BAB II
LANDASAN TEORI

2.1 Data Mining

2.1.1 Pengertian Data Mininig

Data mining adalah langkah analisis terhadap proses penemuan pengetahuan didalam basisdata atau *knowledge discovery in databases* yang disingkat KDD. Pengetahuan bisa berupa pola data atau relasi antar data yang *valid* (yang tidak diketahui sebelumnya). Data mining merupakan gabungan sejumlah disiplin ilmu komputer yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode -metode yang merupakan irisan dari *artificial intelligence, machine learning, statistics, dan database systems* (Suyanto, 2017).

Data mining ditujukan untuk mengekstrak (mengambil intisari) pengethuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basisdata dan managemen data, pemprosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pasca pemprosesan terhadap struktur yang ditemukan, visualisasi, dan online *updating* (suyanto, 2017).

2.1.2 Metode Data Mining

Secara umum, metode data mining dapat dibagi menjadi dua : deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakan untuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi (Suyanto, 2017).

Metode yang ada dalam data mining adalah sebagai berikut :

1. *Classification*

   Klasifikasi merupakan proses untuk menemukan sekumpulan model yang dijelaskan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui

pada sebuah objek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih. Sedangkan data uji digunakan untuk mengetahui tingkat akurasi dan model yang telah dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai dari suatu objek data.

2. *Clustering*

   Pengelompokan data yang tidak diketahui label kelasnya kedalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripanya. Metode inilah yang digunakan dalam tugas akhir ini.

3. *Association*

   Tujuan dari metode ini yaitu untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang terhubung kuat dengan yang lainya.

4. *Regression*

   Regression mirip dengan klasifikasi. Perbedaan utamanya adalah terletak pada atribut yang diproduksi nilai yang kontinyu.

5. *Forecasting*

   Prediksi (*forecasting*) berfungsi untuk melakukan prediksi kejadian yang akan diproses berdasarkan data sejarah yang ada.

6. *Sequence Analiysis*

   Tujuan dari metode ini adalah untuk mengenali pola dari data *diskrit* sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.

7. *Deviation Analiysis*

   Tujuan dari metode ini adalah untuk menemukan penyebab perbedaan antara data yang satu dengan data yang lain dan biasa disebut sebagai *outlier detection.* Sebagai contoh adalah apakah sudah terjadi penipuan terhadap pengguna kartu kredit dengan melihat catatan transaksi yang tersimpan dalam basis data perusahaan tersebut.

2.2 *Clustering*

Salah satu metode yang diterapkan dalam KDD adalah *clustering. Clustering* adalah membagi data kedalam grup-grup yang mempunyai objek yang karakteristiknya yang sama. *Clustering* memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan *text mining*, dan analisis web (Zainul & Sarjono, 2016).

Dengan menggunakan klasterisasi, kita dapat mengidentifikasi daerah yang padat, menentukan pola-pola distribusi secara keseluruhan dan menemukan ketertarikan yang menarik antara atribut-atribut data. Dalam data mining, usaha difokuskan pada metode-metode penemuan untuk klaster pada basis data berukuran besar secara efektif dan efisien (Zainul & Sarjono, 2016). Data didapat dari hasil pengamatan suatu proses, klaster secara umum merupakan wujud himpunan bagian dari suatu himpunan data dan metode *clustering* dapat diklasifikasikan berdasarkan himpunan bagian yang dihasilkan. Pada penelitian tugas akhir ini akan dibahas mengenai penggunaan metode *K-Means* dengan pengelompokan profil mahasiswa.

2.3 Algoritma K-Means

Algoritma K-Means merupakan algoritma pengelompokan iterative yang melakukan partisi set data ke dalam jumlah K cluster yang sudah ditetapkan diawal. Algoritma K-Means sederhana untuk diimplementasikan dan dijalankan, relative cepat, mudah beradaptasi, umum dalam penggunannya dalam praktek (Prasetyo, 2014).

Teknik *clustering* yang paling sederhana dan umum dikenal adalah *clustering K-Means.* Dalam teknik ini kita ingin mengelompokan obyek kedalam K kelompok atau Cluster. Untuk melakukan *clustering,* nilai K harus ditentukan terlebih dahulu. Biasanya user atau pemakai sudah mempunyai informasi awal tentang objek yang sedang dipelajari, termasuk beberapa jumlah *cluster* yang paling tepat. Secara detail kita bisa menggunakan ukuran ketidak miripan untuk mengelompokan obyek kita.

Ketidak miripan bisa diterjemahkan dalam konsep jarak. Jika jarak dua objek atau dua titik cukup dekat maka dua objek itu mirip. Semakin dekat berarti semakin tinggi kemiripanya, semakin tinggi jarak semakin tinggi ketidakmiripanya.

