BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Penelitian terkait metode clustering atau algoritma k-means pernah di lakukan oleh Muhammad Toha dkk (2013), Sylvia Pretty Tulus (2014), Johan Oscar Ong (2013), Nurhayati dan Luigi Ajeng Pratiwi (2015), dan Ari Muzakir (2014). 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 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. 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 Nurhayati dan Luigi Ajeng Pratiwi (2015) mengelompokan jurusan siswa dengan dua cluster yang akan diberi label IPA dan IPS. Pada penelitiannya Ari Muzakir (2014) menentukan penerimaan beasiswa dengan patokan nilai Matematika, bahasa Inggris dan komputer dengan 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.

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

Tabel 2.1 Perbandingan penelitian

Peneliti (tahun)	**Metode**	**Objek Yang Diteliti**	**Hasil**
Muhammad toha, dkk (2013)	Clustering dan algoritma K-MEANS	Pencapaian karakter siswa	Mengelompokkan karakter siswa dalam empat cluster
Johan ocar (2013)	Clustering dan algoritma K-MEANS	strategi marketing president university	Dalam penelitian ini data-data yang ada akan dikelompokan mejadi tiga *cluster*
Sylvia Pretty Tulus, Hendry (2014)	Clustering dan algoritma K-MEANS berbasis heatmap	Data potensi hasil tambang, berupa data spasial	dalam penelitian ini data dikelompokkan menjadi empat cluster.
Nurhayati dan Luigi, Ajeng Pratiwi (2015)	Algoritma k-means dalam data mining	Peminatan jurusan bagi siswa	Dibentuk dalam dua cluster.
Ari Muzakir (2014)	Clustering dan algoritma k-means	Penentuan beasiswa	Dibentuk dalam tiga cluster

2.2. Dasar Teori

2.2.1. Data Mining

Prasetyo Eko (2013) mengatakan bahwa *Data mining* merupakan disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data. Data mining sering juga disebut knowledge discovery in

database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan.

Secara umum data mining memiliki empat tugas utama:

1. Klasifikasi (*Classification)*

Klasifikasi bertujuan untuk mengklasifikasikan item data menjadi satu dari beberapa kelas standar. Sebagai contoh, suatu program email dapat mengklasifikasikan email yang sah dengan email spam. Beberapa algoritma klasifikasi antara lain pohon keputusan, nearest neighbor, naïve bayes, neural networks dan support vector machines.

2. Regresi (*Regression*)

Regresi merupakan pemodelan dan investigasi hubungan dua atau lebih variabel. Dalam analisis regresi ada satu atau lebih variabel independentt / prediktor yang biasa diwakili dengan notasi x dan satu variabel respon yang biasa diwakili dengan notasi.

3. Pengelompokan (*Clustering*)

Clustering merupakan metode pengelompokan sejumlah data ke dalam klaster (group) sehingga dalam setiap klaster berisi data yang semirip mungkin.

4. Pembelajaran Aturan Asosiasi (*Association Rule Learning*)

Pembelajaran aturan asosiasi mencari hubungan antara variabel. Sebagai contoh suatu toko mengumpulkan data kebiasaan pelanggan dalam berbelanja. Dengan menggunakan pembelajaran aturan asosiasi, toko tersebut dapat

menentuan produk yang sering dibeli bersamaan dan menggunakan informasi ini untuk tujuan pemasaran.

Proses dari data mining mempunyai prosedur umum dengan langkah-langkah sebagai berikut :

1. Merumuskan permasalahan dan hipotesis

Pada langkah ini dispesifikasikan sekumpulan variabel yang tidak diketahui hubungannya dan jika memungkinkan dispesifikasikan bentuk umum dari keterkaitan variabel sebagai hipotesis awal.

2. Mengoleksi data

Langkah ini menitikberatkan pada cara bagaimana data dihasilkan dan dikoleksi. Secara umum ada dua kemungkinan yang berbeda. Yang pertama adalah ketika proses pembangkitan data dibawah kendali dari ahli. Pendekatan ini disebut juga dengan percobaan yang dirancang (designed experiment). Kemungkinan yang kedua adalah ketika ahli tidak memiliki pengaruh pada proses pembangkitan data, dikenal sebagai pendekatan observasional.

3. Pra pengolahan data

Pra pengolahan data melibatkan dua tugas utama yaitu:

a. Deteksi dan pembuangan data asing (outlier)

Data asing merupakan data dengan nilai yang tidak dibutuhkan karena tidak konsisten pada sebagian pengamatan. Biasanya data asing dihasilkan dari kesalahan pengukuran, kesalahan pengkodean dan pencatatan dan beberapa nilai abnormal yang wajar. Ada dua strategi untuk menangani data asing, yang pertama mendeteksi dan berikutnya membuang data asing sebagai bagian dari fase pra

pengolahan. Yang kedua adalah mengembangkan metode pemodelan yang kuat yang tidak merespon data asing.

b. Pemberian skala, pengkodean dan seleksi fitur

Pra pengolahan data menyangkut beberapa langkah seperti memberikan skala variabel dan beberapa jenis pengkodean. Sebagai contoh, satu fitur dengan range [0, 1] dan yang lain dengan range [-100, 100] tidak akan memiliki bobot yang sama pada teknik yang diaplikasikan dan akan berpengaruh pada hasil akhir data mining. Oleh karena itu, disarankan untuk pemberian skala dan membawa fitur-fitur tersebut ke bobot yang sama untuk analisis lebih lanjut.

