BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Sebelumnya sudah ada beberapa penelitian yang telah dilakukan berkaitan dengan analisis pengelompokkan data penduduk miskin menggunakan algoritma *k-means*. Penelitian sebelumnya dilakukan oleh (Margolang dkk., 2019) dengan judul "Implementasi *Data Mining* dalam Mengelompokkan Rumah Tangga Kumuh di Perkotaan Berdasarkan Provinsi Menggunakan Algoritma *K-Means*". Penelitian ini membahas tentang penerapan *data mining* dalam pengelompokkan data jumlah rumah tangga kumuh tingkat provinsi di Indonesia menggunakan algoritma *k-means*. Penggunaan metode ini, data-data yang telah diperoleh dapat dikelompokkan ke dalam beberapa *cluster*. Data diolah dan dibagi dalam 3 *cluster* yaitu, *cluster* tinggi (C1), *cluster* sedang (C2), dan *cluster* rendah (C3). Proses iterasi penelitian ini terjadi sebanyak 4 kali iterasi, sehingga diperoleh penilaian dalam mengelompokkan rumah tangga kumuh berdasarkan provinsi. Hasil yang diperoleh bahwa terdapat 2 provinsi dengan *cluster* tertinggi (C1), 16 provinsi dengan *cluster* sedang (C2), dan 16 provinsi dengan *cluster* terendah (C3). Data ini dapat menjadi masukkan kepada pemerintah setempat agar dapat menangani penyebarluasan bantuan pembangunan untuk daerah rumah tangga kumuh.

Penelitian selanjutnya dilakukan oleh (Nasution dkk., 2020) dengan judul "Penerapan Algoritma *K-Means* dalam Pengelompokan Data Penduduk Miskin Menurut Provinsi". Penelitan ini membahas tentang penerapan algoritma *k-means* dalam pengelompokan data penduduk miskin berdasarkan provinsi di Indonesia. Penelitian ini bertujuan untuk membantu pemerintah dalam mengambil kebijakan yang tepat dan efektif dalam menanggulangi kemiskinan di Indonesia. Dalam penelitian ini, data yang digunakan bersumber dari Badan Pusat Statistik tahun 2012-2018. Jumlah *record* yang digunakan sebanyak 34 provinsi dengan menghasilkan 2 *cluster* yakni *cluster* tinggi sebanyak 8 provinsi dan *cluster* rendah sebanyak 26 provinsi. Berdasarkan hasil pengujian *k-means,* untuk kasus persentase data penduduk miskin menggunakan *tools Rapid Miner,* diperoleh hasil yang sama dengan analisis perhitungan algoritma, dimana diperoleh 8 provinsi dengan *cluster* tinggi yang menjadi pusat perhatian bagi pemerintah dalam melakukan sosialisasi dan pemetaan dalam pemberian bantuan pada provinsi tersebut.

Penelitian selanjutnya dilakukan oleh (Massie & Padilah, 2021) dengan judul " Klasterisasi Angka Usia Muda Melek TIK Berdasarkan Algoritma *K-Means* Menurut Jumlah Provinsi Indonesia". Penelitian ini membahas tentang klasterisasi angka usia muda yang melek TIK (Teknologi Informasi dan Komunikasi) di berbagai provinsi Indonesia dengan menggunakan algoritma *k-means*. *Clustering* dilakukan hanya untuk mengelompokan provinsi menjadi 2 jenis kelompok, yang nantinya dapat dijadikan bahan evaluasi terhadap pemerintah dalam rangka pemerataan TIK disetiap provinsinya. Hasil akhir

dari penelitian ini adalah terdapat 25 provinsi yang termasuk ke dalam kelompok *cluster* 1 dan 9 provinsi yang termasuk ke dalam kelompok *cluster* 2. Dengan demikian perlu adanya usaha peningkatan angka melek TIK untuk *cluster* provinsi yang nilainya masih tertinggal dengan berbagai provinsi lainnya.

