yang dapat dikonversi ke numerik sehingga dapat dihitung rata-ratanya. Algoritma k-means tidak mungkin diaplikasikan pada himpunan objek dengan atribut bernilai nominal. Untuk mengatasi masalah ini, dapat dimodifikasi

algoritma *k-means* yang tadinya menggunakan rata-rata *k-modes* yang menggunakan modus atau *modes* (nilai yang paling sering muncul). Konsekuensinya, tentu saja harus menggunakan ukuran *dissimilarity* yang berhubungan dengan objek data bernilai nominal dan menggunakan suatu metode berbasis frekuensi untuk menjadi memperbarui modus dalam setiap klaster.

3. *K-Medoids*

   Algoritma *k-medoids* menggunakan teknik berbasis objek representatif (perwakilan) yang disebut *medoids* untuk mengatasi kelemahan *k-means* yang sensitif terhadap derau dan pencilan. Caranya dengan menghilangkan penggunaan rata-rata untuk memperbarui *centroid* dan menggantinya dengan objek aktual sebagai representasi dari suatu klaster. Jadi, algoritmia *k-medoids* melakukan partisi dengan cara meminimalkan jumlah *dissimilarity* antara setiap objek *p* dan objek representatif terdekat, yaitu menggunakan jumlah kesalahan absolut.

4. *Fuzzy C-Means*

   Fuzzy C-Means (FCM), yang dikenal juga sebagai *Fuzzy* ISODATA (Wu dalam Suyanto, 2017: 267) Pada dasarnya, cara kerja FCM mirip dengan *k-means*. Dua konsep fundamentalma yang membedakan FCM dengan k-means adalah (Bezdek et al dalam Suyanto, 2017: 268)

   a. Pada FCM, setiap objek dibiarkan menjadi anggota dari semua *k* klaster dengan derajat keanggotaan berbeda-beda yang jika dijumlahkan sama dengan 1.

b. FCM menggunakan fungsi objektif yang dapat dipandang sebagai total variansi objek dari centoid c_i.

2.5.2 Metode Berbasis Hirarki

Sesuai dengan namanya, metode klasterisasi hirarki (*hierarchical clustering*) bekerja dengan cara mengelompokkan objek-objek data ke dalam sebuah hirarki klaster. Namun, bukan berarti objek-objek data memliki struktur bertingkat-tingkat seperti dalam struktur organisasi perusahaan atau institusi. Hirarki di sini hanya untuk merangkum dan merepresentasikan data secara ringkas agar mudah dalam visualisasi. Algoritma-algoritma yang termasuk ke dalam metode berbasis hirarki adalah:

1. BIRCH

   Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) bekerja dengan cara mempartisi objek secara hirarki menggunakan struktur pohon, di mana simpul daun (*leaf node*) atau simpul bukan daun (*nonleaf node*) yang berada di tingkat bawah dipandang sebagai "klaster mikro". Kemudian melakukan klasterisasi makro terhadap klaster-klaster mikro tersebut menggunakan algoritma klasterisasi.

   BIRCH dirancang untuk mengklasterisasi data numerik yang berukuran besar dengan menggabungkan teknik klasterisasi hirarki dan metode metode klasterisasi lain seperti partisi iteratif. Dengan strategi ini, BIRCH dapat mengatasi kelemahan metode *agglomerative clustering* dalam skalabilitas dan ketidakmampuan untuk membatalkan apa yang telah dilakukan di langkah sebelumnya.

2. *Chameleon*

   Metode *chameleon* bekerja dengan cara mengeksploasi pemodelan dinamis dalam klasterisasi hirarki Pemodelan dinamis digunakan untuk menghitung *similarity* antar klaster, yaitu berdasatkan: seberana kuat objek-objek yang terhubung dalam suatu klaster dan kedekatan klaster-klaster. Dua klaster digabungkan jika memiliki tingkat keterhubungan yang tinggi dan saling berdekatan. Berbeda dengan BIRCH, *Chameleon* tidak bergantung pada model statis yang dibuat oleh *user*. *Chameleon* dapat secara otomatis beradaptasi tehadap karakteristik internal dari klaster-klaster yang digabungkan. Proses penggabungan dapat menemukan klaster-laster yang homogen dan alami. Yang menarik proses penggabungan dapat digunakan untuk senua jenis atribut data asalkan ukuran *similarity* dapat ditentukan.