4. Mengestimasi model

Pemilihan dan implementasi dari tehnik data mining yang sesuai merupakan tugas utama dari fase ini. Proses ini tidak mudah, biasanya dalam pelatihan, implementasi berdasarkan pada beberapa model dan pemilihan model yang terbaik merupakan tugas tambahan.

5. Menginterpretasikan model dan menarik kesimpulan

Pada banyak kasus, model data mining akan membantu dalam pengambilan keputusan. Metode data mining modern diharapkan akan menghasilkan hasil akurasi yang tinggi dengan menggunakan model dimensi-tinggi.

Pengetahuan yang baik pada keseluruhan proses sangat penting untuk kesuksesan aplikasi. Tidak perduli seberapa kuat metode data mining yang digunakan, hasil dari model tidak akan valid jika pra pengolahan dan pengkoleksian data tidak benar atau jika rumusan masalah tidak berarti.

3.1.1. Metode 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 tersubut nilai label nya belum diketahui shingga diharapkan setelah melakukan pengelompokan data dapat diketahui label dari data tersebut. Metode clustering juga sering disebut tahapan awal sebelum melakukan metode lain seperti klasifikasi.

Cluster analysis adalah mengelompokan 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 hugungan satu sama lain dan sebaliknya yaitu kelompok yang tidak berhubungan dengan objek dalam kelompok yang lain.

Clustering cocok digunakan untuk menjelajahi data. Jika ada banyak kasus tapi tidak ada pengelompokan 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.

3.1.2. K-Means

Prasetyo Eko (2013) megatakan bahwa Algoritma *K-Means clustering* merupakan teknik *cluster* berbasis jarak yang berusaha mempartisi data kedalam beberapa *cluster*. Metode ini mempartisi data kedalam *cluster* menurut karakteristik yang dimiliki setiap data, setiap data yang memiliki karakteristik

sama dikelompokkan kedalam satu cluster yang sama begitu juga dengan data yang mempunyai karakteristik berbeda dikelompokkan kedalam *cluster* lain.

Pada algoritma ini, yang menjadi pusat cluster dinamakan *centroid*, *centroid* merupakan nilai acak dari seluruh kumpulan data yang dipilih pada tahap awal, kemudian *K-Means* menyeleksi masing-masing komponen dari seluruh data dan memisahkan data tersebut kedalam salah satu *centroid* yang sudah diuraikan sebelumnya berdasarkan jarak terdekat antara komponen data dan pusat masing-masing *centroid* dengan syarat tidak ada lagi data yang berpindah kelompok. Algoritma penglompokan data *K-means* adalah sebagai berikut (Eko, P ., 2013):

1. Tentukan jumlah kelompok
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat cluster (centroid/rata-rata) dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Langkah 3,
    - apabila masih ada data yang berpindah cluster,
    - atau apabila perubahan nilai centroid ada yang di atas nilai threshold yang ditentukan,
    - atau apabila perubahan nilai pada fungsi obyektif yang digunakan masih di atas nilai threshold yang ditentukan menggunakan rumus persamaan (2.4).

Pada langkah 3 dalam Algoritma di atas, lokasi centroid (titik pusat) setiap kelompok yang diambil dari rata-rata (mean) semua nilai data pada setiap fiturnya

harus dihitung kembali. Jika M menyatakan jumlah data dalam sebuah cluster, i menyatakan fitur ke-i dalam sebuah cluster dan p menyatakan dimensi data, maka untuk menghitung centroid fitur ke-i digunakan persamaan 2.1.

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \quad \text{..... (2.1)}$$

Formula tersebut dilakukan sebanyak p dimensi sehingga i mulai 1 sampai p.

Cara mengukur jarak data ke pusat cluster menggunakan Euclidean (Bezdek, 1981) pada persamaan 2.2.

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \quad \text{..... (2.2)}$$

Keterangan persamaan 2.2:

D = jarak antara data x2 dan x1, dan | . | adalah nilai mutlak.