Penelitian selanjutnya dilakukan oleh (Ananda dkk., 2022) dengan judul "Implementasi *K-Means* dalam Pengelompokkan Data Akta Kelahiran di Indonesia". Penelitian ini membahas tentang penggunaan algoritma *k-means* dalam mengelompokkan data kepemilikan akta kelahiran di Indonesia sehingga dapat sebanding dengan jumlah penduduk di Indonesia. Penelitian ini menggunakan teknik *clustering* dan *Davies Bouldin Index* (DBI) untuk menguji validitas hasil klasterisasi. Hasil dari penelitian ini yaitu, klasterisasi terbaik diperoleh dengan menggunakan algoritma *k-means* dengan nilai k=4. Hasil klasterisasi telah diuji validitasnya menggunakan teknik *Davies Bouldin Index* (DBI) dengan nilainya sebesar 0,059. Pola yang didapatkan dari hasil klasterisasi dapat dijadikan sebagai acuan bagi Dinas Kependudukan dan Pencatatan Sipil (Disdukcapil) dalam melakukan pemetaan data akta kelahiran di Indonesia.

Penelitian selanjutnya dilakukan oleh (Deviana & Putro, 2023) dengan judul "Pengelompokkan Wilayah di Indonesia Berdasarkan Indikator Kerawanan Ekonomi Pasca Pandemi *Covid-19* (*K-Means Cluster Algorithm*)". Penelitian ini membahas tentang penerapan algoritma *k-means* untuk mengelompokkan provinsi-provinsi di Indonesia ke dalam 4 kelompok

berdasarkan tingkat kerawanan ekonomi pasca pandemi. Tujuannya adalah memberikan informasi dan pertimbangan kepada pemerintah dalam merancang kebijakan ekonomi yang berlandaskan pada karakteristik wilayah-wilayah yang memiliki kerawanan ekonomi serupa. Hasil penelitiannya yaitu pengelompokan provinsi-provinsi di Indonesia menjadi 4 kelompok kerawanan ekonomi berdasarkan analisis *k-means cluster algorithm*, hasilnya yaitu:

1. *Cluster* ke-1, merupakan kelompok provinsi yang digambarkan dengan warna hijau. Kelompok ini memiliki karakteristik TPAK (Tingkat Partisipasi Angkatan Kerja) dan TPAK lansia rendah dan memiliki tingkat pengangguran tertinggi dibandingkan kelompok lain.
2. *Cluster* ke-2, dengan warna kuning, memiliki ciri persentase penduduk miskin di bawah rata-rata. TPAK dan TPAK lansia pada kelompok ini lebih rendah dari rata-rata dan sejalan dengan itu, tingkat pengangguran pada kelompok ini berada di atas rata-rata.
3. *Cluster* ke-3, disimbolkan dengan warna oranye, memiliki karakteristik PDRB (Produk Domestik Regional Bruto) perkapita atas dasar harga berlaku yang lebih rendah dari rata-rata dan persentase penduduk miskin di atas rata-rata. TPAK dan TPAK lansia pada kelompok ini tinggi dan tingkat pengangguran rendah sehingga dapat mengurangi angka ketergantungan.
4. *Cluster* ke-4, merupakan kelompok provinsi dengan kerawanan ekonomi tertinggi yang disimbolkan dengan warna merah. Kelompok ini memiliki ciri PDRB perkapita atas dasar harga berlaku yang paling rendah diantara

wilayah lain. Kelompok ini memiliki TPAK dan TPAK lansia yang tinggi serta tingkat pengangguran yang rendah.

Perbandingan antara setiap penelitian dapat dilihat pada Tabel 2.1.

Tabel 2.1 Perbandingan Penelitian Terdahulu

No	Nama	Judul Penelitian	Metode	Hasil
1	(Margolang dkk., 2019)	Implementasi *Data Mining* dalam Mengelompokkan Rumah Tangga Kumuh di Perkotaan Berdasarkan Provinsi Menggunakan Algoritma *K-Means*	*K-Means Clustering*	Hasil yang diperoleh yaitu terdapat 2 provinsi dengan *cluster* tertinggi (C1), 16 provinsi dengan *cluster* sedang (C2), dan 16 provinsi dengan *cluster* terendah (C3).
2	(Nasution dkk., 2020)	Penerapan Algoritma *K-Means* dalam Pengelompokan Data Penduduk Miskin Menurut Provinsi	*K-Means Clustering*	Hasil yang didapatkan yaitu menghasilkan 2 *cluster* yakni *cluster* tinggi sebanyak 8 provinsi dan *cluster* rendah sebanyak 26 provinsi. Diperoleh 8 provinsi dengan *cluster* tinggi yang menjadi pusat perhatian bagi pemerintah dalam melakukan pemberian bantuan.