2.5.3 Metode Berbasis Kepadatan

Secara konsep metode klasterisasi yang berbasis partisi maupun hirarki cenderung lemah untuk klaster kalster yang berbentuk bebas dan acak (tidak bulat). Apalagi jika terdapat derau atau pencilan pada klaster-kalster tersebut. *Chameleon* sebagai metode berbasis hirarki dapat menangani klaster-kalster yang berbentuk bebas dan acak, namun kompleksitas komputasinya cenderung besar. Untuk mengatasi kedua masalah ini, dapat menggunakan strategi lain yang disebut dengan metode berbasis kepadatan Pada subbab ini kita akan membahas tiga algoritma yang mengzunakan metode berbasis kepadatan, yaitu DBSCAN, OPTICS, dan DENCLUE.

1. *DBSCAN*

Sesuai denngan namanya, *Density-Based Clustering Based on Connected Regions with High Density*, DBSCAN menemukan klaster berdasarkan region-region dengan kepadatan tinggi yang terhubung (*connected regions with high density*). Kepadatan dari suatu objek o diukur berdasarkan jumlah objek yang dekat dengan o.

DBSCAN merupakan metode pertama yang berbasis kepadatan, dengan konsep klasterisası pada DBSCAN yang sederhana. Pertama DBSCAN mencari objek objek inti (*core objects*), yaitu objek yang memililki ketetanggaan padat. Kedua, DBSCAN menghubungkan objek-objek inti tersebut dengan objek-objek tetangganya untuk membentuk region-region padat. Region padat itulah yang dinyatalan sehagai klaster (Krirgel et al dalam Suyanto, 2017: 279)

2. *OPTICS*

Sesuai dengan namanya, *Ordering Points to Identify the Clustering Structure,* metode ini mengeluarkan sebuah pengurutan klaster atau *cluster ordering* (tidak secara eksplisit menghasillan klaster). *Cluster ordering* ini berupa sebuah daftar linier dari semua objek dan merepresentasikarn struktur klasterisasi berbasis kepadatan Objek-objek dalam suatu klaster yang lebih padat dimasukkan ke daftar secara lebih berdekatan satu sama lain di dalam *cluster ordering*. Sementara itu, objek-objek dalam suatu klaster dengan kepadatan rendah dimasukkan ke daftar secara lebih berjauhan satu sama lain di dalam cluster *ordering*. Pengurutan ini seolah

seolah sama dengan klasterisasi DBSCAN yang dihasilkan dari berbagai variasi parameter ε dan *Minobj*.

3. *DENCLUE*

   Untuk mengatasi sensitivitas DBSCAN dan OPTICS terhadap radius ketetanggaan ε. Alexander Hinneburg dan Daniel A Keim mengusulkan metode lain yang diberi nama DENCLUE atau *DENsity-based CLUstEring* (Hinneburg et al dalam Suyanto, 2017: 286). DENCLUE menggunakan fungsi distribusi kepadatan yang dikenal sebagail *kernel density estimation*, sebuah pendekatan estimasi kepadatan tanpa parameter.

2.5.4 Metode Berbasis Kisi

Metode klasterisasi yang berbasis partisi, hirarki, dan kepadatan memiliki kompleksitas komputasi yang tinggi karena menggunakan pendekatan *data driven* (disetir oleh data), di mana klasterisasi dilakukan dengan mengikuti distribusi objek data dalam ruang penggabungan (*embedding space*). Jika cenderung menghindari metode metode dengan kompleksitas, maka dapat menggunakan metode berbasis kisi (*grid-based method*). Metode berbasis kisi menggunakan pendekatan *space-driven* (disetir ruang) dengan mempartisi *embedding space* ke dalam sel-sel yang yang tidak bergantung pada distribusi objek data. Metode ini menggunakan struktur data kisi multiresolusi. Ruang objek dikuantisasi ke dalam sejumlah sel yang membentuk struktur kisi di mana semua operasi klasterisasi dilakukan. Strategi ini memberikan dua keuntungan: komputasi yang sangat cepat dan tidak bergantung pada jumlah objek data (melainkan hanya bergantung pada jumlah sel).

