

BAB II

TINJAUAN PUSTAKA

2.1 Peneliatian Terkait

Tinjauan studi yang menjadi sumber referensi dari penelitian yang penulis buat ini berasal dari beberapa penelitian sejenis sebelumnya. Dari beberapa penelitian tersebut didapatkan berbagai hasil pandangan tentang penerapan klusterisasi data dari masing-masing jurnal. Berikut merupakan beberapa penelitian yang terkait dan relevan dengan penelitian ini.

Penelitian pertama dilakukan oleh Lizda Iswari dan Ervina Gita Ayu yang membuat penelitian tentang Pemanfaatan Algoritma K-Means Untuk Pemetaan Hasil Klasterisasi Data Kecelakaan Lalu Lintas. Dimana dalam penelitian ini, peneliti menggunakan metode clustering untuk mengelompokkan daerah rawan kecelakaan lalu lintas berdasarkan ruas-ruas jalan yang memiliki kesamaan karakteristik dan visualisasi hasil clustering dalam bentuk peta dua dimensi. Dalam penelitian ini masih terbatas pada penggunaan data temporal dalam periode bulanan. Sedangkan dalam penentuan parameter klasterisasi yang digunakan dalam penelitian ini masih bersumber pada deskripsi kejadian dan belum melibatkan data geometri (kondisi fisik) jalan raya.

Penelitian kedua dilakukan oleh Silvi Agustina, Dhimas Yhudo, Hadi Santoso, Nofiandi Marnasusanto, Arif Tirtana, Fakhris Khusnu tentang Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means. Dalam penelitian ini penulis menggunakan metode k-means dan manhattan distance sebagai distance measure (perhitungan jarak). Penelitian ini menggunakan 20 data uji, dimana ke-20 data tersebut dibagi menjadi 3 *cluster* dengan *cluster* 1 merupakan beras kualitas buruk, *cluster* 2 beras kualitas sedang, dan *cluster* 3 beras kualitas baik. Dari hasil penelitian, didapatkan 3 pusat *cluster* akhir yaitu pusat *cluster* 1 (5,89333;2,05), pusat *cluster* 2 (6,28199;2,546), dan pusat *cluster* 3 (6,96583;2,999167) serta dihasilkan validasi sebesar 92,8% yang menunjukkan bahwa program ini dapat dijadikan sebagai acuan dalam klasterisasi kualitas beras.

Selanjutnya penelitian dilakukan oleh Erga Aprina Sari meneliti tentang Penerapan Algoritma K-Means Untuk Menentukan Tingkat Kesehatan Bayi dan Balita Pada Kabupaten dan Kota di Jawa Tengah. Penelitian ini menggunakan teknik *data mining* dengan algoritma K-Means untuk mengelompokkan atau mengklaster kabupaten-kabupaten yang ada di Provinsi Jawa Tengah berdasarkan kemiripan karakteristik daerah yang ditinjau dari nilai indikator kesehatan yaitu angka kematian bayi dan balita, angka kesakitan bayi dan balita, dan status gizi bayi dan balita.

Dari hasil uji coba didapat kabupaten/kota yang memiliki hasil analisa indikator kesehatan tinggi yang tingkat kesehatan pada indikator tersebut buruk karena jumlah penderitanya banyak. Penanganan masalah dapat difokuskan pada kabupaten/kota dengan indikator kesehatan tinggi

Penelitian keempat dilakukan oleh Firlir Irhamni, Fitri Damayanti, Bain Khusnul K, Miftachul A tentang Optimalisasi Pengelompokan Kecamatan Berdasarkan Indikator Pendidikan Menggunakan Metode Clustering dan Davies Bouldin Index. Penelitian ini tentang pengelompokan kecamatan untuk pemerataan pendidikan menggunakan indikator pendidikan yang terdapat pada suatu kecamatan sebagai salah satu organisasi pemerintah. Parameter penyebab keberhasilan pendidikan dapat dilihat dari indikator pendidikan di suatu daerah, salah satu tolak ukurnya adalah rendahnya nilai Angka Partisipasi Murni (APM) dan nilai Angka Partisipasi Kasar (APK). Indikator lain yang mempengaruhi pemerataan pendidikan adalah sarana dan prasarana pendidikan yang meliputi jumlah sekolah, ruang kelas, dan tenaga pengajar. Pengelompokan kecamatan berdasarkan tingkat pendidikan SMA/SMK/MA tersebut menggunakan metode clustering yaitu Self Organizing Map (SOM) dan hasil clustering tersebut kemudian diolah dengan metode Davies Bouldin Index (DBI) untuk menunjukkan seberapa baik cluster yang diperoleh.