3	(Massie & Padilah, 2021)	Klasterisasi Angka Usia Muda Melek TIK Berdasarkan Algoritma *K-Means* Menurut Jumlah Provinsi Indonesia	*K-Means Clustering*	Hasil akhir dari penelitian ini adalah terdapat dua puluh lima provinsi yang termasuk ke dalam kelompok *cluster* 1 dan sembilan provinsi yang termasuk ke dalam kelompok *cluster* 2.
4	(Ananda dkk., 2022)	Implementasi *K-Means* dalam Pengelompokkandata Akta Kelahiran di Indonesia	*K-Means Clustering*	Hasil dari penelitian ini yaitu, klasterisasi terbaik diperoleh dengan menggunakan algoritma *k-means* dengan nilai k=4. Hasil klasterisasi telah diuji validitasnya menggunakan teknik *Davies Bouldin Index* (DBI) dengan nilai sebesar 0,059.
5	(Deviana & Subuh, 2023)	Pengelompokkan Wilayah di Indonesia Berdasarkan Indikator Kerawanan Ekonomi Pasca Pandemi *Covid-19* (*K-Means Cluster Algorithm*)	*K-Means Clustering*	Hasilnya yaitu pengelompokan provinsi di Indonesia menjadi 4 kelompok berdasarkan analisis *k-means cluster algorithm*.

Pada penelitian yang akan dilakukan mengenai klasterisasi data kemiskinan menurut kabupaten/kota di Provinsi NTT menggunakan algoritma *k-means*, akan mengacu pada penelitian kedua (Nasution dkk., 2020) dengan judul "Penerapan Algoritma *K-Means* dalam Pengelompokan Data Penduduk Miskin Menurut Provinsi". Penelitian yang dilakukan yaitu, mengelompokkan provinsi di seluruh Indonesia berdasarkan jumlah penduduk miskin yang serupa menggunakan algoritma *k-means*. Data yang digunakan pada penelitian tersebut dari tahun 2012 sampai tahun 2018 yang bersumber dari Badan Pusat Statistik.

Penelitian yang akan dilakukan memiliki beberapa perbedaan Perbedaannya yaitu, penelitian sebelumnya fokus pada pengelompokan provinsi di seluruh Indonesia menggunakan 1 variabel data yaitu, data jumlah penduduk miskin setiap provinsi dari tahun 2012 sampai 2018. Sedangkan penelitian yang akan dilakukan, berfokus pada pengelompokkan tingkat kabupaten/kota di Provinsi NTT menggunakan 4 variabel data yaitu, data jumlah penduduk miskin, rata-rata lamanya sekolah rata-rata pengeluaran perkapita sebulan makanan dan *non* makanan, dan persentase penduduk yang sulit mengakses layanan kesehatan. Data yang digunakan dari tahun 2017 sampai tahun 2023.

2.2 Landasan Teori

Penelitian yang dilakukan berdasarkan pada beberapa penelitian lain yang telah dilakukan sebelumnya, pada poin ini akan dibahas secara singkat teori-teori penunjang sebagai berikut:

2.2.1 *Data Mining*

Pengertian *data mining* mencakup penggunaan teknik analisis statistik untuk menggali pengetahuan yang tersembunyi dalam volume data besar, ini merupakan alat yang memungkinkan pengguna untuk mengakses dan mengolah data yang luas, yang sebelumnya tidak dapat diperoleh secara langsung. Dalam era, saat data tersedia dalam jumlah yang besar, *data mining* menjadi penting dalam mengubah data mentah menjadi informasi berharga dan pemahaman yang mendalam.

Secara umum, *data mining* dapat didefinisikan dengan menggabungkan dua kata kunci, yaitu "*data*", yang merujuk pada sekumpulan fakta atau entitas yang tercatat, dan "*mining*" yang merujuk pada proses penambangan. Oleh karena itu, data *mining* dapat dipahami sebagai proses pengeksplorasian data untuk menghasilkan *output* berupa pengetahuan yang berguna (Riadi & Mesran, 2023).

