BAB 2

TINJAUAN PUSTAKA

2.1 Analisis Klaster

Analisis cluster merupakan salah satu analisis data mining yang digunakan untuk mengelompokkan objek multidimensional, yaitu objek yang dapat digambarkan dengan sejumlah karakteristik. Salah satu tugas dalam data mining adalah proses *clustering*. Tujuan utama dari proses ini adalah untuk pengelompokan sejumlah data/obyek ke dalam sebuah cluster sehingga dalam setiap cluster akan berisi data yang semirip mungkin (Nurjanah, dkk, 2014). Pada analisis cluster memungkinkan untuk pengelompokkan pada data yang besar sehingga penggunaannya menjadi lebih praktis (Kucherov & Kurenkov, 2017). Pada analisis klaster terdapat beberapa teknik diantaranya hirarki, partisi, *density based* dan struktur grid (Jiang, Li, Min, Qi, & Rao, 2017). Ada dua jenis analisis *cluster* berbasis partisi yaitu *hard clustering* (tegas) dan *crisp clustering* (*fuzzy*) (Lin, Huang, Kuo, & Lai, 2014). Salah satu contoh *hard partition* yang sederhana adalah K-Means (Celebi, Kingravi, & Vela, 2013) dan *crisp partititon* adalah Fuzzy C-Means (Pimentel & de Souza, 2016).

2.2 Fuzzy C-Means

Logika *fuzzy* didasarkan pada himpunan *fuzzy* yang pertama kali diperkenalkan oleh Lotfi Zadeh (1965). *Fuzzy clustering* merupakan analisis pengelompokkan yang didasarkan pada himpunan *fuzzy*. Konsep dasar *fuzzy clustering* adalah bahwa sebuah objek dapat menjadi bagian lebih dari satu kelompok (Grekousis & Thomas, 2012). Pada *Fuzzy clustering* setiap objek memiliki derajat keanggotaan dalam semua kelompok. Derajat keanggotaan tersebut bersifat kontinyu yakni nilainya berada dalam rentang [0,1] (Saha & Das, 2017).

Fuzzy C-Means (FCM) pertama kali diperkenalkan oleh Dunn (1973) kemudian dikembangkan oleh Bezdek (1981). FCM merupakan metode yang banyak dikenal dan banyak digunakan dalam penelitian (Simhachalam & Ganesan, 2015). FCM menghubungkan derajat keanggotaan dan jarak suatu objek pada pusat kelompoknya. Suatu objek akan cenderung menjadi anggota suatu kelompok jika objek tersebut memiliki nilai derajat keanggotaan tertinggi (Hadi, 2017). *Fuzzy clustering* merupakan metode yang sering digunakan dalam GDA. *Fuzzy clustering* memberikan nilai keanggotaan pada masing-masing wilayah dan dapat membantu mengurangi kekeliruan ekologis. Algoritma *fuzzy clustering* yang digunakan dalam GDA adalah algoritma *Fuzzy C-Means* (Son, Lanzi, et al., 2013). Fungsi objektif dalam FCM didefiniskan sebagai berikut (Bezdek, Ehrlich, & Full, 1984):

$$J_m(\tilde{U}, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2 \qquad (2.1)$$

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m}\left(x_{kj} - v_{ij}\right)^{\frac{1}{2}}\right] \quad (2.2)$$

Dengan:

μ_{ik} : nilai keanggotaan dari ke-k pada kelompok ke-i,$0 \leq \mu_{ik} \geq 1$

d_{ik} : jarak dari titik data x_k ke pusat kelompok v_i

v_i : nilai pusat kelompok ke-i

x_k : titik data

n : jumlah objek penelitian

c : jumlah kelompok yang diinginkan

m : fuzziness,parameter yang digunakan untuk mengukur tingkat kesamaran dari hasil pengelompokkan, m > 1

2.3 Geo-Demographic Analysis (GDA)

Istilah geo-demografi cukup lama digunakan untuk mengacu pada pembangunan dan penerapan tipologi *small area* yang dapat digunakan untuk memahami tingkat variasi pola perilaku konsumen dan kondisi medis, permasalahan sosial dan karakteristik kehidupan lain yang diamati antara kelompok sosial ekonomi yang berbeda (Brown, Hirschfield, & Batey, 1991).

