BAB II

TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Pembentukan cluster merupakan salah satu teknik yang digunakan dalam mengekstrak pola kecenderungan suatu data.Teknik ini ini digunakan dalam proses Knowledge discovery in database (KDD). Data mining biasanya identik dengan proses penggalian data-data yang cukup besar dan dikelompokkan menjadi data yang tersusun rapi. Dalam hal ini penulis mengelompokkan data mahasiswa baru tahun ajaran 2014/2015 dengan teknik clustering. Pengelompokkan yang penulis terapkan menggunakan algoritma K-Means Clustering, algoritma K-Means Clustering mampu mengelompokkan data pada kelompok yang sama dan data yang berbeda pada kelompok yang berbeda. Sehingga akan terlihat kelompok data mahasiswa baru tahun ajaran 2014/2015 pada Universitas Potensi Utama yang tidak terstruktur menjadi terstruktur.Tujuan dari penelitian ini adalah menerapkan algoritam Kmeans Clustering pada data penerimaan mahasiswa baru tahun ajaran 2014/2015(studi kasus : Universitas Potensi Utama). Hasil K-Means Clustering yang diperoleh ada dua kelompok, pusat cluster dengan Cluster 1 = 1 ; 1.75; 1.5 dan Cluster 2 = 2.95; 1.65; 1.4, Cluster pertama jika asal sekolaha dalah SMA atau Sekolah Menengah Pertama maka rata-rata jurusan yang diambil adalah Sistem Informasi dan kedua jika asal Sekolahnya adalah SMK rata-rata jurusan yang diambil adalah Teknik Informatika[1].

B. Data Mining

Data mining merupakan pemilihan atau “menambang” dari jumlah data yang banyak[2].

Data mining yang juga dikenal dengan istilah patern recognition merupakan suatu metode yang digunakan untuk megolah data yang menentukan pola tersembunyi dari data yang diolah. Data yang diolah dengan teknik data mining ini kemudian menghasilkan suatu pengetahuan baru yang bersumber dari data lama, hasil dari pengolahan data tersebut dapat digunakan dalam menentukan keputusan dimasa depan[3].

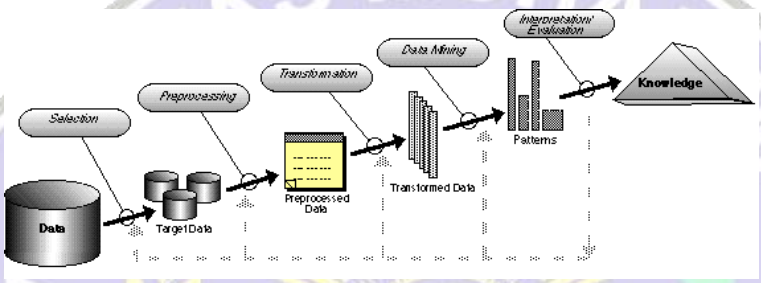
Data mining juga bisa diartikan sebagai rangkaian kegiatan untuk menemukan pola yang menarik dari data dalam jumlah besar, kemudian data-data tersebut dapat disimpan dalam database, data warehouse atau penyimpanan informasi. Ada beberapa ilmu yang mendukung teknik data mining diantaranya adalah data analisis, signal processing, neural network pengenalan pola[4].

Data mining adalah bagian dari proses KDD (knowledge discovery in database) yang terdiri dari beberapa tahapan seperti pemilihan data,pra pengolahan, transformasi, data mining, dan evaluasi[5].

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas dalam data mining adalah klastering. Tujuan utama dari klastering adalah pengelompokan sejumlah data/obyek ke

dalam klaster sehingga dalam setiap klaster akan berisi data yang semirip mungkin[6].

Metode analisis data yang digunakan adalah klastering. Adapun untuk menganalisis data dalam penerapan data mining ini menggunakan tahapan Knowledge Discovery in Databases (KDD) yang terdiri dari beberapa tahapan, [7]yaitu :



Gambar 2.1 Tahapan knowledge discovery in databases

1. Data Selection

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

2. Preprocessing atau Cleaning

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus knowledge data discovery. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti

kesalahan cetak juga dilakukan proses enrichment, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi.

3. Transformation

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam knowledge data discovery merupakan proses kreatif dan sangat tergantung 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 tergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation atau evaluation

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 informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada pada sebelumnya.

C. Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu[8] :

1. Deskripsi

Deskripsi adalah menggambarkan pola dan kecenderungan yang terdapat dalam data yang memungkinkan memberikan penjelasan dari suatu pola atau kecenderungan tersebut.

2. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numeric dari pada kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai variabel target sebagai nilai prediksi.

3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, akan tetapi dalam prediksi nilai dari hasil akan ada di masa mendatang.

4. Klarifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksikan kelas untuk data yang tidak diketahui kelasnya.

5. Pengklusteran

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain.

6. Asosiasi

Asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

D. Definisi matematis jarak Euclidean

Euclidean distance adalah metrika yang seing digunakan menghitung kesamaan dua vector. Semakin besar jarak antara dua vector, maka tingkat kesamaan atau kemiripanya semakin kecil. Dan sebaliknya, semakin kecil jarak antara dua vector, maka tingkat kesamaan atau kemiripanya semakin besar. Perancangan clustering data menggunakan algoritma k means berbasis heatmap[9]

E. Clustering

Klaster adalah salah satu teknik dalam data mining yang berkaitan dengan pengelompokan objek sesuai dengan karakteristik atau kesamaan. Dimana dalam satu klaster memiliki tingkat kesamaan karakteristik yang

tinggi (homogen) dan antar klaster memiliki perbedaan yang tinggi (heterogen). Klaster juga termasuk dalam data mining yang bersifat unsupervised learning[10].

Clustering akan melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada[11].

Clustering mengacu pada pengelompokan catatan, pengamatan, atau kasus ke dalam benda serupa. Klaster adalah kumpulan catatan yang mirip satu sama lain, dan berbeda dengan catatan di kelompok lain. Clustering berbeda dari klasifikasi dalam hal itu tidak ada target variabel untuk Clustering. Tugas pengelompokan tidak mencoba untuk mengklasifikasikan, memperkirakan, atau memprediksi nilai dari variabel target. Sebagai gantinya, algoritma pengelompokan dicari untuk mengelompokan keseluruhan data dalam sub kelompok atau kelompok yang relatif homogen, dimana kesamaan catatan dalam klaster dimaksimalkan dan kesamaanya untuk catatan diluar klaster diminimalkan[8].

F. Algortima K-Means

Metode k-Means Clustering cukup efektif diterapkan dalam proses pengklasifikasain karakteristik terhadap objek penelitian. Algoritma K-Means juga tidak terpengaruh terhadap urutan objek yang digunakan, hal ini di buktikan ketika penulis mencoba menentukan secara acak titik awal pusat klaster dari salah satu objek permulaan perhitungan[12].

Metode K-Means adalah metode yang termasuk dalam algoritma clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numeric[13].

Permasalahan yang dikaji dalam tulisan ini adalah bagaimana penggunaan metode *K-Means Cluster Analysis* dalam pengklasifikasian karakteristik suatu objek, tujuan yang ingin penulis capai adalah mengkaji metode *K-Means Cluster Analysis* dalam pengklasifikasian karakteristik berdasarkan set variabel yang dibentuk. Metode ukuran jarak yang digunakan dalam menghitung jarak objek terhadap *centroid* yaitu persamaan jarak *Euclidian*. Pada algoritma *K-Means Cluster Analysis* terdapat beberapa langkah yang harus dilakukan yaitu sebagai berikut:

1. Tentukan jumlah *cluster*.
2. Alokasikan objek ke dalam *cluster* secara random.
3. Hitung *centroid* sampel yang ada di masing-masing *cluster*.
4. Alokasikan masing-masing objek ke *centroid* terdekat.
5. Kembali ke langkah 3 apabila masih ada objek yang berpindah *cluster* atau masih ada perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan[14].

G. Microsoft Excel

Microsoft Office Excel adalah Program aplikasi pada Microsoft Office yang digunakan dalam pengolahan angka (Aritmatika). Microsoft Office Excel sangrat membantu dalam proses penyelesaian permasalahan

khususnya dalam bidang administrative. Microsoft Office Excel sangat dikenal dengan penggunaan rumus-rumus atau formula dalam lembar kerjanya. Penggunaan rumus yang efektif akan memudahkan dalam membuat laporan pekerjaan dengan menggunakan Microsoft Office Excel. Formula atau rumus pada Microsoft Office Excel adalah keunggulan tersendiri untuk aplikasi ini, dengan kemampuannya dalam mengolah data melalui perhitungan matematis yang sangat beragam fungsinya.

H. Perangkat Lunak Weka

Software yang digunakan dalam penelitian ini adalah Weka. Tujuan dari penggunaan *software* ini adalah membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan dengan proses di Weka *Interface* ini. Alat penelitian Weka *Interface*, seperti tampak pada Gambar 2.2 adalah aplikasi data mining *open source* berbasis Java. Aplikasi ini dikembangkan pertama kali oleh Universitas Waikato di Selandia Baru. Weka memiliki banyak algoritma *machine learnin*g yang dapat digunakan untuk melakukan generalisasi atau formulasi dari sekumpulan data sampling. Salah satunya adalah *klastering* dengan menggunakan algoritma K-Means.