Pada saat data sudah dihitung ketidakmiripan terhadap setiap *centroid*, maka selanjutnya dipilih ketidakmiripan yang paling kecil sebagai *cluster* yang akan diikuti sebagai relokasi data pada *cluster* di sebuah iterasi. Relokasi sebuah data dalam *cluster* yang diikuti dapat dinyatakan dengan nilai keanggotaan *a* yang bernilai 0 atau 1. Nilai 0 jika tidak menjadi anggota sebuah *cluster* dan 1 jika menjadi anggota sebuah *cluster*. Karena K-Means mengelompokan secara tegas data hanya pada satu *cluster*, maka dari nilai *a* sebuah data pada semua *cluster*, hanya satu yang bernilai 1, sedangkan lainya 0 seperti dinyatakan oleh persamaan berikut :

$$\begin{cases} 1 & \arg\min \{d(X_i, C_j)\} \\ 0 & \text{lainnya} \end{cases} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. \; 2.1$$

d(X_i, C_j) menyatakan ketidakmiripan (jarak) dari data ke-i ke *cluster* C_j.

Menghitung jarak setiap data ke centroid terdekat menggunakan Persamaan Euclidean yang dapat dilihat pada Persamaan 2.2

$$D(X_1, X_1) = ||x1 - x2|| = \sqrt{\sum_{j=1}^{p} \left| X_{2j} - X_{1j} \right|^2} \ldots\ldots\ldots\ldots\ldots\ldots..2.2$$

Sementara untuk mendapatkan titik *centroid* C didapatkan dengan menghitung rata-rata setiap fitur dari semua data yang tergabung dalam setiap *cluster*. Rata – rata sebuah fitur dari semua data dalam sebuah *cluster* dinyatakan oleh persamaan berikut :

$$C_j = \frac{1}{NK} \sum_{i=1}^{NK} X_{j1} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \; 2.3$$

Nk adalah jumlah data yang tergabung dalam sebuah *cluster*.

Jika diperhatikan dari langkahnya yang selalu memilih *cluster* terdekat, maka sebenarnya K-Means berusaha untuk meminimalkan fungsi objektif/fungsi biaya non-negatif, seperti dinyatakan oleh persamaan berikut :

$$J = \sum_{i=1}^{N}\sum_{i=1}^{K} a_{ic} d(x_{i}, c_{i})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \; 2.4$$

Dengan kata lain, K-Means berusaha untuk meminimalkan total jarak kuadrat di antara setiap titik X_i dan representasi *cluster* C_j terdekat.

langkah-langkah pengerjaan algoritma K-Means (Prasetyo, 2014) yaitu :

1. Inisialisasi : tentukan nilai K sebagai jumlah cluster yang diinginkan dan metrik ketidakmiripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang batas perubahan fungsi objektif dan ambang batas perubahan posisi centroid.
2. Pilih K data dari set data X sebagai centroid.
3. Alokasikan semua data ke centroid terdekat dengan metrik jarak yang sudah ditetapkan.
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah dibawah ambang batas yang diinginkan; atau (b) tidak ada data yang terpindah cluster; atau (c) perubahan posisi centroid sudah dibawah ambang batas yang ditetapkan.

2.4 Evaluasi

Terdapat beberapa metode atau indeks evaluasi yang dapat digunakan untuk mengukur kualitassebuah algoritma *clustering.* Hasil yang didapat dari beragam metode tersebut dapat berbeda karene pendekatan yang didapat pun berbeda. Ada beberapa pendekatan kriteria yang dapat digunakan dalam evaluasi, yaitu pendekatan dengan kriteria eksternal dan pendekatan dengan kriteria internal.

Salah satu metode yang biasa digunakan adalah metode *purity*, dengan menggunakan persamaan pengujian *purity. Cluster* dikatakan murni (pure) semua objek dengan *class* yang sama berada pada *cluster* yang sama. Untuk mengukur tingkat akurasi *clustering* atau 'r', pengukuran nilai 'r' ini menggunakan persamaan berikut ini :

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \qquad \ldots\ldots\ldots\ldots\ldots\ldots \text{2.4}$$

Semakin tinggi nilai r (semakin mendekati 1), semakin baik kulitas *cluster*. Sedangkan untuk menghitung *error cluster* atau 'e' seperti persamaan berikut ini :

e = 1 – r

dimana r adalah nilai tingkat kemurnian *cluster.*

2.5 Penelitian Sebelumnya

Berikut ada beberapa paper yang digunakan sebagai refrensi pembelajaran yaitu sebagai berikut :

1. " Implementasi algoritma k-means dalam pengklasteran mahasiswa pelamar beasiswa " oleh Nurul Rohmawati W, Sofi Defiyanti dan Mohamad Jajuli. Dalam penelitian ini penulis menjelaskan masalah yang terjadi dalam penelitian ini adalah banyaknya mahasiswa mengajukan cuti akademik bahkan *dropout* yakni mengenai tingginya

biaya perkuliahan yang mempengaruhi kelangsungan kegiatan belajar disebuah instansi Pendidikan tinggi. Beasiswa adalah bantuan yang diberikan kepada mahasiswa yang kurang mampu untuk memenuhi kewajibannya selama masa studi.

Metode yang digunakan adalah metode K-Means, hasil dari penelitian ini adalah membandingkan hasil *cluster* dari masing-masing format atribut dalam menentukan mahasiswa penerimaan beasiswa. Untuk mengukur kinerja algoritma, pengukuran ini dilihat dari hasil cluster dengan menghitung nilai kemurnian (*purity measure*) dari masing-masing cluster yang dihasilkan (Nurul, Sofi, Jajuli, 2015).