BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

1.1 Tinjauan Pustaka

Tinjauan Pustaka dalam penelitian ini merupakan referensi penulisan dalam melakukan Implementasi Data Mining Pengelompokkan Pelanggan Menggunakan RFM dan K-Means Clustering. Referensi Penelitian terkait metode clustering atau algoritma k-means pernah di lakukan oleh Ari Muzakir (2014), Johan Oscar Ong (2013) ,Muhammad Toha dkk (2013), Nurhayati dan Luigi Ajeng Pratiwi (2015), dan Sylvia Pretty Tulus (2014).

Pada penelitiannya Ari Muzakir (2014) menentukan penerimaan beasiswa dengan patokan nilai Matematika, bahasa Inggris dan komputer pada sekolah SMK Ndengan tiga cluster proses menggunakan algoritma k-means sehingga akan didapatkan hasil nilai yang masuk dalam kriteria baik. Dalam pembahasan ini yang dikatakan nilai baik adalah nilai yang diatas 70.

Pada penelitiannya Johan Oscar Ong (2013) mengumpulkan seluruh data kemudian menginisialisasikan ke dalam bentuk angka agar data bisa diolah dengan menggunakan metode k-means clustering.

Pada penelitiannya Muhammad Toha, dkk (2013) melakukan pengelompokan siswa dengan melalui karater siswa, dalam penelitian ini siswa dikelompokkan dalam

4 cluster yaitu kelompok siswa berkarakter unggul, berkembang, mulai terlihat, dan kelompok siswa berkarakter lemah.

Pada penelitiannya Nurhayati dan Luigi Ajeng Pratiwi (2015) mengelompokan jurusan siswa dengan dua cluster yang akan diberi label IPA dan IPS dengan menggunakan Algoritma k-means dalam data mining.

Pada penelitiannya Sylvia Pretty Tulus (2014) mengelompokkan data spasial melalui proses normalisasi dan dikelompokkan menggunakan Algoritma K-Means. Data dikelompokkan berdasarkan jarak terdekat objek bukan berdasarkan karakteristik objek.

Perbedaan antara penelitian yang pernah dilakukan dapat di lihat pada tabel 2.1

Tabel 2.1 *Tinjauan Pustaka*

No	Nama Pengarang	Tahun	Metode	Objek Penelitian	Hasil
1	Ari Muzakir	2014	Clustering dan Algoritma K-Means	Penentuan Beasiswa	Dibentuk dalam tiga cluster
2	Johan ocar	2013	Clustering dan Algoritma K- Means	Strategi Marketing President University	Black box testing metode boundary

					value analysis
3	Muhammad toha, dkk	2013	Clustering dan Algoritma K-Means	Pencapaian Karakter Siswa	Mengelompokkan karakter siswa dalam empat cluster
4	Nurhayati dan Luigi, Ajeng Pratiwi	2015	Algoritma K-Means dalam data mining	Peminatan Jurusan bagi siswa	Dibentuk dalam dua cluster
5	Sylvia Pretty Tulus, Hendry	2014	Clustering dan Algoritma K- Means berbasis Heatmap	Data potensi hasil tambang, berupa data spesial	Dalam penelitian ini data dikelompokkan menjadi empat cluster.