Dalam teknik *klastering* memiliki penggunaan yang luas dan dengan saat ini memiliki kecenderungan yang semakin meningkat seiring dengan jumlah data yang terus berkembang, oleh karena itu algoritma K-means adalah teknik sederhana untuk analisis klastering. Tujuannya adalah dengan mudah untuk menemukan divisi terbaik entitas n ke dalam kelompok k

(disebut klaster), sehingga total jarak antara anggota kelompok dan entroid sesuai, terlepas dari kelompok diminimalkan. Setiap entitas milik *klaster* dengan *mean* terdekat. Ini hasil ke partisi ruang data ke Voronoi Sel.

Gambar 2.2 Interface aplikasi WEKA

Weka (waikato environment for knowledge analysis) adalah aplikasi mining open source berbasis java. Aplikasi ini dikembangkan pertama kali di selandia baru sebelum menjadi bagian dari pentaho. Weka terdiri dari koleksi algoritma machine learning yang dapat digunakan untuk melakukan generalisasi/formulasi dari sekumpulan data sampling. Walaupun kekuatan weka terletak algoritma yang makin lengkap dan canggih, kesuksesan datamining terletak pada faktor pengetahuan manusia impelentatornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan serta penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan forulais yang diharapkan.

I. Pengertian flowchart (bagan alir)

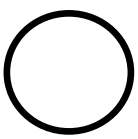
Flowchart atau bagan alir merupakan tenik analisis yang digunakan untuk menjelaskan aspek-aspek system informasi secara jelas, tepat, dan logis. Bagan alir menggunakan serangkaian symbol standar untuk menguraikan prosedur pengolahan transaksi yang digunakan oleh sebuah perusahaan, sekaligus menguraikan aliran data dalam sebuah system[15].

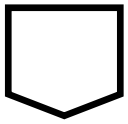
Secara garis besar flowchart atau bagan alir dapat diartikan sebagai urutan langkah-langkah/prosedur pada sebuah system guna untuk menyelesaikan permasalahan flowchart atau bagan alir memiliki symbol dan fungsinya masing-masing diantaranya sebagai berikut :

1. Flow direction symbol (symbol penghubung atau alur)

Sesuai dengan namanya symbol ini memiliki fungsi untuk menghubungkan symbol-simbol yang terdapat dalam flowchart. Symbol ini disebut juga connecting line.

Tabel 2.1 symbol penghubung atau alur

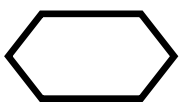
Symbol	Nama symbol	Fungsi
	Arus/flow	Menunjukan jalanya arus suatu proses.
	Communication link	Menggambarkan suatu transisi data atau transformasi dari lokasi satu ke lokasi lainya.
	On page connector	Menggambarkan sambungan satu proses dengan proses lainya masih dalam satu lembar atau halaman yang

		sama.
	Off page connector	Menggambarkan sambungan satu proses dengan proses lainya yang berbeda lembar atau halaman.

2. Processing symbol (symbol proses)

Simbol ini menggambarkan operasi dalam suatu prosedur atau proses yang ada pada flowchart.

Tabel 2.2 symbol proses

Symbol	Nama symbol	Fungsi
	Proses	Untuk menggambarkan proses atau operasi yang ada pada flowchart.
	Processing manualy	Untuk mengambarkan sautu tindakan atau proses yang dilakukan secara manual atau bisa tidak dilakukan oleh computer.
	Decision	Untuk menggambarkan suatu kondisi tertentu yang memiliki dua pilihan Ya/Tidak.
	Predefined press	Menyatakan penyediaan tempat penyimpanan suatu pengolahan untuk member nilai awal.

	Terminal	Menyatakn permulaan dan akhir suatu program atau proses.
	Off-line storage	Menyatakan permulaan dan akhir suatu program atau proses.
	Manual input	Menggambarkan proses input yang dilakukan secara manual.

3. Input-output symbol

Symbol-simbol ini menggambarkan proses input atau output dan jenis peralatan sebagai medianya.

Tabel 2.3 input-output symbol

Symbol	Nama symbol	Fungsi
	Input-output	Menggambarkan proses input atau output tanpa melihat jenis alat yang digunakan.
	Punched card	Menggambarkan input yang berasal dari kartu atau output yang ditulis ke dalam media kartu.
	Magnetic-tape unit	Menggambarkan input yang bersumber dari pita magnetic atau output disimpan ke dalam pita magnetic.
	Disk storage	Menggambarkan input bersumber dari disk atau output yang disimpan di

		disk.
	Document	Menggambarkan proses pencetakan laporan ke printer
	Display	Menggambarkan peralatan output yang dipakai merupakan layar.