Geo-Demographic Analysis (GDA) merupakan perpaduan antara *Geographical Information System* (GIS) dan algoritma *data mining* (Son et al., 2012). GDA menggunakan teknik *clustering* untuk mengklasifikasikan data *geodemographic* menjadi beberapa kelompok sehingga mempermudah dalam proses analisis (Son, Lanzi, et al., 2013). Ada dua asumsi utama dalam GDA yaitu;

pertama, dua orang yang hidup di wilayah yang sama memiliki kesamaan karakteristik dibandingkan dua orang dari wilayah yang berbeda. Kedua adalah dua wilayah dapat dikategorikan berdasarkan populasi yang dimilikinya (Palmer, 2008).

2.4 Fuzzy Geographically Weighted Clustering (FGWC)

Fuzzy Geographically Weighted Clustering (FGWC) pertama kali diperkenalkan oleh (Mason & Jacobson, 2007). FGWC merupakan perbaikan dari algoritma FCM yang lebih peka terhadap geografis karena melibatkan efek populasi dan jarak dalam perhitungan derajat keanggotaan pada tiap observasinya (Hadi, 2017). Pengaruh wilayah satu terhadap wilayah lain dianggap sebagai hasil dari jumlah populasi dan jarak antar wilayah tersebut. Penentuan keanggotaan kelompok pada FGWC yang dihitung pada tiap iterasi ditunjukkan oleh rumus berikut (Mason & Jacobson, 2007).

$$\mu'_i = \alpha\mu_i + \beta\frac{1}{A}\sum_{j=1}^{n} w_{ij}\mu_j \qquad (2.3)$$

Dengan:

μ'_i : nilai keanggotaan baru dari objek-i

μ_i : nilai keanggotaan lama dari objek-i

w_{ij} : ukuran penimbang sejumlah interaksi antar wilayah

A : nilai untuk memastikan nilai penimbang interaksi antar wilayah agar tetap dalam range 0-1

α dan β merupakan faktor pengali untuk nilai keanggotaan yang lama dan nilai penimbang dari rerata keanggotaan unit observasi lain.

Nilai α dan β didefinisikan sebagai berikut

$$\alpha + \beta = 1$$

Penimbang keanggotaan w_{ij} didefinisikan sebagai berikut.

$$w_{ij} = \frac{(m_i m_j)^b}{{d_{ij}}^a} \quad (2.4)$$

Dengan:

m_i : jumlah populasi dari wilayah - i

m_j : jumlah populasi dari wilayah - j

d_{ij} : jarak antara wilayah – i dan wilayah - j

a dan b merupakan parameter yang ditentukan oleh peneliti. Jika pengaruh populasi dianggap sama pentingnya dengan pengaruh jarak, maka a=b=1. Analisis *Fuzzy geo-demographic* telah diteliti oleh Feng dan Flowerdew (Feng & Flowerdew, 1998) dengan menggabungkan *fuzzy clustering* dan *Neighbourhood Effects* untuk menghitung nilai derajat keanggotaanya. Penelitian tersebut menggabungkan efek ketetanggaan setelah proses *fuzzy clustering* yang hasilnya mempengaruhi nilai pusat klaster. Namun penelitian Feng dan Flowerdew tersebut masih memiliki keterbatasan yaitu metode tersebut mengabaikan efek kewilayahan yang tidak memiliki batas umum dan mengabaikan efek populasi yang merupakan kunci dari pertimbangan geografi (Mason & Jacobson, 2007). Untuk mengatasi keterbatasan tersebut maka disusunlah algoritma *Fuzzy Geographically Weighted Clustering* (FGWC).

2.5 Fungsi Objektif Fuzzy Geographically Weighted Clustering (FGWC)

Algoritma FGWC memiliki keunggulan karena sifatnya *geographic aware*. Namun memiliki keterbatasan pada tahap inisialisasi pusat klaster yang ditentukan secara acak dan jumlah kelompok geodemografis ditentukan manual oleh peneliti. Kedua hal tersebut dapat menyebabkan proses iterasi gagal mencapai solusi optimal. Untuk mengatasi keterbatasan tersebut maka algoritma GSA digunakan untuk memilih pusat klaster atau matriks keanggotaan pada fase inisialisasi FGWC. Fungsi objektif FGWC yang akan diminimumkan adalah (Wijayanto & Purwarianti, 2014):

$$J_{FGWC}(U,V;X) = \sum_{i=1}^{c}\sum_{k=1}^{n} {u_{ik}}^{m}|v_i - x_k|^2 \rightarrow \min \quad (2.5)$$