2.2.2 *Clustering*

Clustering atau klasifikasi merupakan pendekatan yang digunakan untuk mengorganisir kumpulan data ke dalam beberapa kelompok berdasarkan kesamaannya. "Kelompok" dalam konteks ini adalah sekelompok objek data yang menunjukkan kesamaan satu sama lain dan berbeda dari objek-objek yang ada di kelompok yang berbeda. Objek data akan ditempatkan dalam satu atau lebih kelompok sehingga objek yang berada dalam kelompok yang sama memiliki kesamaan yang signifikan satu sama lain (Rahma, 2020).

2.3 K-Means Clustering

K-Means adalah suatu teknik pengelompokan data *non* hirarki yang bertujuan untuk memisahkan data menjadi *cluster* atau kelompok-kelompok. Dalam metode ini, data yang memiliki kesamaan karakteristik dikelompokkan ke dalam *cluster* yang sama, sementara data dengan karakteristik yang berbeda ditempatkan dalam *cluster* yang berbeda .(Margolang dkk., 2019).

Ada beberapa langkah dalam klasterisasi data menggunakan *algoritma k-means*:

1. Tentukan jumlah *cluster (k)* dalam *dataset*.
2. Kemudian, nilai *centroid* ditentukan pada tahap awal secara acak, sementara pada tahap iterasi, digunakan Persamaan 2.1.

   $V_{ij} = \frac{1}{Ni} \sum_{k=0}^{Ni} Xk$ .....(2.1)

   Keterangan:

   V_{ij} = rata-rata pusat kelompok pertama (*cluster* ke-i) untuk variabel ke-j

   Ni = jumlah anggota dalam kelompok pertama (*cluster* ke-i)

   k = indeks yang merujuk ke *cluster*

   j = indeks yang merujuk ke variabel

   Xkj = nilai variabel ke-j pada data ke-k dalam kelompok tersebut
3. Tentukan jarak setiap data ke pusat *cluster*. Untuk mengukur jarak antara setiap data dengan *centroid* digunakan rumus *euclidean*

distance (D) sebagaimana dijelaskan dalam Persamaan 2.2 berikut ini (Solichin & Khairunnisa, 2020):

$$D_{(i,j)} = \sqrt{\sum_{k=1}^{n}\left(X_{ik} - C_{jk}\right)^2} \quad \text{.....(2.2)}$$

Keterangan:

D = jarak *cluster*

X_{ik} = nilai data (i,k)

C_{jk} = nilai *centroid* (j,k)

n = jumlah *cluster*

4. Pengelompokan data berdasarkan *cluster* terdekat atau kelompokkan data berdasarkan jarak terdekat ke *centroid*. Mengamati *cluster* yang memiliki jarak terdekat dengan data, lalu masukkan data ke dalam *cluster* yang sesuai.
5. Bandingkan *cluster* baru dengan *cluster* sebelumnya. Apabila ada perubahan dalam *cluster* baru atau pengelompokkan data yang berbeda dengan *cluster* sebelumnya, maka proses akan diulang kembali ke tahap 2. Jika *cluster* atau pengelopokkan data telah sama dengan *cluster* sebelumnya, proses dapat dihentikan dan hasil pengelompokan akhir telah diperoleh.

2.3.1 *Rapid Miner*

Rapid Miner adalah alat pemrograman sumber terbuka yang dapat digunakan untuk menganalisis data dan teks dengan berbagai cara. Dengan metode yang informatif, *Rapid Miner* membantu pengguna

mendapatkan wawasan untuk membuat keputusan yang optimal. *Suite Rapid Miner* mencakup sekitar 500 operator yang dapat digunakan untuk berbagai tahap dalam proses penambangan informasi (*information mining*), mulai dari *input* hingga *output*, *information pra-peocessing*, dan representasi. *Suite Rapid Miner* adalah kumpulan alat atau modul yang terintegrasi dalam perangkat lunak *Rapid Miner* yang digunakan untuk menganalisis data, dikembangkan menggunakan bahasa *java.*

Rapid Miner dinaungi oleh Lisensi Publik Affero GNU (AGPL). *Rapid Miner* merupakan perangkat lunak sumber terbuka untuk penambangan data atau *data mining* (Ananda dkk., 2023).