BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Data mining telah banyak diterapkan baik pada perusahaan-perusahaan, instansi pemerintahan dan pendidikan. Penelitian terkait *data mining* juga sudah banyak dilakukan, tinjauan pustaka bertujuan sebagai bahan referensi dan rujukan terhadap hasil penelitian sebelumnya yang berhubungan dengan penelitian yang akan dilakukan.

M.J. Carvalho, P. Melo-Goncalves (2016) melakukan penelitian dengan judul "*Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation*". Penelitian yang dilakukan berdasarkan pada perubahan iklim dari berbagai variabel. Diutamakan pada perubahan jangka panjang dalam curah hujan, suhu maksimum dan minimum. Dengan menerapkan analisis pengelompokkan menggunakan *k-means* dengan perbedaan iklim harian untuk setiap variabel independen dan yang terpenting menggunakan setiap variabel sebagai fitur (versi *multivariate*). Hasilnya adalah peta dimana setiap titik grid untuk *cluster* (wilayah). Ketika menganalisis perbedaan klimatologi musiman rata-rata untuk masing-masing daerah menjadi jelas bahwa daerah tersebut memiliki musiman rata-rata. Pada kenyataannya, karakteristik yang berbeda mengenai curah hujan, suhu maksimum dan minimum diproyeksikan perubahan. Meskipun tidak signifikan secara statistik untuk setiap variabel ketika pasangan wilayah dibandingkan. Hasil penelitian menunjukkan bahwa, ketika meningkatkan jumlah *cluster* dianggap ada peningkatan rinci dalam fitur spasial yang diperoleh. Meskipun perubahan iklim musiman terdeteksi tidak jelas pada fungsi probalitas distribusi variabel asli dan

beberapa daerah ditemukan tidak signifikan berbeda satu sama lain mengenai perubahan variabel, metodologi ini menyajikan sebuah novel cara untuk mendekati subjek mengindentifikasi daerah perubahan iklim yang koheren. Selain itu, menciptakan kemungkinan untuk menentukan daerah berdasarkan beberapa variabel, bukan hanya satu.

Penelitian yang dilakukan oleh Johan Oscar Ong (2013) dengan judul penelitiannya "Implementasi Algoritma *K-Means Clustering* Untuk Menentukan Strategi Marketing President University". Penelitian yang dilakukan dari data hasil *clustering* dapat dibuat beberapa strategi promosi yang dilakukan oleh pihak *marketing President University* agar promosi yang dilakukan lebih efektif dan effisien. Adapun strategi promosi pertama yang dilakukan yaitu melakukan promosi dengan mengirim tim *marketing* yang sesuai dengan jurusan yang paling banyak diminati dan strategi promosi kedua yaitu melakukan promosi pada kota-kota di Indonesia yang didasarkan pada tingkat kemampuan akademik dari calon mahasiswa.

Josi Aranda, Wirda Astari Galvani Natasya (2016) dengan judul penelitian "Penerapan Metode *K-Means Cluster Analysis* Pada Sistem Pendukung Keputusan Pemilihan Kosentrasi Untuk Mahasiswa *International Class* STMIK AMIKOM Yogyakarta". Pengujian yang dilakukan pada penelitian ini, iterasi *clustering* data mahasiswa terjadi sebanyak 3 kali iterasi maka ditemukan hasil 4 dari 12 mahasiswa diarahkan untuk mengambil kosentrasi Pemograman dan 4 mahasiswa mengambil konsentrasi Multimedia sementara 3 – 5 mahasiswa mengambil konsentrasi Jaringan Komputer. Hasil *cluster* juga dipengaruhi dari nilai *cendroid* awal yang dipakai dan jumlah data yang dipakai, perbedaan pengambilan data pusat *cendroid* awal yang dipakai juga akan mempengaruhi hasil *cendroid* akhirnya.

Penelitian yang dilakukan oleh Fina Nasari, Surya Darma (2015) dengan penelitiannya berjudul "Penerapan *K-Means Clustering* Pada Data Penerimaan Mahasiswa Baru (Studi Kasus: Universitas Potensi Utama). Pengujian yang dilakukan iterasi *clustering* data mahasiswa terjadi sebanyak 2 kali iterasi. Berdasarkan dari hasil *cluster* kesimpulan yang dapat diambil adalah bahwa jika asal Sekolah dalam Sekolah Menengah Pertama maka rata-rata jurusan yang diambil adalah Sistem Informasi dan jika asal Sekolahnya adalah SMK rata-rata yang diambil adalah Teknik Informatika. Hasil *cluster* juga dipengaruhi dari nilai *cendroid* awal yang dipakai dan jumlah data yang dipakai, perbedaan pengambilan data pusat *cendroid* awal yang dipakai juga akan mempengaruhi hasil *cendroid* akhir.

