

## BAB II

### TINJAUAN PUSTAKA

#### A. Tinjauan Pustaka

Metode *k-means* telah diterapkan dalam berbagai bidang ilmu misal yang berkaitan dengan kesehatan seperti penelitian pengembangan sistem untuk simulasi pemodelan aliran pasien di rumah sakit (Isken & Rajagopalan, 2002), penelitian untuk mencari pola diet dan faktor sindrom metabolik terhadap populasi (Leite & Nicolosi, 2009), mengklasifikasikan perawat menurut hasil tes kesehatan mental (Jamali & Ayatollahi, 2015), maupun penelitian untuk analisis kualitas air sungai (Zou, et al., 2015),

Pendekatan baru berdasarkan implementasi *k-means* yang disajikan untuk mendeteksi secara otomatis penyakit *erythematous-squamous* bertujuan untuk menemukan struktur untuk data yang diberikan dengan menemukan kesamaan antara data menurut data sifat *teristics*, metode ini mengelompokkan data atribut dari pasien untuk mendeteksi 5 penyakit *erythematous-squamous* untuk ditentukannya skema klasifikasi optimal (Ubeyli & Dogdu, 2010).

Penelitian yang bertujuan untuk memberikan gambaran perbandingan penggunaan alkohol di kalangan perempuan di 20 negara Afrika, data yang didapatkan berasal dari Survei Kesehatan Dunia WHO menggunakan kuesioner standar, secara total terdapat 40.739 wanita dewasa yang dilibatkan dalam penelitian ini, penggunaan metode *k-means*

bertujuan untuk mengidentifikasi negara – negara yang memiliki karakteristik serupa, hasil yang dapat disimpulkan dari penelitian adalah bertambahnya usia dikaitkan dengan peningkatan kemungkinan menjadi seorang peminum, berbagai pola minum yang hadir di antara perempuan Afrika, negara – negara dengan pola konsumsi yang berbahaya memerlukan perhatian serius untuk mengurangi bahaya yang berhubungan dengan alkohol, beberapa kesamaan dalam faktor yang berhubungan dengan penggunaan alkohol dapat diidentifikasi antara negara – negara Afrika yang berbeda, meskipun ini terbatas dan menyoroti keragaman kontekstual konsumsi alkohol para wanita di Afrika (Martinez, et al., 2011).

Penelitian untuk analisis kanker juga menggunakan metode *k-means*, penelitian ini dikarenakan adanya kebutuhan yang mendesak metodologi yang lebih efektif untuk memahami, mencegah dan menyembuhkan kanker, teknologi *microarray* memberikan dasar yang berguna untuk mencapai tujuan ini, dengan analisis kluster data ekspresi gen pada pasien kanker lalu identifikasi sub tipe tumor dan perawatan individual, *k-means* dipilih untuk kesederhanaan dan efisiensi (Iam-On & Boongoen, 2012).

Pada penelitian diagnosik kriteria untuk penyakit gunung kronis (*chronic mountain sickness/CMS*) data didapatkan dari 1.029 penduduk Cina berjenis kelamin laki – laki yang melakukan migrasi ke dataran tinggi Qinghai di Tibet data yang diteliti seperti *prevalensi* CMS dan korelasi dengan konsentrasi Hb, ketinggian dan lama tinggal, metode *k-*

*means* di penelitian ini digunakan untuk mencari nilai dari konsentrasi Hb dan titik *cut-off*, hasil penelitian berupa klaster rendah Hb ( $Hb < 200$  g/L) dengan saturasi oksigen tetap stabil dan klaster tinggi Hb ( $Hb \geq 200$  g/L) dengan saturasi oksigen menurun serta beberapa faktor yang mempengaruhi pengembangan CMS yang tinggal di ketinggian 4.500 m dan lama tinggal 60 bulan (Jiang, et al., 2014)

Penelitian kesehatan lain yang menggunakan metode *k-means* seperti penelitian mengenai identifikasi pembentukan makanan sub-jenis dan mengevaluasi prevalensi diabetes, obesitas di daerah Washington DC, dengan mengidentifikasi kelompok berdasarkan proporsi toko, restoran, makanan khusus, makanan cepat saji, toko – toko, dan apotek, hasil pengklasteran ada 2 yaitu kategori sehat yang terbagi menjadi 3 sub-tipe kategori yaitu makanan khusus, toko dan restoran lalu kategori tidak sehat yang terbagi menjadi 2 sub-tipe kategori yaitu toko – toko, dan makanan cepat saji (Frankenfield, et al., 2015).