1.2 Dasar Teori

2.2.1 Data Mining

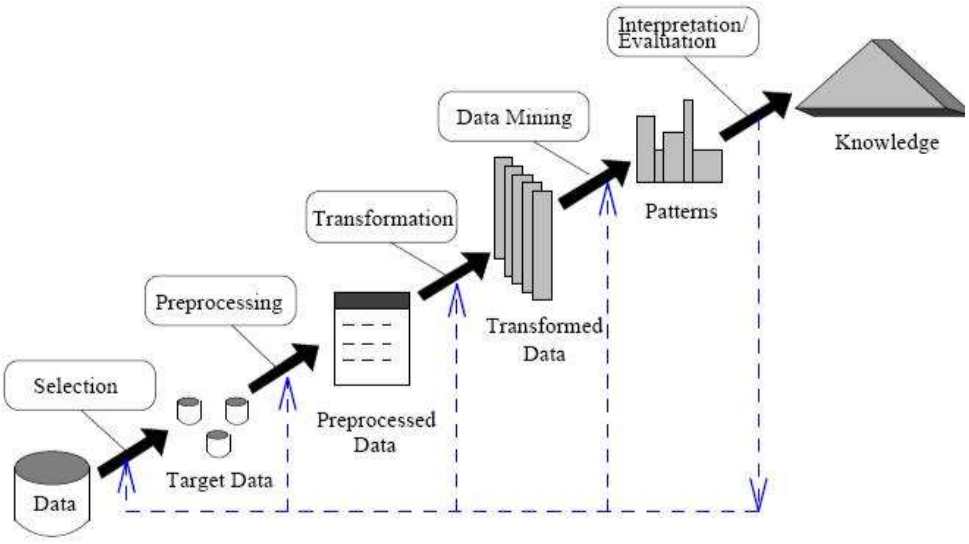
Data mining adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya. Secara teknis, data mining dapat disebut sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuan field dari sebuah relasional database yang besar(Mabrur A.G, 2012). Kemampuan Data mining untuk mencari informasi bisnis yang berharga dari basis data yang sangat besar, dapat dianalogikan dengan penambangan logam mulia dari lahan sumbernya, teknologi ini dipakai untuk:

1. Prediksi trend dan sifat-sifat bisnis, dimana data mining mengotomatisasi proses pencarian informasi pemprediksi di dalam basis data yang besar.
2. Penemuan pola-pola yang tidak diketahui sebelumnya, dimana data mining "menyapu" basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapuan.

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 (Nasari et al., 2015).

Proses KDD secara garis besar dapat dijelaskan sebagai berikut:

1. Data Selection Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining disimpan dalam suatu berkas, terpisah dari basis data operasional.
2. Pre- processing / Cleaning Sebelum proses data mining dapat dilaksanakan, perluh dilakukan proses pembersihan pada data yang menjadi fokus KDD. Proses pembersihan mencakup antara lain membuang duplikasi data, memeriksa data yang inkosisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).
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 tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.



Gambar 2. 1 *Skema dari Proses KDD*

2.2.2 Clustering

Salah satu teknik yang dikenal dalam data mining yaitu clustering. Pengertian clustering keilmuan dalam data mining adalah pengelompokan sejumlah data atau objek ke dalam cluster (group) sehingga setiap dalam cluster tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam cluster yang lainnya. Sampai saat ini, para ilmuwan masih terus melakukan berbagai usaha untuk melakukan perbaikan model cluster dan menghitung jumlah cluster yang optimal sehingga dapat dihasilkan cluster yang paling baik. Ada dua metode clustering yang kita kenal, yaitu hierarchical clustering dan partitioning. Metode hierarchical clustering sendiri terdiri dari complete linkage clustering, single linkage clustering, average linkage clustering dan centroid linkage clustering.Sedangkan metode partitioning sendiri terdiri dari k-means dan fuzzy k-means (Alfina, Santosa, & Barakbah, 2012). Pengelompokan (clustering) merupakan bagian dari ilmu data mining yang bersifat tanpa arahan (unsupervised) . Clustering adalah proses pembagian data ke dalam kelas atau cluster berdasarkan tingkat kesamaannya. Dalam clustering, data yang memiliki kesamaan dimasukkan ke dalam cluster yang sama, sedangkan data yang tidak memiliki kesamaan dimasukkan dalam cluster yang berbeda (Khotimah, Teknik, Studi, Informatika, & Kudus, 2014).

2.2.3 K-Means

K-Means Clustering adalah, K dimaksudkan sebagai konstanta jumlah cluster yang diinginkan, Means dalam hal ini berarti nilai suatu ratarata 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 kedalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain. Algoritma K-means merupakan algoritma yang membutuhakan parameter input sebanyak k dan membagi sekumpulan n objek kedalam k cluster sehingga tingkat kemiripan antar anggota dalam suatu cluster tinggi sedangkan tingkat kemiripan dengan anggota pada cluster lain sangat rendah. Kemiripan anggota tarhadap cluster diukur dengan kedekatan objek terhadap nilai mean pada cluster atau dapat disebut sebagai centroid cluster atau pusat massa. (Khotimah et al., 2014).