1. *STING*

   STatistical INformation Grid (STING) menggunakan teknik klasterisasi multiresolusi berbasis kisi di mana ruang spasial dari objek-objek data dibagi ke dalam sel-sel persegi yang bertingkat dan rekursif (yang menunjukkan level resolusi yang berbeda dan membentuk struktur hirarki, di mana setiap sel pada tingat yang tinggi dipartisi untuk membentuk sejumlah sel pada level yang lebth rendah. Informasi statistik yang berkaitan dengan atribut dalam setiap sel kisi, seperti rata-rata, nilai maksimum, dan nilai minimum, dihitung lebih dulu dan disimpan sebagai parameter statistik (yang berguna untuk proses *query* dan untuk analisis data yang lain).

2. *CLIQUE*

   CLustering In QUEst atau CLIQUE dirancang untuk mencari klaster berbasis kepadatan dalam sub-sub ruang. CLIQUE mempartisi setiap dimensi ke dalam interval-interval yang tidak tumpang tindih sehingga mampu membagi seluruh ruang objek data ke dalam sel-sel. Sebuah batas kepadatan *l* digunakan untuk mengidentifikasi sel-sel yang padat dan jarang, Sebuah sel dikatakan padat jika jumlah objek yang dipetakan lebih besar dari *l*. (Suyanto 2017: 293).

2.6 ALGORITMA *K-MEANS*

Berdasarkan metode dari pengelompokkan *data mining* di atas, penulis menggunakan metode algoritma *k-means* karena metode ini paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya dan juga proses dalam metode *k-means* dapat dipahami dengan baik oleh penulis.

Menurut Kamber et al (2012 : 369) mendefinisikan bahwa "Algoritma k-means merupakan centroid dari sebuah cluster sebagai nilai rata-rata dari poin-poin di dalam cluster".

Menurut Retno Tri Vulandari (2017: 54) mengatakan bahwa:

"*K-Means* merupakan algoritma *clustering* yang berulang-ulang. Algoritma *k-means* menetapkan nilai-nilai *cluster* secara random, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid, mean* atau *means*."

Menurut Fajar Astuti Hermawati dalam jurnal Elmayanti (2017: 358) mengatakan bahwa:

"*K-Means* adalah menggunakan pendekatan partitional clustering. Tiap cluster dihubungkan dengan sebuah centroid (titik pusat). Tiap titik ditempatkan ke dalam cluster dengan centroid terdekat. Jumlah cluster, K. harus ditentukan".

Sarwono didalam jurnal Nurul Rohmawati et al (2015:93) mengemukakan secara lebih detail, algoritma *k-means* adalah sebagai berikut:

1. Menentukan k sebagai jumlah kluster yang ingin di bentuk.

2. Membangkitkan nilai random untuk pusat cluster awal (centroid) sebanyak k.

3. Menghitung jarak setiap data input terhadap masing – masing centroid menggunakan rumus jarak Euclidean (Euclidean Distance) hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Berikut adalah persamaan Euclidian Distance:

   $d(xi, \mu j) = \sqrt{\sum(xi, \mu j)^2}$ ……2.1

   Dimana :

   xi : data kriteria,

   µj : centroid pada cluster ke-j

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).

5. Memperbaharui nilai centroid. Nilai centroid baru di peroleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus:

   $\mu j\ (t\ + 1) = \frac{1}{Nsj} \sum_{j \in sj} xj$ ……2.2

   Dimana:

   µj(t+1) : centroid baru pada iterasi ke (t +1)

   Nsj : banyak data pada cluster Sj.

6. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap cluster tidak ada yang berubah.

Jika langkah 6 telah terpenuhi, maka nilai pusat cluster (µj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

Menurut Anggoro Eko Wicaksono (2016:3) didalam jurnalnya: "*k-means* merupakan salah satu metode clustering non-hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster. Algoritma ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokan ke dalam satu *cluster* yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam *cluster* yang lain. Algoritma ini sederhana untuk diterapkan dan dijalankan, relatif cepat, mudah diadaptasi, dan umum digunakan dalam praktek. Berikut ini adalah tahap-tahap algoritma *k-means clustering* :

1. *Input* data yang digunakan dalam *clustering*. Data ini digunakan untuk menentukan nilai rata-rata *data point* yang berada dalam satu *cluster* dan menentukan jarak dari setiap *data point* ke *centroid*.
2. Alokasi ke *cluster* secara acak, dalam tahap ini pertama kalinya *data point* dialokasikan ke *cluster* secara acak tanpa ada kriteria tertentu.
3. Hitung *centroid data point* yang ada pada setiap *cluster*. Nilai *centroid* pada *k-means* digunakan sebagai pusat *cluster*. Dengan menentukan anggota *cluster* secara acak pada tahap sebelumnya, maka terbentuk iterasi awal sebagai pusat *cluster* acak.
4. Alokasi ke *centroid* terdekat, pada tahap ini hasil *centroid* dari setiap *cluster* sudah diketahui, kemudian *datapoint* dialokasikan pada *centroid* terdekat berdasarkan nilai jarak *similarity data point* terhadap *centroid*. Jarak

similarity dari *data point* ke *centroid* pada masing masing *cluster* diperoleh dari perhitungan *euclidean distance*. Kemudian nilai jarak setiap *data point* ke *centroid cluster* dibandingkan, dan *data point* menjadi anggota dari *cluster* berdasarkan jarak *data point* ke *centroid* terdekat.

5. Konvergen, mengalokasikan *data point* ke *centroid* dengan nilai jarak terdekat, dengan menguji apakah *cluster* yang terbentuk telah membentuk *cluster* yang konvergen atau tidak. *Cluster* dinyatakan konvergen jika anggota dari masing masing"

2.7 *DATA MINING TOOLS*

Ada banyak *tools* yang tersedia untuk *data mining.* Tujuan utama *data mining tools* adalah untuk menemukan data, mengekstrak data, menyaring data, mendistribusikan informasi dan memonetisasinya. Berikut beberapa penjelasan *data mining tools* antara lain:

1. RAPIDMINER

Rapid Miner adalah sebuah software untuk pengolahan data mining. Rapid Miner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. "RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik" (Setiawan dalam jurnal Purwanto & Darmadi, 2018:45). Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang diintegrasikan pada

produknya sendiri. RapidMiner ditulis menggunakan bahasa java sehingga dapat bekerja disemua sistem operasi.

Beberapa fitur dari Rapid Miner sebagai berikut :

a. Banyaknya algoritma data mining, seperti decision tree.
b. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D scatter plots.
c. Banyaknya variasi plugin, seperti plugin untuk melakukan analisis teks.
d. Menyediakan prosedur data mining dan machine learning termasuk ETL (Extraction, Transformasi, Loading), data preprocessing, visualisasi, modeling dan evaluasi.
e. Proses data mining tersusun atas operatoroperator yang nestable,

dideskripsikan dengan XML dan dibuat dengan GUI.

2. SPSS

Dari berbagai program olah data statistik lainnya, *SPSS* merupakan Program yang paling banyak digunakan. Menurut Ahmad Sujana dan Hana Zainab Mukarromah (2017 : 51) “SPSS merupakan sebuah aplikasi dari Microsoft yang berfungsi sebagai aplikasi statistik untuk mengolah data. SPSS for Windows adalah sebuah program aplikasi yang dirilis pada tahun 1992 dengan kemampuan analisis statistik yang cukup tinggi. SPSS yang pertama kali dirilis adalah SPSS/PC+ berbasis teks pada tahun 1984. Aplikasi ini dapat menggunakan

program atau kode eksternal, yang hanya membutuhkan software bantu lain berupa editor".

