

BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Penelitian mengenai implementasi dari Algoritma K-Means Clustering terdahulu guna sebagai referensi dan bahan pembandingan penelitian yang saat ini dapat dilakukan, diantaranya sebagai berikut:

Penelitian pertama dilakukan oleh Asroni, Hidayatul Fitri dan Eko Prasetyo (2018) dengan judul penelitian Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). Hasil penelitian yang dilakukan dapat diambil kesimpulan Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa Pendidikan Dokter dan Ilmu Hubungan Internasional menjadi jurusan pilihan calon mahasiswa baru Universitas Muhammadiyah Yogyakarta. Hasil penelitian ini menjadi acuan pihak universitas untuk melakukan strategi promosi ke calon mahasiswa baru.

Penelitian kedua dilakukan oleh Suhartini dan Ria Yuliani (2021) dengan judul penelitian Penerapan Data Mining untuk Mengcluster Data Penduduk Miskin Menggunakan Algoritma KMeans di Dusun Bagik Endep Sukamulia Timur. Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan bahwa menentukan centroid (titik pusat) pada tahap awal algoritma K-Means sangat berpengaruh pada hasil cluster seperti pada hasil pengujian yang dilakukan dengan menggunakan 200 dataset dengan centroid yang berbeda menghasilkan hasil cluster yang berbeda pula. Berdasarkan 3 cluster yang telah dilakukan pengujian menggunakan Rapid Miner

bahwa Data Penduduk terbanyak terdapat pada cluster ke-3 yang tergolong pada penduduk ekonomi rendah (Miskin).

Penelitian ketiga dilakukan oleh Parjito dan Permata (2022) dengan judul penelitian Penerapan Data Mining untuk Clustering Data Penduduk Miskin Menggunakan Metode K-Means. Berdasarkan hasil penelitian menggunakan metode K-Means yang telah dilakukan maka didapat hasil berupa kategori masyarakat miskin dan tidak miskin yang dapat dipergunakan sebagai salah satu alat yang digunakan untuk mengambil keputusan dalam menentukan kelompok masyarakat yang mendapatkan bantuan dengan melihat variable Kepemilikan Aset. Hal ini terjadi karena dalam pembahasan yang dilakukan nilai variabel kepemilikan aset mempunyai pengaruh pembuatan cluster terbaik. Berdasarkan hasil pengujian data training menggunakan Weka dapat disimpulkan dari 812 data training penduduk, 103 data digolongkan menjadi miskin dan 709 menjadi tidak miskin dengan iterasi sebanyak 4 kali. Persentase 13% pada kategori miskin dan 83% pada kategori tidak miskin dari hasil penerapan pada tools weka sedangkan pada validitas yang dilakukan menggunakan metode DBI dihasilkan jumlah cluster yang cocok adalah 2 cluster dengan nilai 0,1643 yang berarti cluster tersebut cocok untuk digunakan.

Penelitian keempat dilakukan oleh Yuni Radana Sembiring, Saifullah dan Riki Winanjaya (2021) dengan judul penelitian Implementasi Data Mining Dalam Mengelompokkan Jumlah Penduduk Miskin Berdasarkan Provinsi Menggunakan Algoritma K-Means. Dari hasil penelitian yang telah dilakukan, pengelompokkan jumlah penduduk miskin yang ada diprovinsi indonesia berhasil diterapkan menggunakan Algoritma K-Means Clustering. Pengujian yang dilakukan dalam penelitian ini, iterasi clustering pada data jumlah penduduk miskin terjadi sebanyak 3 kali iterasi. Dari Hasil Perhitungan k-means dengan Aplikasi Rapidminer hasil yang didapat bernilai sama yaitu Cluster Pertama dengan jumlah data penduduk miskin sebanyak 10 provinsi yaitu

Sumatera Utara, Sumatera Selatan, Lampung, Banten, NTB, NTT, Suawesi Selatan dan Papua. Cluster Kedua jumlah data penduduk miskin sebanyak 22 Provinsi yaitu, Sumatera Barat, Riau, Jambi, Bengkulu, Kep Bangka Belitung, Kep Riau, DKI Jakarta, Di Yogyakarta, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, dan Papua Barat. Cluster Ketiga jumlah data penduduk miskin sebanyak 3 Provinsi yaitu, Jawa Barat, Jawa Tengah dan Jawa Timur.

