BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Elsa Renalia (2019) pernah dilakukan penelitian dengan topik implementasi *K-Means* untuk pengelompokan peminatan mahasiswa Teknik Informatika. Penelitian ini digunakan untuk melakukan pengelompokan data menggunakan kriteria data nilai mutu mahasiswa.

Fitri Larasati Sibuea & Andy Sapta (2017) pernah melakukan penelitian dengan topik, Pemetaan siswa berprestasi menggunakan metode "Algoritma *K-Means Clustering* Berbasis Desktop". Penelitian ini digunakan untuk melakukan pengelompokan data nilai rapot siswa SMK Yapim Simpang Kawat kelas X AK2 tahun Ajaran 2016/2017. Indikator yang digunakan meliputi : nilai tugas, nilai UTS, nilai UAS, absen dan nilai sikap.

Joko Waluyo (2019) pernah melakukan penelitian dengan topik Penerapan metode *K-Means Clustering* dalam penilaian kedisiplinan siswa untuk layanan bimbingan dan konseling. Penelitian ini digunakan untuk melakukan pengelompokan data menggunakan kriteria data siswa yang meliputi nama siswa, absensi, nilai kerapihan, dan nilai prilaku.

Siti Nur Arofah & Fitri Marisa (2018) pernah melakukan penelitian dengan topik penerapan data mining untuk mengetahui minat siswa pada matapelajaran matematika. Penelitian ini digunakan untuk melakukan pengelompokan data siswa menggunakan kriteria Nama siswa, Nilai tugas, Nilai UH, dan Nilai UAS.

Teguh Wibowo (2018) pernah melakukan penelitian dengan topik Penerapan *Data Mining* pemilihan siswa kelas unggulan dengan Metode *K-Means Clustering* sebagai Alat Bantu Pemilihan Siswa Kelas unggulan di SMP Negeri 02 Tasikmadu. Penelitian ini digunakan untuk melakukan pengelompokan data nilai siswa menggunakan kriteria nilai rapot siswa kelas tujuh yang terdiri dari tujuh kelas.

Tabel 2. 1 Perbandingan Tinjauan Pustaka

Peneliti, Tahun	Objek/ Data	Metode	Teknologi	Hasil
Elsa Ranalia, 2019	Data Nilai Mutu Mahasiswa, Sebanyak 70 siswa (AKAKOM Angkatan 2015 & 2016)	*GA K-means*	Desktop	Analisa Peminatan Mahasiswa Teknik Informatika.
Fitri Larasati Sibuea & Andy Sapta, 2017	Nilai Tugas, Nilai UTS, Nilai UAS, Absen dan Nilai Sikap (SMK Yapim Simpang Kawat kelas X AK2 tahun Ajaran 2016/2017)	*K-means Clustering*	Desktop	Pemetaan Siswa Berprestasi Berdasarkan Nilai Rapot.
Joko Waluyo, 2019	Nama Siswa, Absensi, Nilai Kerapihan, dan Nilai Prilaku (SMP Negeri 3 Temanggung)	*K-Means Clustering*	Desktop	Layanan Bimbingan dan Konseling. Berdasarkan Nilai Kedisiplinan

Siti Nur Arofah & Fitri Marisa, 2018	Nama Siswa, Nilai Tugas, Nilai UH, dan Nilai UAS	*K-Means Clustering*	Desktop	Meneliti Minat Siswa Pada Matapelajaran Matematika
Teguh Wibowo, 2018	Data Nilai Rapot Siswa Kelas Tujuh (SMP Negeri 02 Tasikmadu)	*K-Means Clustering*	Desktop	Menentukan Kelas Unggulan Berdasarkan Nilai Rapot Siswa
Usulan	Data Nilai Try Out Siswa (SMK Muhammadiyah 3 YK Jurusan TKJ Angkatan 2019)	*K-Means Clustering*	Desktop	Mengelompokan Data Nilai Siswa Untuk Melihat Keriteria Nilai Siswa Pada setiap Kelas, Melalui Data Uji Nilai Try Out

Tabel 2.1 Tabel Lanjutan

2.2 Dasar Teori

2.2.1 Data Mining

Data Mining adalah langkah analisis terhadap proses penemuan pengetahuan di dalam basis data atau *knowledge discovery in databases* yang disingkat KDD (Fayyad et al.1996). Pengetahuan bias berupa pola data atau relasi antar data yang valid (yang tidak diketahui sebelumnya). Data Mining merupakan gabungan sejumlah disiplin ilmu computer (ACM 2006), (Clifton 2010), 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* (ACM 2006).