## **B. Landasan Teori**

### **1. Malaria**

Malaria adalah penyakit parasitik yang dicirikan oleh adanya demam, menggigil, dan anemi, penyakit ini disebabkan oleh parasit yang dipindahkan dari seorang penderita ke orang lain melalui gigitan nyamuk *Anopheles* yang telah terinfeksi *Plasmodium*, di dalam tubuh manusia, parasit pada stadium sporozoit akan menyerang sel hati dan

berkembang di dalamnya menjadi bentuk dewasa yaitu *skizon eksoeritrositer* yang mengandung beribu – ribu merozoit, proses ini membutuhkan waktu satu sampai dengan tiga minggu (Arsin, 2012). Penyakit malaria juga dapat dikatakan sebagai penyakit yang muncul kembali (*re-emerging disease*), hal ini disebabkan oleh pemanasan global yang mengakibatkan penyebaran penyakit parasitik yang ditularkan melalui nyamuk dan serangga lainnya semakin ganas. Perubahan temperatur, kelembaban udara, dan curah hujan yang ekstrim mengakibatkan nyamuk lebih sering bertelur sehingga *vector* sebagai penular penyakit pun bertambah dan sebagai dampak muncul berbagai penyakit diantaranya malaria (Harijanto, et al., 2000). Penderita yang terserang malaria biasanya melalui tiga tahap yakni menggigil, diikuti oleh demam kemudian berkeringat, saat menggigil cenderung akan sakit kepala, malaise, kelelahan dan nyeri otot, selain itu kadang muntah, mual, dan diare (NIAD, 2007). Gejala malaria yang lain adalah membesarnya limpa dan anemia yang diakibatkan oleh hemolisa semua sel baik yang sehat maupun terinfeksi yang menyebabkan urin berwarna hitam (Tjay & Rahardja, 2007).

## 2. Pengenalan Pola

Banyak definisi pengenalan pola telah diusulkan, dapat didefinisikan sebagai suatu proses yang mengarah ke keputusan, juga dapat didefinisikan pengenalan pola sebagai bidang ilmu yang

bersangkutan dengan objek diskriminatif atas dasar informasi yang tersedia (Abed et al., 2010). Pengenalan pola dapat terdiri dari *supervised classification* pola masukan diidentifikasi sebagai anggota dari kelas yang telah ditetapkan, atau *unsupervised classification* kelas yang belum ditetapkan (Basu et al., 2010). Langkah dalam pengenalan pola adalah memilih satu set fitur atau atribut dari fitur semesta (keseluruhan data) yang digunakan untuk pengenalan pola, pada pola asli diubah menjadi data representasi agar dapat dengan mudah dilakukan modifikasi untuk menghilangkan *noise* yang ada, setelah *noise* dihilangkan fitur dalam data yang didefinisikan sebagai relevan melakukan pengenalan pola (Ghorpade et al., 2010).

Sistem pengenalan pola terdiri dari 3 unit dasar (Parasher et al., 2011)

- a. *Data Perception*, input visual data dari lingkungan dan dilakukan preprosesing
- b. *Feature Extraction*, proses penggalian informasi yang relevan dari data yang diolah
- c. *Classification/Recognition*, proses mengklarifikasi/mengenal informasi

### 3. Clustering

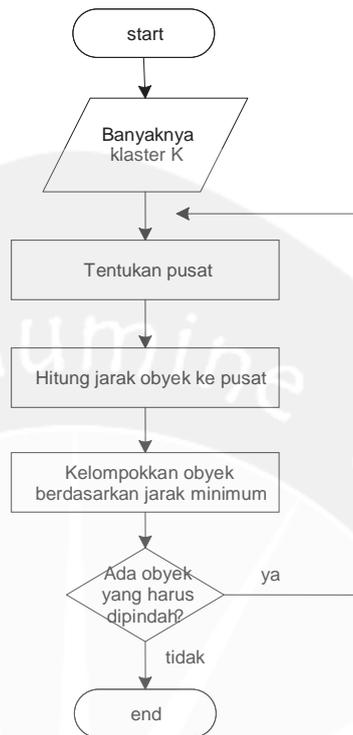
*Clustering* digunakan dalam *data mining*, sedangkan *data mining* sendiri merupakan proses penggalian informasi dari kumpulan data besar menggunakan algoritma dan teknik yang diambil dari