Table 2. 1 Perbandingan Penelitian

No	Peneliti	Judul penelitian	Metode	Keterangan
1	Asroni, Hidayatul Fitri dan Eko Prasetyo (2018)	Metode Clustering dengan Algoritma K-Means pada Pengelompokan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik).	K-Means Clustering	Hasil penelitian yang dilakukan dapat diambil kesimpulan Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa Pendidikan Dokter dan Ilmu Hubungan Internasional menjadi jurusan pilihan calon mahasiswa baru Universitas Muhammadiyah Yogyakarta. Hasil penelitian ini menjadi acuan pihak universitas untuk melakukan strategi promosi ke calon mahasiswa baru.
2	Suhartini dan Ria Yuliani (2021)	Penerapan Data Mining untuk Mengcluster Data Penduduk Miskin Menggunakan Algoritma	K-Means Clustering	Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan bahwa menentukan centroid (titik pusat) pada tahap awal algoritma K-Means sangat

		KMeans di Dusun Bagik Endep Sukamulia Timur.		berpengaruh pada hasil cluster seperti pada hasil pengujian yang dilakukan dengan menggunakan 200 dataset dengan centroid yang berbeda menghasilkan hasil cluster yang berbeda pula. Berdasarkan 3 cluster yang telah dilakukan pengujian menggunakan Rapid Miner bahwa Data Penduduk terbanyak terdapat pada cluster ke-3 yang tergolong pada penduduk ekonomi rendah (Miskin).
3	Parjito dan Permata (2022)	Sistem Penerapan Data Mining untuk Clustering Data Penduduk Miskin Menggunakan Metode K-Means.	K-Means Clustering	Berdasarkan hasil penelitian menggunakan metode K-Means yang telah dilakukan maka didapat hasil berupa kategori masyarakat miskin dan tidak miskin yang dapat dipergunakan sebagai salah satu alat yang digunakan untuk mengambil keputusan dalam menentukan kelompok masyarakat yang mendapatkan bantuan dengan melihat variable Kepemilikan Aset. Berdasarkan hasil pengujian data training menggunakan Weka dapat disimpulkan dari 812 data training penduduk, 103 data digolongkan menjadi miskin dan 709 menjadi tidak miskin dengan iterasi sebanyak 4 kali. Persentase 13% pada kategori miskin dan 83% pada kategori tidak miskin dari hasil penerapan pada tools weka sedangkan pada validitas yang

				dilakukan menggunakan metode DBI dihasilkan jumlah cluster yang cocok adalah 2 cluster dengan nilai 0,1643 yang berarti cluster tersebut cocok untuk digunakan.
4	Yuni Radana Sembiring, Saifullah dan Riki Winanjaya (2021)	Implementasi Data Mining Dalam Mengelompokkan Jumlah Penduduk Miskin Berdasarkan Provinsi Menggunakan Algoritma K-Means.	K-Means Clustering	Dari hasil penelitian yang telah dilakukan, pengelompokkan jumlah penduduk miskin yang ada diprovinsi indonesia berhasil diterapkan menggunakan Algoritma K-Means Clustering. Pengujian yang dilakukan dalam penelitian ini, iterasi clustering pada data jumlah penduduk miskin terjadi sebanyak 3 kali iterasi. Dari Hasil Perhitungan k-means dengan Aplikasi Rapidminer hasil yang didapat bernilai sama yaitu Cluster Pertama dengan jumlah data penduduk miskin sebanyak 10 provinsi yaitu Sumatera Utara, Sumatera Selatan, Lampung, Banten, NTB, NTT, Suawesi Selatan dan Papua. Cluster Kedua jumlah data penduduk miskin sebanyak 22 Provinsi yaitu, Sumatera Barat, Riau, Jambi, Bengkulu, Kep Bangka Belitung, Kep Riau, DKI Jakarta, Di Yogyakarta, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara,

				dan Papua Barat. Cluster Ketiga jumlah data penduduk miskin sebanyak 3 Provinsi yaitu, Jawa Barat, Jawa Tengah dan Jawa Timur.
--	--	--	--	--

2.2 Dasar Teori

2.2.1 Data Mining

Larose (2006), mengartikan data mining adalah sebuah proses menemukan sesuatu bermakna dengan memilah data melalui repository dengan bantuan teknologi sosialisasi pola, statistik, serta matematika.

Grup Gartner berpendapat (Larose, 2005: 2), data mining adalah sebuah proses menemukan korelasi baru, pola, dan tren bermakna pada penyimpanan data dengan jumlah besar, menggunakan teknologi pengenalan pola seperti statistic dan teknik matematika. Menurut Han, J., et al (2012: 8), data mining adalah proses menemukan pola dan pengetahuan dari data yang berjumlah besar.