Data Mining ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia (ACM 2006) serta meliputi basisdata infrensi, ukuran ketertarikan, pertimbangan

kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi dan online updating.

Berdasarkan fungsionalitasnya, tugas-tugas *data mining* bisa dikelompokkan ke dalam enam kelompok berikut ini (Fayyad et al, 1996):

1. Klasifikasi (*classification*): men-generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru. Misalkan, klasifikasi penyakit ke dalam sejumlah jenis, klasifikasi email ke dalam spam atau bukan.
2. Klasterisasi *(clustering*): mengelompokkan data, yang tidak diketahuilabel kelasnya, ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripannya.
3. Regresi (*regression*): menemukan suatu fungsi yang memodelkan data dengan galat (*kesalahan prediksi*) seminimal mungkin.
4. Deteksi anomaly (*anomaly detection*): mengidentifikasi data yang tidak umum, bisa berupa outlier (*pencilan*), perubahan atau deviasi yang mungkin sangat penting dan perlu investigasi lebih lanjut.
5. Pembelajaran aturan asosiasi (*association rule learning*) atau pemodelan kebergantungan (*dependency modeling*): mencari relasi antar variable.
6. Perangkuman (*summarization*): menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

2.2.2 Clustering

Prasetyo Eko (2013) mengatakan bahwa *Clustering* adalah teknik menemukan sekelompok data dari pemecahan atau pemisahan sekumpulan data menurut karakteristik tertentu yang telah ditentukan. Dalam pengelompokan tersebut nilai label nya belum diketahui sehingga diharapkan setelah melakukan pengelompokkan data dapat diketahui label dari data tersebut. Metode clustering juga sering disebut tahapan awal sebelum melakukan metode lain seperti klasifikasi.

Cluster analysis adalah mengelompokkan data objek pada informasi yang mirip atau memiliki kesamaan antara satu dengan yang lainnya, tujuannya agar dapat menemukan kelompok yang berkualitas seperti kelompok yang merupakan objek –objek yang mirip atau memiliki hubungan satu sama lain dan sebaliknya yaitu kelompok yang tidak berhubungan degan objek dalam kelompok yang lain.

Clustering cocok digunakan untuk menjelajahi data. Jika ada banyak kasus tapi tidak ada pengelompokkan yang jelas, algoritma *clustering* dapat digunakan untuk mencari pengelompokan dari data tersebut. *Clustering* juga dapat berguna sebagai data-preprocessing yaitu langkah untuk mengidentifikasi kelompok-kelompok yang berhubungan dalam membangun model.

Teknik *clustering* termasuk ke dalam teknik *unsupervised learning* dimana kita tidak perlu melatih metode tersebut atau dengan kata lain, tidak ada fase pembelajaran (*learning*). Santosa (2007) menjelaskan bahwa teknik

unsupervised learning adalah metode-metode yang tidak membutuhkan label ataupun keluaran dari setiap data yang diinvestigasi.

Tujuan utama dari *clustering* adalah pengelompokan objek-objek yang mirip kedalam satu klaster dan berusaha membuat jarak antar klaster dapat dilihat dengan membandingkan jarak objek ke *centroid* satu dengan *centroid* lainnya. Terdapat beberapa metode yang sering digunakan untuk pencarian jarak, diantaranya Manhattan dan Euclidean. Euclidean sering digunakan karena perhitungan jarak dalam *distance space* merupakan jarak terpendek yang bias didapatkan antara dua titik yang di perhitungkan, sedangkan Manhattan sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaan *outliers* dengan lebih baik.

2.2.3 K-Means

Metode K-Means pertama kali di perkenalkan oleh MacQueen JB pada tahun 1967. Metode ini adalah salah satu metode *non hierarchi* yang umum digunakan. Metode ini termasuk dalam teknik penyekatan (*partition*) yang membagi atau memisahkan objek ke k daerah bagian yang terpisah. Pada K-Means, setiap objek harus masuk dalam kelompok tertentu, tetapi dalam satu tahapan proses tertentu, objek yang sudah masuk dalam satu kelompok, pada satu tahapan bereikutnya objek akan berpindah ke kelompok lain.