bidang statistik, *machine learning and database management system* (Balasubramanian & Umarani, 2012). *Clustering* adalah teknik data mining dari pengelompokan objek data ke dalam beberapa kelompok atau kluster sehingga objek dalam kluster memiliki kemiripan yang tinggi tetapi sangat berbeda dengan objek yang berada di kluster lainnya (Joshi & Kaur, 2013). *Clustering* tidak seperti klasifikasi karena tidak memiliki kelas yang ditetapkan dalam mengelompokan objek data yang besar, pendekatan *clustering* digunakan untuk mengidentifikasi kesamaan antara titik data, setiap titik data dalam kluster yang sama mengalami kesamaan yang lebih besar dibandingkan titik data dalam kluster yang lain (Tomar & Agarwal, 2013). Metode *cluster* dapat diklasifikasikan ke dalam 4 kategori yaitu *partitioning methods*, *hierarchical methods*, *density-based methods* dan *gridbased methods*, untuk metode kluster lain yang tidak termasuk klasifikasi ini tetapi telah dikembangkan seperti *fuzzy klustering*, *artificial neural network* dan *generic algorithms* (Rajagopal, 2011). Analisis kluster digunakan untuk mengidentifikasi baik benda, individu atau variabel atas dasar kesamaan, hasil dari analisis kluster adalah sejumlah kelompok heterogen dengan isi homogen (Priya & Ghosh, 2013). Analisis kluster dapat diaplikasikan secara luas seperti dalam data mining, pencarian informasi yang berkaitan dengan ilmu biologi, ilmu kedokteran, pemasaran atau juga pada *segmentasi citra*, sebagai contoh *clustering* dapat membantu produsen menemukan

kelompok yang berbeda dan ciri kelompok konsumen berdasarkan data pembelian atau dalam ilmu biologi, dapat digunakan untuk memperoleh data tanaman dan hewan taksonomi, mengkategorikan gen dengan fungsi serupa dan mendapatkan informasi mengenai populasi (Pande et al., 2012).

#### 4. Metode K-Means

*K-means* merupakan salah satu metode pengelompokan data yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok, data yang berkarakteristik sama dimasukkan ke dalam satu kelompok (Afrisawati, 2013). Algoritma *k-means* pada dasarnya melakukan 2 proses yaitu proses pendeteksian lokasi pusat klaster dan proses pencarian anggota dari tiap – tiap klaster (Agustin, et al., 2015). Algoritma *k-means* dimulai dengan pemilihan secara acak nilai K, nilai K disini merupakan banyaknya klaster yang ingin dibentuk, kemudian tetapkan nilai K secara random, untuk sementara nilai tersebut menjadi pusat dari klaster atau bisa disebut dengan *centroid* menggunakan rumus hingga ditemukan jarak paling dekat dari setiap data dengan *centroid*, klasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (Nugraha, et al., 2014).



Gambar 2.1. *Flowchart K-Means*

Algoritma *k-means* memiliki langkah – langkah sebagai berikut (Gunawan, et al., 2014):

1. Menentukan banyaknya/jumlah kluster K
2. Menentukan nilai pusat (*centroid*)

Dalam menentukan nilai *centroid* untuk awal iterasi, nilai awal *centroid* dilakukan secara acak. Sedangkan jika menentukan nilai *centroid* yang merupakan tahap dari iterasi, maka digunakan rumus

dimana:

$$\bar{V}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \dots \dots \dots (1)$$

- $V_{ij}$  adalah *centroid* atau rata – rata kluster ke-i untuk variable ke-j
- $N_i$  adalah jumlah data yang menjadi anggota kluster ke-i
- $i, k$  adalah indeks dari kluster
- $x_{kj}$  adalah nilai data ke-k yang ada di dalam kluster tersebut untuk variabel ke-j

3. Menghitung jarak antara titik *centroid* dengan titik tiap objek

Rumus untuk menghitung jarak tersebut dapat menggunakan *euclidean distance*, yaitu

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \dots \dots \dots (2)$$

Dimana:

- $D_e$  adalah *euclidean distance*
- $i$  adalah banyaknya objek
- $(x, y)$  merupakan koordinat objek
- $(s, t)$  merupakan koordinat *centroid*

4. Pengelompokan objek

Teknik untuk menentukan anggota kluster adalah dengan memperhitungkan jarak minimum objek.

5. Kembali ke tahap 2, lakukan perulangan atau iterasi hingga nilai *centroid* yang dihasilkan tetap dan anggota kluster tidak berpindah ke kluster lain.