3. Rattle

Rattle adalah sebuah aplikasi penggalian data grafis yang dibangun di atas bahasa statistik pemahaman R. R tidak diperlukan untuk menggunakan Rattle. Rattle ini mudah digunakan, cepat untuk menyebarkan, dan memungkinkan kita untuk bekerja dengan cepat melalui pengolahan data, permodelan, dan evaluasi dari proyek *data mining*.

4. Orange Ailab

Orange Ailab adalah perangkat lunak open source yang memungkinkan pengguna yang tidak memahami sedikitpun tentang pemrograman dapat melakukan visualisasi dan analisis data. Fitur-fitur yang dimiliki diantaranya scatterplots, bar charts, trees, dendrograms, networks dan heatmaps (Mujiasih, 2015 : 191).

2.8 RAPID MINER

Dalam penelitian ini menggunakan *tools Rapidminer* karena mampu mengolah data mining, untuk melakukan analisis terhadap data mining dan memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Pekerjaan yang dilakukan oleh *RapidMiner text mining* adalah berkisar dengan analisis teks, mengekstrak pola-pola dari data set yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan *database*. *Rapid Miner* merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technologi Blanchardstown* dan Ralf Klinkenberg dari

rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Tujuan dari analisis teks ini adalah untuk mendapatkan informasi bermutu tertinggi dari teks yang diolah.

2.9 P3AP2KB

P3AP2KB atau pemberdayaan Perempuan, Perlindungan Anak, Pengendalian Penduduk, dan Keluarga berencana. Kementerian Pemberdayaan Perempuan dan Perlindungan Anak (dahulu Kementerian Negara Pemberdayaan Perempuan dan Perlindungan Anak, disingkat PP & PA) adalah kementerian dalam Pemerintah Indonesia yang membidangi urusan pemberdayaan perempuan dan perlindungan anak. Kementerian PP & PA dipimpin oleh seorang Menteri Pemberdayaan Perempuan dan Perlindungan Anak (Meneg PP & PA) yang sejak tanggal 23 Oktober 2019 dijabat oleh I Gusti Ayu Bintang Darmawati.

Kementerian Pemberdayaan Perempuan dan Perlindungan Anak mempunyai tugas menyelenggarakan urusan di bidang pemberdayaan perempuan dan perlindungan anak dalam pemerintahan untuk membantu Presiden dalam menyelenggarakan pemerintahan negara. Dalam melaksanakan tugas sebagaimana dimaksud di atas, Kementerian Pemberdayaan Perempuan dan Perlindungan Anak menyelenggarakan fungsi:

1. Perumusan dan penetapan kebijakan di bidang pemberdayaan perempuan dan perlindungan anak;

2. Koodinasi dan sinkronisasi pelaksanaan kebijakan di bidang pemberdayaan perempuan dan perlindungan anak;
3. Pengelolaan barang milik/kekayaan negara yang menjadi tanggung jawab Kementerian Pemberdayaan Perempuan dan Perlindungan Anak; dan
4. Pengawasan atas pelaksanaan tugas di lingkungan Kementerian Pemberdayaan Perempuan dan Perlindungan Anak

2.10 TINJAUAN STUDI

Penulis memulai penelitian ini dengan terlebih dahulu melakukan studi kepustakaan dari penelitian-penelitian dan sumber-sumber lain. Beberapa penelitian telah banyak dilakukan dengan menggunakan teknik *data mining* untuk menggali berbagai informasi dari sebuah *database*. Penelitian tersebut membahas tentang topik yang terkait dengan penelitian penulis, antara lain adalah penelitian mengenai algoritma yang akan digunakan penulis.