Dengan:

U : Matriks keanggotaan

V : Matriks untuk pusat klaster

X : Matriks untuk data

v_i : Pusat klaster untuk objek - i

u_i : Elemen dari matriks keanggotaan

x_k : Titik data

m : Fuzziness, parameter yang digunakan untuk mengukur tingkat kesamaran dari hasil pengelompokkan, m > 1 Pusat klaster didefinisikan sebagai berikut:

$$v_i = \frac{\sum_{k=1}^{n} {u_{ik}}^{m} x_k}{\sum_{k=1}^{n} {u_{ik}}^{m}} \quad (2.6)$$

Matriks keanggotaan dapat dihitung menggunakan rumus berikut ini:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|v_i - x_k\|}{\|v_j - x_k\|}\right)^{\frac{2}{m-1}}} \quad (2.7)$$

2.6 Gravitational Search Algorithm (GSA)

GSA merupakan salah satu algoritma yang dikembangkan oleh Rashedi (Rashedi, Nezamabadi-pour, & Saryazdi, 2009). Tujuan dari algoritma ini adalah untuk memperbaiki ekplorasi dan eksploitasi populasi berbasis algoritma untuk mencapai solusi optimal. GSA terinspirasi dari hukum gerakan dan gravitasi Newton. Dalam GSA, setiap objek disebut agen dan kemampuan masing-masing agen ditentukan oleh massanya. Tiap-tiap agen saling mempengaruhi karena hokum gravitasi. Agen dengan kemampuan yang kecil akan berpindah menuju agen yang memiliki kemampuan yang besar. Langkah awal dalam GSA adalah inisialisasi secara acak N solusi dan m dimensi. Posisi agen dirumuskan sebagai berikut:

$$X_i = (X_{i1}, \dots, X_{id}, \dots, X_{im})$$

Pada tiap iterasi, total *force* pada masing-masing agen (F) dirumuskan sebagai berikut:

$${F_{ij}}^d(t) = G(t)\frac{M_i(t)M_j(t)}{R_{ij}(t)}({x_i}^d(t) - {x_i}^d(t)) \quad (2.8)$$

$${F_{ij}}^d(t) = \sum_{j=1,j\neq 1}^{N} rand_i {F_{ij}}^d(t) \quad (2.9)$$

Dimana ${x_i}^d$ menunjukkan posisi agen, (t) adalah konstan gravitasi pada t,M_i adalah massa dari agen i dan $R_{ij}(t)$ adalah jarak *Euclidean* di antara agen.

$$R_{ij}(t) = \left\| X_i(t), X_j(t) \right\|_2 \quad (2.10)$$

(t) selalu *update* pada setiap iterasi dengan rumus berikut

$$G(t) = G(G_0,\mathrm{t}) \quad (2.11)$$

Dimana G_0 adalah gravitasi konstan. Massa agen $M_i(t)$ didefinisikan sebagai berikut:

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \quad (2.12)$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \quad (2.13)$$

fit_i (t) merupakan nilai *fitness* dari solusi. *The best* dan *the worst* ditentukan oleh nilai *fitness* tersebut. Berikut ini ada dua fungsi untuk meminimumkan *the best* dan *the worst*

$$best(t) = \min_{j \in \{1 \ldots N\}} fit_j(t)$$

$$worst(t) = \max_{j \in \{1 \ldots N\}} fit_j(t)$$

Percepatan (a) dan kecepatan (v) masing-masing agen didefinisikan sebagai berikut:

$$a_i{}^d(t) = \frac{F_i{}^d(t)}{M_i(t)} \quad (2.14)$$

$$v_i{}^d(t+1) = rand_i \; x \; v_i{}^d(t) \; x \; a_i{}^d \quad (2.15)$$

Langkah terakhir adalah *update* posisi masing-masing agen x.

$$x_i{}^d(t+1) = x_i{}^d(t) + v_i{}^d(t+1) \qquad (2.16)$$

Ulangi langkah di atas hingga iterasi maksimum atau hingga kriteria berhenti terpenuhi.