P = Dimensi data

X_{2j} = Koordinat dari obyek i pada dimensi k

X_{1j} = Koordinat dari obyek j pada dimensi k

Pada langkah 4 pada Algoritma, pengalokasian kembali data ke dalam masing-masing kelompok dalam metode *K-means* didasarkan pada perbandingan jarak antara data dengan sentroid setiap kelompok yang ada. Data dialokasikan ulang secara tegas ke kelompok yang mempunyai sentroid dengan jarak terdekat dari data tersebut. Pengalokasian data ke cluster menggunakan persamaan 2.3 (MacQueen, 1967):

$$a_{il} = \begin{cases} 1 & d = \min\{D(x_i, C_l)\} \\ 0 & lainnya \end{cases} \quad \text{….....(2.3)}$$

Dimana a_{il} adalah nilai keanggotaan titik x_i ke pusat cluster C_l, *d* adalah jarak terpendek dari data x_i ke K cluster setelah dibandingkan, dan C_l centroid (pusat cluster) ke-l.

Fungsi objektif berdasarkan jarak dan nilai keanggotaan data dalam cluster

$$J = \sum_{i=1}^{N}\sum_{l=1}^{K} a_{ic} D(x_i, C_l)^2 \qquad \text{..... (2.4)}$$

Dimana N adalah jumlah data, K adalah jumlah cluster, a_{il} adalah nilai keanggotaan titik data x_i ke pusat cluster C_l, C_l adalah pusat cluster ke-*l*, $D(x_i,C_l)$ adalah jarak titik x_i ke cluster C_l yang diikuti.

Untuk *a* mempunyai nilai 0 atau 1. Apabila suatu data merupakan anggota suatu kelompok maka nilai a_{il} =1 , jika tidak, akan maka nilai a_{il} =0

3.1.3. Penjurusan di SMA

Penjurusan di SMA dilakukan dengan mempertimbangkan orientasi siswa yakni sebagai berikut :

1. Melanjutkan ke pendidikan yang lebih tinggi ke program studi Ilmu Alam, Ilmu Sosial, atau Bahasa sesuai dengan minat setelah lulus dari SMA.
2. Bekerja di masyarakat; penjurusan merupakan salah satu proses penempatan atau penyaluran dalam pemilihan program pengajaran para siswa SMA. Dalam penjurusan ini, siswa diberi kesempatan memilih jurusan yang paling cocok dengan karakteristik dirinya. Ketepatan memilih jurusan dapat menentukan keberhasilan belajar siswa. Sebaliknya, kesempatan yang sangat baik bagi siswa akan hilang karena kekurangtepatan menentukan jurusan.

Tujuan penjurusan antara lain :

1. Mengelompokkan siswa sesuai kecakapan, kemampuan, bakat, dan minat yang relatif sama.
2. Membantu mempersiapkan siswa melanjutkan studi dan memilih dunia kerja.
3. Membantu memperkokoh keberhasilan dan kecocokan atas prestasi yang akan dicapai di waktu mendatang (kelanjutan studi dan dunia kerja).

Siswa yang naik kelas XI dan akan mengambil program studi tertentu (IPA, IPS dan Bahasa) boleh memiliki nilai tidak tuntas paling banyak tiga pelajaran. Mata pelajaran IPA lebih menitik beratkan pada penguasaan konsep-konsep IPA untuk kepentingan siswa menyelesaikan masalah dalam kehidupan sehari-hari. Fungsi yang lain adalah memberikan makna pembekalan agar siswa tersebut dapat survive di percaturan kompetisi perkembangan sains dan teknologi bagi kepentingan kesejahteraan masyarakatnya. Dengan demikian penilaian akademik lebih terfokus pada penguasaan konsep-konsep IPA dan keterampilannya dalam melakukan observasi, memahami atau menemukan konsep-konsep IPA.

Untuk mata pelajaran IPS menitikberatkan pengembangan keterampilan ilmu sosial. Penilaian akademik menitiberatkan pada keterampilan sosial seperti membuat peta, maket rumah, interaksi sosial, dan adaptif terhadap lingkungan sosial. Mata pelajaran Bahasa menitikberatkan pengembangan keterampilan bahasa seperti membuat surat, menyusun karya tulis, mengerjakan instruksi lisan, dialog dan berpidato.

IPA dan IPS sama-sama membutuhkan keahlian tersendiri dan sama-sama memerlukan minat dan kecerdasan. Maka orang tua dan guru seyogyanya bersikap arif dalam penjurusan ini. Ajaklah anak-anak kita mengenali minat dan

potensi mereka sendiri sekaligus arahkanlah sesuai hal tersebut. Bila sang anak berminat memasuki jurusan IPS, maka guru dan orang tua patut mendorong dan mendukungnya demikian pula sebaliknya. Bagi para guru BK/BP di pundak andalah tanggung jawab untuk membimbing para siswa mengenali potensi dirinya masing-masing.

Berdasarkan buku pedoman Pelaporan Hasil belajar Peserta didik untuk Kurikulum berbasis kompetensi dari Dirjen Didasmen Jakarta tahun 2006 dan keputusan rapat wali kelas, Komite dan BK, sekolah menetapkan sementara membuka 2 program yaitu IPA dan IPS.