Penelitian ini memberikan kontribusi terhadap pengambilan kebijakan dari pihak berwenang

Penelitian kelima dilakukan oleh Mario Anggara, Henry Sujiani, Helfi Nasution melaporkan bahwa penelitiannya mendapatkan hasil pengujian terhadap pengelompokan member di alvaro fitness menggunakan *k-means*

clustering dengan menggunakan 3 macam *distance measure*. Didapatkan bahwa dalam pengujian silhouette coefficient clustering, *Chebyshev Distance* memiliki nilai silhouette coefficient-nya sebesar 0.242821. Sedangkan *Euclidean Distance* dan *Manhattan Distance* memiliki nilai silhouette coefficient sebesar 0.232149 dan 0.240016.

2.2 Kecelakaan Lalu Lintas

Kecelakaan Lalu Lintas adalah suatu peristiwa di Jalan yang tidak diduga dan tidak disengaja melibatkan Kendaraan dengan atau tanpa Pengguna Jalan lain yang mengakibatkan korban manusia dan/atau kerugian harta benda (UU nomor 22 Tahun 2009).

2.3 Penentuan Lokasi Rawan Kecelakaan Lalu Lintas

Suatu tempat dikatakan “daerah” atau “lokasi” apabila diketahui letak dan batas-batasnya. Antara Direktorat Keselamatan Transportasi Darat dengan Departemen Pemukiman dan Prasarana Wilayah terdapat perbedaan dalam penyebutan tempat yang tergolong rawan kecelakaan lalu lintas. Direktorat Keselamatan Transportasi Darat menyebutnya dengan “daerah rawan kecelakaan”, sedangkan Departemen Pemukiman dan Prasarana Wilayah menyebutnya dengan “lokasi rawan kecelakaan”.

Daerah yang memiliki angka kecelakaan tinggi, resiko kecelakaan tinggi serta potensi kecelakaan tinggi pada suatu ruas jalan dapat disebut juga dengan daerah rawan kecelakaan. (Bayu Pramadya K.S, 2012)

Departemen Permukiman dan Prasarana Wilayah (2004) menjelaskan suatu lokasi dinyatakan sebagai lokasi rawan kecelakaan lalu lintas apabila:

1. Memiliki angka kecelakaan yang tinggi.
2. Lokasi kejadian kecelakaan relatif bertumpuk.
3. Lokasi kecelakaan berupa persimpangan, atau segmen ruas jalan sepanjang 100 – 300 m untuk jalan perkotaan, atau segmen ruas jalan sepanjang 1 km untuk jalan antar kota.
4. Kecelakaan terjadi dalam ruang dan rentan waktu yang relatif sama.
5. Memiliki penyebab kecelakaan dengan faktor yang spesifik.

2.4 Pengertian Jalan

Jalan adalah seluruh bagian jalan, termasuk bangunan pelengkap dan perlengkapannya yang diperuntukan bagi lalu lintas umum, yang berada pada permukaan tanah, diatas permukaan tanah, dibawah permukaan tanah dan/atau air, serta diatas permukaan air, kecuali jalan rel dan jalan kabel (UU nomor 22 tahun 2009).

Jalan dibagi kedalam kelas – kelas bukan hanya didasarkan pada fungsinya tetapi juga dipertimbangkan pada besarnya volume serta sifat lalu lintas. Adapun klasifikasi jalan dijelaskan dalam table dibawah ini (PP nomor 30 tahun 2021).

Tabel 2.1. Klasifikasi Kelas Jalan (PP nomor 30 tahun 2021).