2.7 Indeks Validitas

Permasalahan utama pada *clustering* metode partisi adalah menentukan jumlah klaster optimal. Untuk menentukan jumlah klaster yang optimal dapat menggunakan indeks validitas (Mashfuufah & Istiawan, 2018). Indeks validitas merupakan sebuah ukuran validitas untuk menemukan jumlah klaster optimal yang sepenuhnya dapat menjelaskan struktur data (Zhao & Fränti, 2014). Pada penelitian ini digunakan beberapa indeks validitas klaster yaitu (Sara, 2018) (Pamungkas & Pramana, 2018):

2.7.1 Partition Coefficient Index (PCI)

PCI menghitung nilai rata-rata dari derajat keanggotaan pada masing-masing objek dalam matriks keanggotaan (Bezdek, 1981).

$$PCI = \frac{1}{N}\left(\sum_{i=1}^{c}\sum_{j=1}^{N}{\mu_{ij}}^2\right) \qquad (2.17)$$

Dengan

μ_{ij} : Derajat keanggotaan titik data j di dalam kelompok ke i

N : Jumlah titik data

c : Jumlah kelompok

Nilai PCI berada di antara $\left[\frac{1}{c}, 1\right]$. Jumlah kelompok optimal dinyatakan dengan nilai PCI yang maksimum.

2.7.2 Classification Entropy Index (CEI)

CEI digunakan untuk menentukan kesamaran dari partisi kelompok.

$$CEI = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\ log\ \mu_{ij} \quad (2.18)$$

Dengan:

μ_{ij} : Derajat keanggotaan titik data $-j$ di dalam kelompok ke $-i$

N : Jumlah titik data

c : Jumlah kelompok

Dari persamaan tersebut, nilai CEI berada di antara$[0, log_C]$. Jumlah kelompok optimal dinyatakan dengan nilai CEI yang minimum.

2.7.3 Separation Index (SI)

SI menghitung kekompakkan dan separasi pada masing-masing klaster.

$$SI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2\|x_j - v_i\|^2}{N min_{i,k}\|v_k - v_i\|^2} \quad (2.19)$$

Dengan:

μ_{ij} : Derajat keanggotaan titik data $-j$ di dalam kelompok ke $-i$

N : Jumlah titik data

v : Pusat kelompok

c : Jumlah kelompok

x : titik data

Jumlah kelompok yang optimal dinyatakan dengan nilai indeks SI yang minimum.

2.7.4 Xie Beni Index (XBI)

XBI bertujuan menghitung rasio dari total varians dalam klaster dan pemisahan klaster (Xie & Beni, 1991).

$$XBI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^2}{N min_{i,j}\|v_k - v_i\|^2} \quad (2.20)$$

Dengan:

μ_{ij} : Derajat keanggotaan titik data $-j$ di dalam kelompok ke $-i$

N : Jumlah titik data

m : *fuzziness*, parameter yang digunakan untuk mengukur tingkat kesamaran dari hasil pengelompokkan

v : Pusat kelompok

c : Jumlah kelompok

x : Titik data

Jumlah kelompok optimal dinyatakan dengan nilai XBI yang minimum.

2.7.5 IFV Index

IFV sering digunakan untuk memvalidasi pengelompokkan *fuzzy* dengan data spasial, karena sifatnya yang *robust* dan stabil. Ketika nilai IFV maksimum maka kualitas klaster semakin baik. Ukuran persamaannya diuraikan sebagai berikut (Hu, Meng, & Shi, 2008):

$$IFV = \frac{1}{c}\sum_{j=1}^{c}\left\{\frac{1}{N}\sum_{k=1}^{N}\mu^{2}{}_{kj}\left[log_2 c - \frac{1}{N}\sum_{k=1}^{N} log_2 \mu_{kj}\right]^2\right\}\frac{SDmax}{\overline{\sigma D}} \quad (2.21)$$

Jarak maksimum antara pusat klaster diuraikan sebagai berikut:

$$SD_{max} = max_{k\neq j}\|V_k - V_j\|^2 \quad (2.22)$$

Pembagi antara tiap objek dan pusat klaster diuraikan sebagai berikut:

$$\overline{\sigma D} = \frac{1}{c}\sum_{j=1}^{c}\left(\frac{1}{N}\sum_{k=1}^{N}\|V_k - V_j\|^2\right) \quad (2.23)$$