Data mining adalah proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data. Dalam proses ini seringkali memanfaatkan beberapa metode, seperti matematika, statistika dan pemanfaatan teknologi artificial intelligence (AI). Jenis-jenis data dalam Data Mining adalah :

- a. Association.
- b. Clustering.
- c. Classification.

- d. Prediction.
- e. Regression.
- f. Anomaly detection.
- g. Sequential pattern.

2.2.2 Algoritma K-Means

Menurut Vulandari (2017:54) k-means merupakan algoritma yang menetapkan nilai-nilai cluster (k) secara random, untuk sementara nilai tersebut menjadi pusat dari cluster yang biasa disebut centroid. K-means merupakan salah satu algoritma yang bersifat unsupervised learning. K-Means memiliki fungsi untuk mengelompokkan data kedalam data cluster. Algoritma ini dapat menerima data tanpa ada label kategori. **K-Means Clustering** adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan unsupervised learning dan menggunakan metode yang mengelompokkan data berbagai partisi.

Algoritma k-means clustering melakukan dua tugas utama, yakni:

1. Menentukan nilai terbaik untuk titik pusat K atau centroid dengan proses iteratif (perulangan).
2. Menetapkan setiap titik data ke pusat k terdekat. Titik-titik data yang dekat dengan pusat- k tertentu, kemudian dibuatkan sebuah kluster

2.2.3 Clustering

Menurut Tan, 2006 *clustering* adalah sebuah proses untuk mengelompokkan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang

maksimum dan data antar *cluster* memiliki kemiripan yang minimum. *Clustering* merupakan salah satu metode data mining yang bersifat tanpa arahan (unsupervised).

Clustering adalah proses partisi satu set obyek data ke dalam himpunan bagian yang disebut dengan *cluster*. *Clustering* ini merupakan proses mengelompokkan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam suatu cluster memiliki tingkat kemiripan. Teknik ini digunakan untuk mengenali kelompok-kelompok yang dihasilkan dari pengelompokkan unsur-unsur yang lebih kecil berdasarkan adanya kemiripan satu sama lain.

2.2.4 Metode Elbow

Menurut Merliana, Ernawati, & Santoso (2015), Metode Elbow adalah untuk menghasilkan informasi dengan cara melihat perbandingan hasil antara jumlah cluster yang akan membentuk suatu titik terakhir, cluster nilai akan digunakan sebagai model data untuk cluster terbaik. Selain itu, hasil perhitungan akan digunakan untuk membandingkan jumlah cluster. Metode ini digunakan untuk menentukan jumlah cluster terbaik. Metode elbow sendiri merupakan metode yang umum digunakan karena dapat memaksimalkan kualitas hasil cluster. Berikut ini merupakan rumus SSE :

$$SSE = \sum_{k=1}^k \sum_{x_i \in S_k} |x_i - C_k|^2$$

Gambar 2.1 Rumus SSE

Keterangan :

: nilai atribut dari data ke-*i*

: nilai atribut titik pusat Cluster ke- i

Algoritma metode elbow melibatkan beberapa tahap, di antaranya sebagai berikut.

1. Menginisialisasi nilai awal untuk K.
2. Meningkatkan nilai K.
3. Menghitung Sum of Squared Error (SSE) untuk setiap nilai K.
4. Mengamati penurunan SSE yang signifikan dari nilai K.
5. Menentukan nilai K yang membentuk siku pada grafik.

2.2.5 Kemiskinan

Kemiskinan secara etimologis berasal dari kata “ miskin” yang artinya tidak berharta benda dan serba kekurangan. Departemen Sosial dan Biro Pusat Statistik mendefinisikan sebagai ketidakmampuan individu dalam memenuhi kebutuhan dasar minimal untuk hidup layak (BPS dan Depsos, 2002).

Secara harfiah Kamus Besar Bahasa Indonesia Tahun 2008, miskin itu berarti tidak berharta benda. Miskin juga berarti tidak mampu mengimbangi tingkat kebutuhan hidup standar dan tingkat penghasilan dan tingkat ekonomi rendah. Secara singkat kemiskinan dapat didefinisikan sebagai suatu standar tingkat hidup yang rendah yaitu adanya kekurangan materi pada sejumlah atau segolongan orang dibandingkan dengan standar kehidupan yang berlaku dalam masyarakat yang bersangkutan.