Klasifikasi	Keterangan
Kelas I	<ul style="list-style-type: none">a. ukuran lebar tidak melebihi 2.550 (dua ribu lima ratus lima puluh) milimeter;b. ukuran panjang tidak melebihi 18.000 (delapan belas ribu) milimeter;c. ukuran tinggi tidak melebihi 4.200 (empat ribu dua ratus) milimeter; dand. d. ukuran muatan sumbu terberat 10 (sepuluh) ton.
Kelas II	<ul style="list-style-type: none">a. ukuran lebar tidak melebihi 2.550 (dua ribu lima ratus lima puluh) milimeter;b. ukuran panjang tidak melebihi 12.000 (dua belas ribu) milimeter;c. c. ukuran tinggi tidak melebihi 4.200 (empat ribu dua ratus) milimeter; dan

	d. d. ukuran muatan sumbu terberat 8 (delapan) ton.
Kelas III	<ul style="list-style-type: none"> a. ukuran lebar tidak melebihi 2.200 (dua ribu dua ratus) milimeter; b. ukuran panjang tidak melebihi 9.000 (sembilan ribu) milimeter; c. ukuran tinggi tidak melebihi 3.500 (tiga ribu lima ratus) milimeter; dan d. ukuran muatan sumbu terberat 8 (delapan) ton.

2.5 Data Mining

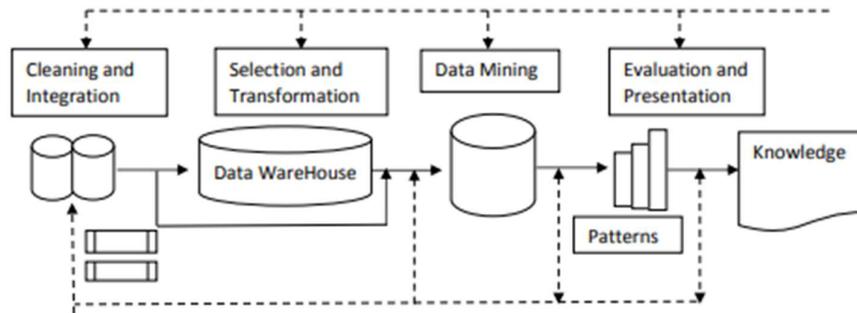
Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahamidan bermanfaat bagi pemilik data (Larose, 2006).

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Description (Deskripsi)
Peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.
2. Estimation (Estimasi)
Estimasi hampir sama dengan klasifikasi, dimana variabel tujuan yang lebih kearah numerik daripada ke arah kategori.
3. Prediction (Prediksi)
Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang (sesuatu yang belum terjadi).
4. Association (Asosiasi)
Asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.
5. Clustering
Clustering merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.
6. Classification (Klasifikasi)
Penyusunan data menjadi beberapa kelompok yang ditentukan.

2.6 Tahap-Tahap Data mining

Rangkaian proses data mining dibagi menjadi beberapa tahap yang bersifat interaktif seperti pada gambar.



Gambar 2.1. Tahap-tahap Data Mining (Irwan Budiman, 2012)

1. *Data Selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Cleaning*

Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan ketik (*tipografi*).

3. *Transformation*

Pengubahan data ke dalam format yang sesuai untuk dapat diproses dalam *data mining*. Misal, beberapa metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima input data kategorikal, maka data berupa angka numerik yang berlanjut perlu dibagi menjadi beberapa interval.