Tabel 2.1 TINJAUAN STUDI

NO	Judul	Penulis dan tahun	Masalah	metode	Hasil
1	IMPLEMENTASI ALGORITMA *C4.5* DALAM MENENTUKAN LOKASI PRIORITAS PENYULUHAN PROGRAM KELUARGA BERENCANA DI KECAMATAN DUMAI TIMUR	Febrina Sari dan David Saro 2017	Bagaimana Algoritma C4.5 dapat menentukan lokasiprioritas Penyuluhan program keluarga berencana , Karena algoritma C4.5 di gunakanuntuk melakukan klasifikasi, jadi hasil dari pengolahan test data set berupa Pengelompokan data dalam kelas-kelasnya, yang mana kelas dibagi menjadi dua yakni tidak prioritas atau ya prioritas.	Menggunakan metode Pohon keputusan algoritma C4.5	bagaimana Algoritma C4.5 dapat menentukan lokasi prioritas penyuluhan program keluarga berencana, Karena algoritma C4.5 digunakan untuk melakukan klasifikasi, jadi hasil dari pengolahan test dataset berupa pengelompokan data dalam kelas-kelasnya, yang mana kelas dibagi menjadi dua yakni tidak prioritas atau ya prioritas.
2	KETEPATAN KLASIFIKASI KEIKUTSERTAAN KELUARGA BERENCANA MENGGUNAKAN	Fajar Heru Setiawan, Rita Rahmawati dan Suparti 2015	Untuk menganalisis pasangan usia subur mengikuti KB atau tidak dapat menggunakan metode regresi logistik biner atau regresi probit	Menggunakan metode Klasifikasi dan Analisis regresi logistik	didapatkan hasil bahwa variabel pendidikan ayah dan tingkat kesejahteraan tidak signifikan terhadap variabel keikutsertaan KB, hal ini ditunjukkan dengan nilai Wj < 1;0,05 2 F = 3,84 atau

	REGRESI LOGISTIK BINER DAN REGRESI PROBIT BINER (Studi Kasus di Kabupaten Semarang Tahun 2014)		biner karena untuk variabel dependennya (keikutsertaan KB) hanya memiliki dua nilai dan bersifat kualitatif .		nilai p-value. .DUHQDDGDGXDYDULDEHO yang tidak signifikan maka dibentuk model baru yang akan diuji lagi dengan uji rasio likelihood ke-2 yang nilainya pada baris kedua dalam Tabel 6 dan uji wald kedua pada Tabel 8. Dari Tabel 8 didapatkan kesimpulan semua variabel yaitu umur ibu, jumlah anak dan pendidikan ibu signifikan yang ditunjukkan dengan Wj > 1;0,05 2 F = 3,84 atau nilai pvalue .
3	IMPLEMENTASI K-MEANS *CLUSTERING* UNTUK PEMETAAN DESA DAN KELURAHAN DI KABUPATEN BANGKALAN BERDASARKAN CONTRACEPTIVE PREVALENCE RATE DAN TINGKAT	Evy Dwi Cahyati 2017	Tingkat pendidikan yang tinggi menjadikan masyakarat lebih memahami pentingnya keluarga berencana dan kesehatan reproduksi. Berdasarkan hal tersebut, keikutsertaan pasangan usia subur (PUS) pada program Keluarga Berencana di desa dan kelurahan dapat dikelompokkan berdasarkan Contraceptive	Menggunakan metode klasterisasi dan algoritma *k-means*	Hasil visualisasi cluster menggunakan Google Maps, cluster 1 yang dilambangkan dengan marker merah cenderung mengelompok, namun terdapat wilayah yang tidak mengelompok pula. Cluster 2 yang ditandai dengan marker biru tersebar secara mengelompok. Cluster 3 yang ditandai dengan marker hijau tersebar secara mengelompok, namun ada pula wilayah pada Kelompok 3 yang tidak ikut