Berikut adalah rumus untuk menentukan jarak data dari masing-masing centrioid :

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

Keterangan :
D = titik dokumen
P = data *record*
Q = data *centroid*

I

Jarak yang terpendek antara centroid dengan dokumen menentukan posisi cluster suatu dokumen. Adapun rumus iterasi lainnya didefinisikan sebagai berikut :

$$C\ (i) = \frac{x1 + x2 + x3 + \cdots .. + xn}{\sum x}$$

Keterangan :
X_1 = Nilai data *record* ke-1
X_2 = Nilai data *record* ke-2
$\sum x$ = jumlah data *record*

2.2.4 Transformasi Normal (Normalisasi)

Suatu teknik untuk mengorganisasikan data kedalam tabel-tabel untuk memenuhi kebutuhan pemakai didalam suatu orgnisasi. Data-data yang dilakukan normalisasi dengan membagi nilai data tersebut dengan nilai range data (nilai data maksimum – nilai data minimum). Tujuan dari normalisasi yaitu:

1. Untuk menghilangan kerangkapan data
2. Untuk mengurangi kompleksitas
3. Untuk mempermudah pemodifikasian data

$$x_n = \frac{x_0 - x_{min}}{x_{max} - x_{min}}$$

Dengan,
X_n = nilai data normal
X_0 = nilai data aktual
X_{min} = nilai minimum data aktual keseluruhan
X_{max} = nilai maksimum data aktual keseluruhan

Normalisasi data input bertujuan untuk menyesuaikan nilai range data dengan fungsi aktivasi dalam sistem ELM. Ini berarti nilai kudrat input harus berada pada range 0 sampai 1. Sehingga range input yang memenuhi syarat adalah nilai data input dari 0 sampai 1. Oleh karena itu output yang dihasilkan pun akan

berada pada range 0 sampai 1. Kemudian untuk nendapatkan nilai sebenarnya dari output perlu dilakukan normalisasi (Hidayat, 2014).

2.2.5 Pelanggan

Pelanggan merupakan bagian penting dari perusahaan karena dapat memberikan keuntungan bagi perusahaan dan meningkatkan faktor pertumbuhan pada suatu perusahaan. Perusahaan akan melakukan segala cara untuk mempertahankan pelanggan yang memberikan keuntungan besar bagi perusahaan tetapi, perusahaan sulit untuk mendapatkan pelanggan yang memberikan keuntungan besar bagi perusahaan. Sifat pelanggan yang selalu pilih- pilih membuat perusahaan sulit untuk membedakan mana pelanggan yang memberikan keuntungan besar bagi perusahaan atau pelanggan yang kurang menguntungkan bagi perusahaan (Fakhri,2015).

2.2.6 Model RFM (Recency Frequency Monetary)

Model Recency Frequency Monetary (RFM) adalah model berbasis perilaku digunakan untuk menganalisis perilaku pelanggan dan kemudian membuat prediksi berdasarkan perilaku database. Model RFM ini merupakan metode yang sudah lama dan populer untuk mengukur hubungan dengan pelanggan(Nurohman, 2014).

Analisa RFM terdiri dari tiga dimensi, yaitu:

1. Recency, yaitu rentang waktu (dalam satuan hari, bulan, tahun) dari transaksi terakhir yang dilakukan oleh konsumen sampai saat ini.

2. Frequency, yaitu jumlah total transaksi atau jumlah rata-rata transaksi dalam satu periode.
3. Monetary, yaitu jumlah rata-rata nilai pembelian konsumen dalam suatu satua waktu.

2.2.7 Python

Python merupakan bahasa pemrograman yang freeware atau perangkat bebas dalam arti sebenarnya, tidak ada batasan dalam penyalinan atau mendistribusikanny.Lengkap dengansource code, debugger, dan profiler, antarmuka yang terkandung didalamnya untuk pelayanan antarmuka, fungsi sistem, GUI (antarmuka pengguna grafis), dan basis datanya. Python dapat digunakan dalam beberapa sistem operasi, seperti kebanyakan Windows, OS/2, UNIX, Machintos dan lainnya.