4. *Data mining*

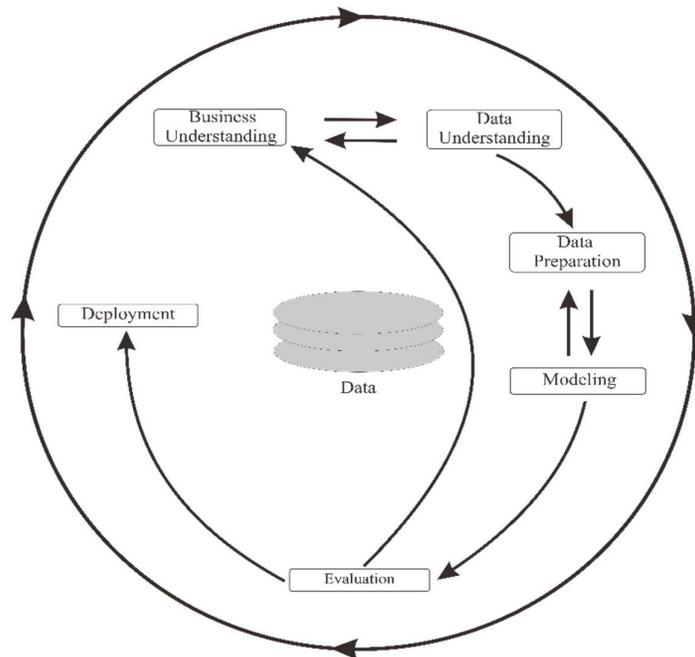
Proses pencarian pola atau informasi yang menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

5. *Pattern Evaluation*

Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya

2.7 CRISP-DM

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah salah satu model proses datamining (*datamining framework*) yang awalnya (1996) dibangun oleh 5 perusahaan yaitu *Integral Solutions Ltd (ISL)*, *Teradata*, *Daimler AG*, *NCR Corporation* dan *OHRA*. *Framework* ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan *methodology standard non-proprietary* bagi data mining. Versi pertama dari *methodologi* ini dipresentasikan pada 4th CRISP-DM SIG Workshop di Brussels pada bulan Maret 1999 (Pete Chapman, 1999); dan langkah langkah proses datamining berdasarkan model ini di publikasikan pada tahun berikutnya (Pete Chapman,2000).



Gambar 2. 2 Gambar siklus hidup CRISP-DM (Larose, Daniel T. 2005)

Berdasarkan CRISP-DM, proses data mining terdiri dari 6 fase yaitu:

1. Fase Pemahaman Bisnis (Business Understanding)

- a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
 - b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.
 - c. Menyiapkan strategi awal untuk mencapai tujuan.
2. Fase Pemahaman Data (Data Understanding Phase)
 - a. Mengumpulkan data.
 - b. menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
 - c. Mengevaluasi kualitas data.
 - d. Jika diinginkan, pilih sebagian kecil grup data yang mungkin mengandung pola dari permasalahan.
 3. Fase Pengolahan Data (Data Preparation Phase)
 - a. Siapkan data dari awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.
 - b. Pilih kasus dan variable yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
 - c. kan perubahan pada beberapa variabel jika dibutuhkan.
 - d. Siapkan data awal sehingga siap untuk perangkat pemodelan.
 4. Fase Pemodelan (Modelling Phase)
 - a. Pilih dan aplikasikan teknik pemodelan yang sesuai.
 - b. Kalibrasi aturan model untuk mengoptimalkan hasil.
 - c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk
 - d. digunakan pada permasalahan data mining yang sama.
 - e. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.
 5. Fase Evaluasi (Evaluation Phase)

- a. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
 - b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
 - c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
 - d. Mengambil keputusan berkaitan dengan penggunaan hasil dari data mining.
6. Fase Penyebaran (Deployment Phase)
- a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikan proyek.
 - b. Contoh sederhana penyebaran: Pembuatan laporan.
 - c. Contoh kompleks penyebaran: Penerapan proses data mining secara paralel pada departemen lain.

2.8 Clustering

Clustering merupakan salah satu teknik data mining yang digunakan untuk mendapatkan kelompok-kelompok dari obyek-obyek yang mempunyai karakteristik yang umum di data yang cukup besar. Pengelompokan sejumlah data/obyek ke dalam cluster merupakan tujuan utama dari metode clustering sehingga dalam setiap cluster akan diisi data yang memiliki tingkat kemiripan yang tinggi. (Santosa, Budi. 2007)

Sebuah cluster adalah kumpulan data yang memiliki kemiripan karakteristik satu sama lain dan tidak memiliki kemiripan dengan cluster lain. Clustering bersifat unsupervised learning karena pengelompokan data yang didasarkan pada kesamaan antar objek.