Asroni, Ronald Adrian (2015) dengan judul penelitian "Penerapan Metode K-Means Untuk *Clustering* Mahasiswa Berdasarkan Nilai Akademik Dengan *WEKA Interface* Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang". Penelitian ini menguji data yang telah ada di *data warehouse* UMM Magelang untuk mencari 5 orang mahasiswa jurusan Teknik Informatika dalam melakukan penyeleksian untuk mengikuti lomba. Adapun lomba yang akan diikuti adalah kompetisi *event Cyberjawara* yang diselenggarakan oleh *indonesia security incident response team on internet infrastructure* (ID SIRTII) Kementerian Komunikasi dan Informatika RI. Data pengujian yang digunakan pada penelitian ini memiliki 5 atribut yaitu nim mahasiswa, nilai mata kuliah algoritma dan pemograman 1, nilai mata kuliah fisika dasar, nilai kalkulus 1, IPK dan jumlah *instance* adalah 124. *Software* yang digunakan adalah *WEKA*, tujuannya untuk membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan pada proses di *WEKA Interface*. Perhitungan jarak menggunakan persamaan *Euclidean*.

Dari data tersebut didapatkan 4 kelompok dengan hasil, *cluster* 0 dengan IPK = 0,5167 sebanyak 9 mahasiswa (7%), *cluster* 1 dengan IPK = 3,4143 sebanyak 28 mahasiswa (23%), *cluster* 2 dengan IPK = 3,3092 sebanyak 40 mahasiswa (32%) dan *cluster* 3 dengan IPK = 3,8991 sebanyak 47 mahasiswa (38%). Maka *cluster* 1 dengan IPK tertinggi bisa digunakan untuk memilih 5 mahasiswa untuk mewakili mengikuti lomba.

Yang menjadi *concern* dalam penelitian ini adalah bagaimana algoritma *k-means* mampu dalam mengelompokkan data calon mahasiswa baru di Universitas Muhammadiyah Yogyakarta berdasarkan jurusan pada Fakultas Kedokteran dan Ilmu Keperawatan (FKIK) dan Fakultas Ilmu Sosial dan Ilmu Politik (FISIPOL) dengan cara *clustering. K-means* digunakan untuk data yang beukuran besar karena memiliki kecepatan yang lebih tinggi, *k-means* juga sebagai alternatif dari metode *clustering*.

2.2. Landasan Teori

2.2.1 *Data Mining*

Data mining merupakan proses iteratif dan interaktif untuk menemukan pola atau model yang sedang tidak diketahui dapat digeneralisasi untuk masa yang akan datang, bermanfaat dan dapat dimengerti dalam suatu *database* yang sangat besar *(massive database)*. *Data mining* merupakan satu langkah dari proses *knowledge discovery in database* (KDD). *Data mining* berisi pencarian *trend* atau pola yang diinginkan dalam *database* besar untuk membantu pengambilan keputusan diwaktu yang akan datang (Fajar Astuti, 2013:3).

Menurut Fayyad dalam buku (Kusrini, 2009) Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya

kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut: (Narwati, 2010)

1. *Data selection*

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

2. *Pre-processing* atau *Cleaning*

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang *inkosisten*, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). juga dilakukan proses *enrichement*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

   Coding adalah transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat bergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

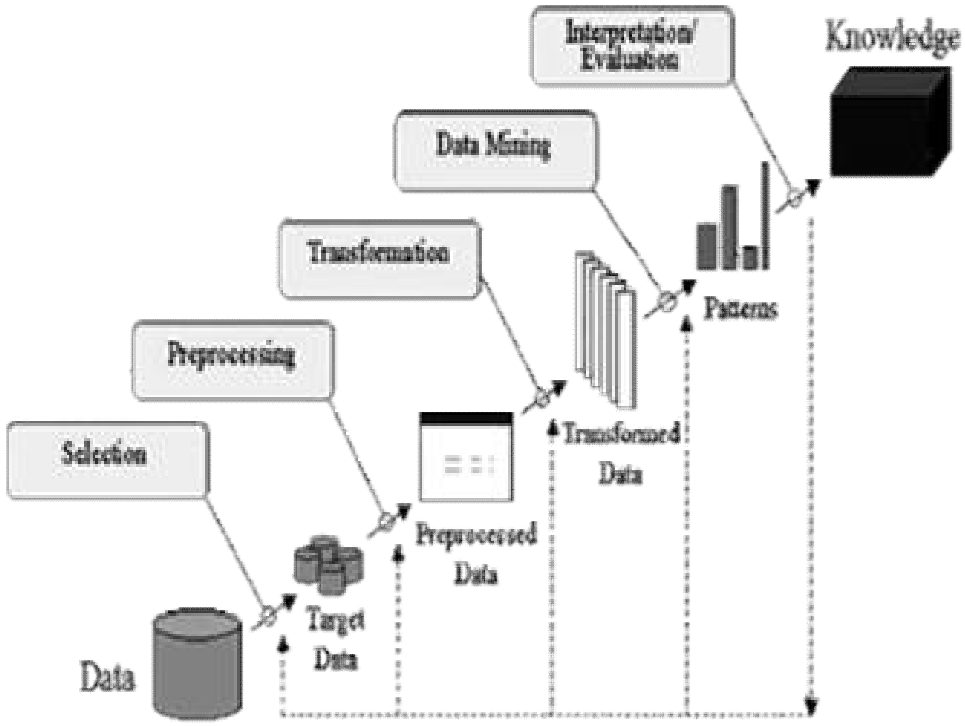
4. *Data mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma

dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation* atau *Evalution*

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



Gambar 2. 1 Proses *data mining* (Fina Nasari, 2015).

Menurut Eko Prasetyo (2012:2), salah satu teknik yang dibuat dalam *data mining* adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Dalam *data mining*, pengelompokkan data juga bisa dilakukan tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada.

Sederhananya, *data mining* adalah proses untuk menggali data dan mendapat informasi yang penting dalam data yang berukuran besar. Menurut Fajar Astuti (2013:14), adapun teknik dan sifat *data mining* adalah sebagai berikut:

a. *Classification (Predictive)*
b. *Clustering (Descriptive)*
c. *AssociationRule Discovery (Descriptive)*
d. *SequentialPattern Discovery (Descriptive)*
e. *Regression (Predictive)*
f. *DeviationDetection (Predictive)*

2.2.2 *Clustering*

Menurut Eko Prasetyo (2012:6), pengelompokkan data-data kedalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada.

Pengelompokkan dapat dibedakan menurut struktur kelompok, keanggotaan data dalam kelompok, dan kekompakan data dalam kelompok. Menurut struktur, pengelompokkan dibagi dua, yaitu hierarki dan *partitioning*. Dalam hierarki, satu data tunggal bisa dianggap sebuah kelompok, dua atau lebih. Pengelompokkan *partitioning*

membagi setiap data hanya menjadi anggota satu kelompok. Menurut keanggotaan data dalam kelompok, dibagi menjadi dua, yaitu ekslusif dan tumpang-tindih. Dalam kategori ekslusif, sebuah data bisa dipastikan hanya menjadi anggota satu kelompok dan tidak menjadi anggota kelompok yang lain. Sedangkan kategori tumpang-tindih adalah metode pengelompokkan yang membolehkan sebuah data menjadi anggota lebih dari satu kelompok. Menurut kategori kekompakan, pengelompokkan terbagi menjadi dua, yaitu komplet dan parsial. Jika semua data bisa bergabung menjadi satu, bisa dikatakan semua data kompak menjadi satu kelompok. Apabila ada satu atau dua data yang tidak ikut bergabung dalam kelompok mayoritas, data tersebut dikatakan mempunyai perilaku menyimpang. Data yang menyimpang dikenal dengan sebutan *outlier*, *noise* atau *uninterested background* (Eko Prasetyo, 2012:177).

2.2.3 *K-Means*

K-means merupakan salah satu metode pengelompokkan data *non-hierarki* yang mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang di *set* dalam suatu kelompok dan memaksimalkan variasi antar kelompok (Eko Prasetyo, 2012:178).

Pengertian dari *k-means clustering* adalah *k* dimaksudkan sebagai konstanta jumlah *cluster* yang diinginkan, *Means* dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai *cluster*, sehingga *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. Metode *k-means* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam suatu kelompok mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dasar algoritma *k-means* adalah sebagai berikut:

1. Tentukan nilai *k* sebagai jumlah klaster yang ingin dibentuk.
2. Inisialisasi *k* sebagai *centroid* yang dapat dibangkitkan secara *random*.
3. Hitung jarak setiap data ke masing-masing *centroid* menggunakan persamaan *Euclidean Distance* yaitu sebagai berikut:

$$d(P,Q) = \sqrt{\sum_{j=1}^{p} (x_j(P) - x_j(Q))^2} \qquad (1)$$

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroidnya*.
5. Tentukan posisi *centroid* baru (*k*).
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.