Algoritma *K-Means Clustering* merupakan salah satu metode data *non-hierarchical clustering* yang dapat mengelompokkan data ke dalam beberapa cluster berdasarkan kemiripan dari data tersebut. Algoritma *K-Means* merupakan

algoritma teknik *cluster* yang berulang-ulang. Algoritma ini dimulai dengan pemilihan secara acak K, yang merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian ditetapkan nilai-nilai K secara random, untuk sementara nilai tersebut menjadi pusat cluster atau bisa disebut dengan centroid / mean. Perhitungan jarak disetiap data yang ada pada masing-masing centroid menggunakan rumus yang sudah disediakan hingga diketemukan jarak yang paling dekat dari setiap data dengan centroid. Kelompokkan setiap data berdasarkan kedekatannya dengan centroid. Langkah tersebut dilakukan terus-menerus sampai nilai centroid stabil.

Proses *clustering* dimulai dengan mengidentifikasi data yang akan diclusterkan, X_{ij} (i=1,...,n;j=1,...,m) dengan n adalah jumlah data yang akan dicluster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap cluster ditetapkan secara bebas (sembarang), C_{kj} (k=1,...,p; j=1,...,m) dengan p adalah jumlah kluster dan m adalah jumlah variabel. Kemudian dihitung jarak antara setiap data dengan setiap pusat cluster. Untuk melakukan perhitungan jarak data ke-i (X_i) pada pusat cluster ke-k (C_k), diberi nama (d_{ik}), dapat digunakan formula Euclidean (Han, Jiawei; & Kamber, Micheline. 2001) seperti pada persamaan (2.1), yaitu

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(X_{ij} - C_{kj}\right)^2} \qquad (2.1)$$

Dimana :

d_{ik} = jarak antara data yang diklusterkan ke i pada kluster ke k

X_{ij} = data ke-i pada variabel ke-j yang akan diclusterkan

C_{kj} = pusat kluster ke-k pada variabel ke-j

Suatu data akan menjadi anggota dari cluster ke-k apabila jarak data tersebut ke pusat cluster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (2.2) Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap cluster.

$$\boldsymbol{Min} \sum_{k=1}^{p} \boldsymbol{d_{ik}} = \sqrt{\sum_{j=1}^{m} (\boldsymbol{X_{ij}} - \boldsymbol{C_{kj}})^2} \qquad (2.2)$$

Dimana :

Min $\sum_{k=1}^{p}$ = Nilai rata-rata dijumlahkan cluster dan kriteria

Nilai pusat cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada cluster tersebut, dengan menggunakan rumus pada persamaan (2.3):

$$\hat{C}_{kj} = \frac{\sum_{i=1}^{r} X_{ij}}{r} \qquad (2.3)$$

Dimana :

$\hat{C}_{kj}$ = rata-rata pusat kluster ke-k pada variabel ke-j

$X_{ij} \in$ cluster ke-k

r = banyaknya anggota cluster ke k

Algoritma dasar dalam *k-means* adalah *Clustering* menggunakan metode *K-Means* secara umum dilakukan dengan algoritma sebagai berikut:

1. Tentukan jumlah cluster (k), yang ingin dibentuk.
2. Bangkitkan *k centroid* (titik pusat *cluster*) awal secara random.
3. Hitung jarak setiap data ke masing-masing *centroid*. Setiap data memilih *centroid* yang terdekat.
4. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data yang memilih pada *centroid* yang sama.
5. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

   (MacQueen JB, 1967).

2.2.4 Java

Java merupakan bahasa pemrograman tingkat tinggi yang dipelopori oleh James Gosling yang merupakan engineer di Sun Microsystem. Java mulai dibangun pada tahun 1991. Java merupakan salah salah satu bahasa yang popular saat ini, dikarenakan java dapat berjalan di berbagai *platform* sistem operasi. (Adam Mukharil Bachtiar, 2018)

2.2.5 MYSQL

MYSQL sering didefinisikan sebagai kumpulan data yang terkait. Secara teknis, yang berada dalam sebuah *database* adalah sekumpulan tabel atau objek lain (indeks, view, dan lain-lain). Tujuan utama pembuatan *database* adalah untuk memudahkan dalam mengakses data. (Abdul Kadir, 2009).