2.9 Algoritma K-Means

Beberapa teknik clustering yang paling sederhana diantara teknik lainnya adalah Klastering K-Means. K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Terdapat dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu Hierarchical dan Non-Hierarchical, dan K-Means merupakan salah satu metode data clustering non-hierarchical atau Partitional Clustering. (informatikalogi, 2016)

Algoritma *K-means* secara umum memiliki tahapan sebagai berikut:

1. Tentukan jumlah kelompok.
2. Inisialisasi titik centroid k (pusat cluster) secara acak.
3. Hitung jarak setiap titik pusat *cluster*, jarak antar satu data dengan satu *cluster* akan menentukan data tersebut masuk dalam cluster mana. Perhitungan jarak yang digunakan adalah sebagai berikut :

Euclidean Distance

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 \dots + (X_{ki} - X_{kj})^2}$$

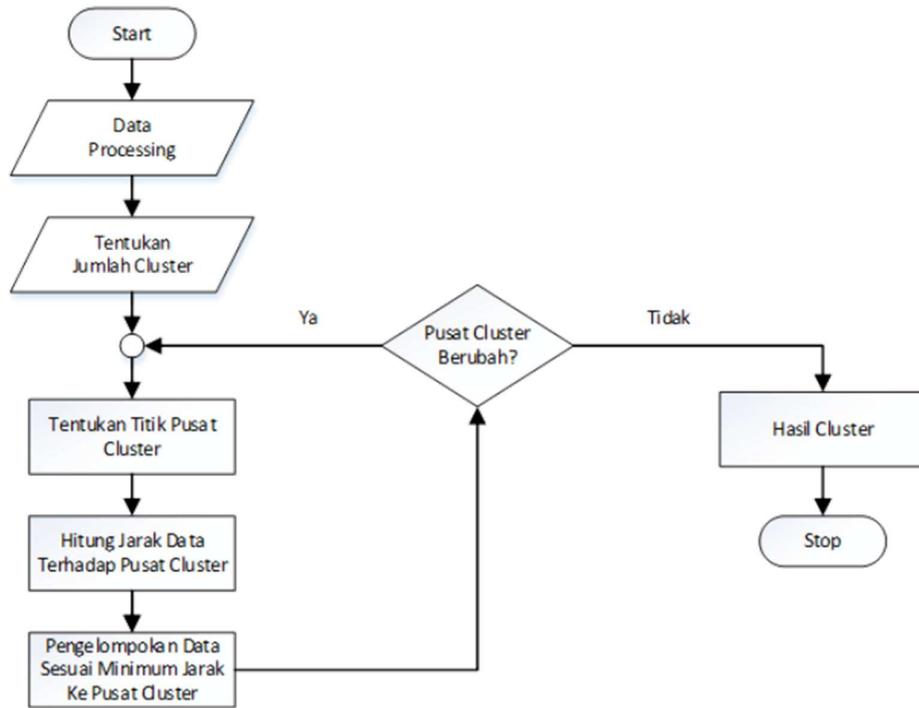
Dimana :

$D_{(i,j)}$ = Jarak data ke i ke pusat *cluster* j

$X_{(ki)}$ = Data ke i pada atribut data ke k

$X_{(kj)}$ = Titik pusat ke j pada atribut ke k

4. Hitung pusat *cluster* yang baru dengan keanggotaan *cluster* yang sekarang, dengan cara mencari rata-rata (mean) dari semua objek ataadata dalam *cluster* tertentu atau dengan menggunakan median dari *cluster* tersebut.
5. Ulangi mulai dari langkah ketiga sampai nilai pusat *cluster* tidak berubah.



Gambar 2.3 Flowchart Algoritma K-Means

2.10 *Hypertext Preprocessor (PHP)*

Menurut Supono & Putratama (2018: 1) mengemukakan bahwa “PHP (PHP: hypertext preprocessor) adalah suatu bahasa pemrograman yang digunakan untuk menterjemahkan basis kode program menjadi kode mesin yang dapat dimengerti oleh komputer yang bersifat server-side yang ditambahkan ke HTML”.