Dengan:

μ_{kj} : Derajat keanggotaan titik data $-k$ di dalam kelompok ke $-j$

N : Jumlah titik data

c : Jumlah kelompok

v_k : Pusat klaster ke k

2.7.6 Demam Berdarah Dengue (DBD)

Demam berdarah dengue (DBD) adalah penyakit yang disebabkan oleh virus dengue yang tergolong *Arthropod-Borne Virus, genus Flavivirus dan Flaciciridae*. DBD ditularkan melalui gigitan nyamuk *genus Aedes*, terutama *Aedes Aegypti* atau *Aedes albopictus*. Pada wilayah tropis dan subtropis penyakit DBD merupakan

endemik yang muncul sepanjang tahun, terutama saat musim hujan karena nyamuk berkembang biak secara optimal.

Demam berdarah dengue memiliki beberapa inveksi virus dengan gejala yang dibagi menjadi 3 yaitu: (a). Demam dengue tanpa gejala spesifik (b). Demam dengue dengan demam di tambah 2 gejala sepesifik yaitu pendarahan dan tanpa pendarahan (c). Demam Berdarah Dengue dengan atau tanpa *shock syndrome*. Variabel yang digunakan dalam penelitian:

a. Jumlah Penderita Penyakit DBD

Demam berdarah dengue (DBD) adalah penyakit yang disebabkan oleh virus dengue dan ditularkan melalui gigitan nyamuk *genus Aedes*, terutama *Aedes Aegypti* atau *Aedes albopictus*. Penyakit ini sebagian besar penyakit ini meyerang anak berumur kurang dari 15 tahun, namun dapat juga menyerang seluruh kelompok umur.

b. Kepadatan penduduk

Kepadatan penduduk adalah perbandingan jumlah penduduk dengan luas wilayahnya. Kepadatan penduduk menunjukkan rasio banyaknya penduduk per kilometer persegi. Kepadatan penduduk dipengaruhi oleh beberapa faktor antara lain: fisiografis, keamanan, pertumbuhan penduduk, biologis, psiologis dan lain-lain.

c. Angka Kesakitan atau *Indicate Rate* (IR)

Berdasarkan Puspitawati (2012), *Indicate Rate* (IR) yaitu jumlah penderita baru suatu penyakit yang ditemukan pada suatu jangka waktu tertentu (umumnya 1 tahun) dibandingkan dengan jumlah penduduk yang mungkin terkena penyakit baru tersebut pada pertengahan jangka waktu bersangkutan.

$$\text{IR DBD} = \frac{\text{Jumlah Penderita DBD}}{\text{Jumlah penduduk pada kurung waktu yang sama}} \times 100.000$$

d. Persentase Rumah Sehat

Rumah sehat yaitu rumah yang memenuhi kriteria minimal: akses air minum, akses jamban sehat, lantai, ventilasi, dan pencahayaan yang dihitung kumulatif dari tahun sebelumnya (Kementrian Kesehatan RI,2017). Kondisi rumah yang tidak sehat dapat beresiko menularkan penyakit, terutama penyakit berbasis lingkungan, karena rumah yang tidak sehat erat hubungannya dengan sanitasi lingkungan, sehingga semakin rendah rumah sehat disuatu wilayah, semakin tinggi peluang terkena penyakit DBD.

$$\%\ \text{Rumah Sehat} = \frac{\text{Jml rumah sehat pada kurun waktu tertentu}}{\text{Jml seluruh rumah pada kurun waktu yang sama}} \times 100\%$$

e. Perilaku hidup bersih dan sehat (PHBS)

Perilaku hidup bersih dan sehat merupakan upaya terhadap rumah tangga agar memberdayakan anggota keluarga agar sadar, mau dan mampu melakukan PHBS dalam memelihara dan meningkatkan kesehatan, mencegah tejadinya resiko penyakit dan melindungi diri dari ancaman penyakit serta berperan aktif dalam kesehatan masyarakat.

$$\text{Persentase Rumah Tangga Ber} - \text{PHB} = \frac{\text{A}}{\text{B}} \times 100\%$$

dengan,

A = Jumlah rumah tangga ber-PHBS di suatu wilayah pada periode waktu tertentu

B = Jumlah rumah tangga yang dipantau atau disurvei di wilayah dan pada waktu yang sama