	PENDIDIKAN		Prevalence Rate (CPR) dan tingkat pendidikan di desa atau kelurahan tersebut		mengelompok
4	*CLUSTER* JUMLAH PENGGUNA ALAT KONTRASEPSI MENGGUNAKAN METODE K-MEANS (Studi Kasus di Kantor KB Kota Yogyakarta)	SHINTYA BUNGA UTAMI 2020.	Berdasarkan latar belakang masalah yang telah diuraikan sebelumnya, maka dirumuskan permasalahan penelitian, yaitu bagaimana menerapkan algoritma K-Means untuk mengcluster data jumlah peserta KB berdasarkan minat peserta KB dalam menggunakan alat kontrasepsi	Menggunakan metode klasterisasi dan algoritma *k-means.*	Hasil pengelompokkan data menggunakan metode k-means clustering menghasilkan 2 kelompok. Kelompok 1 yaitu minat peserta KB yang menggunakan alat kontrasepsi bersifat sementara dengan jumlah data sebanyak 8 data dan kelompok 2 yaitu minat peserta KB yang menggunakan alat kontrasepsi bersifat permanen dengan jumlah data sebanyak 6. tersebut, dengan nilai minimum akhir 4.635.
5	*Implementasi Metode Klastering K-means Untuk Mengelompokan Hasil Evaluasi Mahasiswa.*	Febrizal Alfarasy Syam 2017.	Mengelompokan hasil evaluasi akademik mahasiswa adalah salah satu basis untuk memantau perkembangan kinerja akademik mahasiswa di suatu	Menggunakan metode klasterisasi dan algoritma *k-means.*	Metode Clustering Algoritma K-Means dapat diterapkan pada pengelompokan hasil evaluasi mahasiswa FKIP Universitas Riau. Berdasarkan hasil perhitungan manual dan pengujian dengan software RapidMiner

			universitas. Penelitian yang penulis lakukan adalah di salah satu Perguruan Tinggi di Pekanbaru-Riau yaitu Universitas Riau, tepatnya di Fakultas Keguruan dan Ilmu Pendidikan. Dimana Fakultas tersebut memiliki jumlah mahasiswa hingga tahun 2015 adalah ± 5418 orang.		dengan menggunakan data akademik mahasiswa mendapatkan hasil yang sama. Hasil pengelompokan data akademik mahasiswa dapat berfungsi sebagai acuan bagi perencana akademik untuk memantau dan mengevaluasi perkembangan kinerja akademik setiap mahasiswa.
6	ANALISA DAN IMPLEMENTASI METODE K-MEANS CLUSTERING DALAM PREDIKSI PERSEDIAAN ALAT KONTRASEPSI (STUDI KASUS : KABUPATEN DELISERDANG)	Penda Sudarto Hasugian 2017.	Badan KB ini membutuhkan suatu sistem yang dapat memprediksi tingkat persediaan alat kontrasepsi ditahun yang akan datang dengan melihat dan menganalisa data persediaan alat kontrasepsi yang masuk dan yang keluar dari tahun-tahun sebelumnya	Metode k-means clustering	Dengan menggunakan pemodelan k-means clustering seperti gambar 5.21 di atas, dengan jumlah cluster sebanyak 4 buah, maka didapatkan hasil dengan cluster yang terbentuk adalah 4, sesuai dengan pendefinisian nilai k dengan jumlah cluster_1 ada 3 items, cluster_2 ada 1 items, cluster_3 ada 1 items, cluster_4 ada 2 items, dengan total sejumlah 7 items.

Berdasarkan penelitian terkait diatas, yang membedakan penelitian ini dengan penelitian terkait adalah penulis akan melakukan pengelompokan data keluarga berencana di kecamatan pengabuan, data keluarga berancana tersebut di dapat berdasarkan hasil dari penelitian yang sedang di lakukan, seperti data keluarga yang di dapat dinas P3AP2KB yang akan di kelompokkan berdasarkan desa yang ada. Atribut yang terdapat dalam metode klasterisasi yaitu: Nama Keluarga, Tingkat Kesuburan, Jumlah Anak dan Alamat. Penelitian-penelitian yang telah dilakukan sebelumnya akan menjadi acuan penulis dalam melakukan penelitian.