BAB II
LANDASAN TEORI

2.1. Data Mining

2.1.1. Pengertian Data Mining

Data mining adalah salah satu teknik penelusuran data untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Dalam *data mining*, pengelompokan data juga dilakukan. Tujuannya adalah agar penulis dapat mengetahui pola dan tindak lanjut yang diambil. Semua hal tersebut bertujuan untuk mendukung kegiatan evaluasi agar sesuai dengan yang diharapkan (Prasetyo, 2012).

Data mining ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basis data dan manajemen data, pemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pasca pemrosesan 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*

Analisis kluster atau *clustering* merupakan proses membagi data dalam suatu himpunan kedalam berberapa kelompok yang kesamaan datanya dalam suatu

kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok yang lain.(Kusrini & Emha, 2009).

Potensi *clustering* dapat dignakan untuk menfetahui struktur dalam data yang dapat dipakai lanjut dalam berbagai aplikasi secara luas seperti klasifikasi, pengolahan gambar dan pengenalan pola. *Clustering* dapat diterapkan kedalam data yang kuantitatif (numerik), kualitatif (kategorikal) atau kombinasi keduanya, (Kusrini & Emha. 2009). Data dapat merupakan hasil pengamatan dari suatu prose. *Cluster* secara umum merupakan wujud himpunan bagian dari suatu himpunan data dan metode *clustering* dapat diklasifikasi berdasarkan himpunan bagian yang dihasilkan : apakah *fuzzy* atau *crips (hard).*

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, relatif 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 objek 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 :

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

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

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

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 :

$$C_j = \frac{1}{NK} \sum_{i=1}^{NK} X_{j1} \quad \text{..... 2.3}$$

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 :

$$J = \sum_{i=1}^{N} \sum_{i=1}^{K} a_{ic} d(x_{i,} c_i)^2 \quad \text{..... 2.4}$$

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. Davies-Bouldin Index

Davies-Bouldin Index diperkenalkan oleh David L. Davies and Donald W. Bouldin pada tahun 1979. *Sum-of square within cluster* (SSW) sebagai metrik kohesi dalam sebuah cluster. Separasi dengan *Sum-of-square-between-cluster* (SSWB) dengan mengukur jarak antara centroid C_i dan C_j. $R_{i,j}$ adalah ukuran rasio seberapa baik nilai perbandingan antara cluster ke-i dan cluster ke-j. Rumus SSW adalah :

$$SSW = \frac{1}{N}\sum_{i=1}^{N}\|x_i - C_{Pi}\|^2$$

Rumus SSB :

$$SSB = \frac{2}{M(M-1)}\sum_{I=1}^{M}\sum_{j=1,j\neq 1}^{M}\|C_i - C_j\|^2$$

Rumus R dan DBI

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \qquad DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}(R_{i,j})$$

2.5. MA Kanjeng Sepuh Sidayu

MA KANJENG SEPUH SIDAYU yang terletak di Jalan Pemuda No.75 Kecamatan Bunderan Kabupaten Gresik adalah sebuah instansi sekolah dalam bidang kejuruan. Memiliki jumlah murid yang cukup banyak yang terbagi menjadi 2 jurusan yaitu Ilmu Pengetahuan Alam (IPA) dan Ilmu Pengetahuan Sosial (IPS).

MA Kanjeng Sepuh Sidayu Gresik mungkin sebagai satu-satunya sekolah yang memakai *Dual System Education* di wilayah Jawa Timur. *Dual System Education* ini adalah diberlakukannya sistem pendidikan nasional dan sistem pendidikan pondok pesantren. jadi di samping materi pelajaran yang diajarkan adalah berdasarkan kurikulum pendidikan nasional, MA Kanjeng Sepuh juga mengajarkan materi pelajaran yang diajarkan di banyak pondok-pondok pesantren antara lain Nahwu, Shorof, Manteq, Balaghoh dan banyak muatan lokal yang lain.