Kemiskinan dapat dicirikan keadaan dimana terjadi kekurangan hal-hal yang biasa dipunyai seperti makanan, pakaian, tempat berlindung dan air minum, hal-hal yang berkaitan erat dengan kualitas hidup. Kemiskinan kadang juga berarti tidak adanya akses terhadap pendidikan

dan pekerjaan yang mampu mengatasi masalah kemiskinan dan mendapatkan kehormatan yang layak sebagai warga negara (Perpes No 7 Tahun 2005 tentang RPJMN).

Untuk mengukur kemiskinan, BPS menggunakan konsep kemampuan memenuhi kebutuhan dasar (basic needs approach). Dengan pendekatan ini, kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran. Jadi Penduduk Miskin adalah penduduk yang memiliki rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan.

2.2.6 Garis Kemiskinan

Garis Kemiskinan (GK) merupakan penjumlahan dari Garis Kemiskinan Makanan (GKM) dan Garis Kemiskinan Non Makanan (GKNM). Penduduk yang memiliki rata-rata pengeluaran perkapita per bulan dibawah Garis Kemiskinan dikategorikan sebagai penduduk miskin. Garis Kemiskinan Makanan (GKM) merupakan nilai pengeluaran kebutuhan minimum makanan yang disetarakan dengan 2100 kilokalori perkapita perhari. Paket komoditi kebutuhan dasar makanan diwakili oleh 52 jenis komoditi (padi-padian, umbi-umbian, ikan, daging, telur dan susu, sayuran, kacang-kacangan, buah-buahan, minyak dan lemak, dll). Garis Kemiskinan Non Makanan (GKNM) adalah kebutuhan minimum untuk perumahan, sandang, pendidikan dan kesehatan. Paket komoditi kebutuhan dasar non makanan diwakili oleh 51 jenis komoditi di perkotaan dan 47 jenis komoditi di pedesaan.

Rumus perhitungan garis kemiskinan adalah

$$GK = GKM + GKNM$$

Keterangan

GK = Garis Kemiskinan

GKM = Garis Kemiskinan Makanan

GKNM = Garis Kemiskinan Non Makanan

$$GKM^*_{jp} = \sum_{k=1}^{52} P_{jkp} \cdot Q_{jkp} = \sum_{k=1}^{52} V_{jkp}$$

Dimana :

GKM^*_{jp} : Garis Kemiskinan Makanan daerah j (sebelum disetarakan menjadi 2100 kilokalori) provinsi p .

P_{jkp} : Rata-rata harga komoditi k di daerah j dan provinsi p .

Q_{jkp} : Rata-rata kuantitas komoditi k yang dikonsumsi di daerah j di provinsi p .

V_{jkp} : Nilai pengeluaran untuk konsumsi komoditi k di daerah j provinsi p .

j : Daerah (perkotaan atau pedesaan).

p : Provinsi ke- p .

Gambar 2.2 Rumus GKM

$$GKNM_{jp} = \sum_{k=1}^n r_{kj} V_{kjp}$$

$GKNM_{jp}$: Pengeluaran minimum non-makanan atau garis kemiskinan non-makanan daerah j dan provinsi p

V_{kjp} : Nilai pengeluaran per komoditi/sub-kelompok non-makanan daerah j dan provinsi p

r_{kj} : Rasio pengeluaran komoditi/sub-kelompok non-makanan k menurut daerah (hasil SPKKD 2004) dan daerah j

k : Jenis komoditi non-makanan terpilih

j : Daerah (Perkotaan atau pedesaan)

p : Provinsi ke- p

Gambar 2.3 Rumus GKNM

2.2.7 Persentase Penduduk Miskin

Persentase penduduk miskin merupakan penduduk yang memiliki rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan.

Head Count Index (HCI-P0), adalah persentase penduduk yang berada dibawah Garis Kemiskinan (GK). Dengan rumus perhitungan

$$P_{\alpha} = \frac{1}{n} \sum_{i=1}^q \left[\frac{z - y_i}{z} \right]^{\alpha}$$

Gambar 2.4 Rumus Persentase Penduduk Miskin

Dimana :

$\alpha = 0$

z = garis kemiskinan.

y_i = Rata-rata pengeluaran per kapita sebulan penduduk yang berada dibawah garis kemiskinan ($i=1, 2, 3, \dots, q$), $y_i < z$

q = Banyaknya penduduk yang berada di bawah garis kemiskinan.

n = jumlah penduduk.