Dipakainya *Dual System Education* ini dimaksudkan agar lulusan MA Kanjeng Sepuh di samping menguasai ilmu pengetahuan umum juga agar lulusan MA Kanjeng Sepuh menguasai ilmu agama yang cukup yang sangat dibutuhkan setelah ia lulus dan bergaul di masyarakat.

2.6. Penjurusan

2.6.1. Pengertian Penjurusan

Jurusan adalah satu seri materi pendidikan yang sudah ditentukan secara sistematis sesuai dengan bidangnya (Wiwik Retnowati, 2015). Sistem jurusan di MA dilakukan pada awal masuk semester 1 kelas X, ini merupakan bentuk penempatan dan penyaluran siswa sesuai minat dan bakat serta kemampuan yang dimiliki siswa di sekolah MA ini dalam penjurusan ada 2 jurusan yang harus dipilih yaitu: Jurusan Ilmu Pengetahuan Alam (IPA) dan Ilmu Pengetahuan Sosial (IPS). Dimana setiap jurusan minimal mencapai rata-rata sebagai persyaratan pemilihan jurusan tersebut.

2.6.2. Tujuan Penjurusan

Tujuan penjurusan sendiri adalah agar kelak dikemudian hari pelajaran yang akan diberikan kepada siswa menjadi lebih terarah karena sesuai dengan minat dan bakatnya. Sekolah memegang peranan penting untuk dapat mengembangkan potensi diri yang dimiliki siswa. Kemungkinan yang akan terjadi jika siswa mengalami kesalahan dalam penjurusan. Perlu diingat bahwa sebetulnya antara 2 jurusan IPA dan IPS memiliki karakteristik masing-masing.

2.7. Penelitian Terkait

1) Pada penelitian "Analisis Algoritma K-Means Untuk Sistem Pendukung Keputusan Penjurusan Siswa Di MAN Binong Subang" oleh Wijaya. Penulis menjelaskan masalah yang terjadi dalam penelitian ini adalah proses penentuan jurusan SMA, hal ini dilakukan untuk mengetahui kemampuan siswa sesuai keahliannya sehingga pihak sekolah dapat mengelompokkan kemampuan siswa sesuai dengan bakatnya. Proses Clustering Algoritma K-Means , Pada tahap ini akan dilakukan proses utama yaitu segmentasi data nilai yang diakses dari database yaitu sebuah metode clustering algoritma K-Means dengan asumsi bahwa parameter input adalah jumlah *data set* sebanyak *n data* dan jumlah inisialisasi *centroid* K=2 sesuai dengan jumlah jurusan yang ada di MAN Binong yaitu IPA dan IPS. Adapun tujuan yang ingin dicapai adalah untuk

menguji tepat atau tidaknya algoritma K-Means dalam sistem pendukung keputusan penjurusan siswa. Dari berdasarkan hasil analisis terhadap algoritma K-Means untuk sistem pendukung keputusan penjurusan, maka kesimpulan yang dapat diambil adalah algoritma K-Means kurang tepat untuk sistem pendukung keputusan penjurusan tetapi algoritma K-Means lebih tepat untuk pengelompokan data siswa berdasarkan data nilai yang bisa memberikan gambaran untuk penjurusan siswa (Wijaya, 2011).

2) "Pengelompokan potensi akademik siswa RA TARBIYATUL AULAD dengan metode K-Means" oleh Ba'natus Sa'adah. Penulis menelaskan masalah yang terjadi dalam penelitian ini adalah yakni perkembangan sekolah dasar yang melakukan seleksi bagi siswa TK maupun RA, hal ini dilakukan untuk mengetahui kemampuan dan pengetahuan siswa tersebut sehingga sekolah dasar dapat mngelompokkan kemampuan siswa yang berbeda-beda. Metode yang digunakan adalah metode K-Means, hasil dari penelitian ini menghasilkan bahwasanya metode K-Means dapat digunakan untuk menngelompokkan potensi akademik siswa RA TARBIYATUL AULAD. (Sa'adah, 2014).