

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam penelitian yang dilakukan sebelumnya (Vernanda et al., 2019) penerapan fuzzy c - means untuk menentukan target promosi penerimaan mahasiswa baru bertujuan untuk menentukan sekolah yang menjadi target promosi. Sehingga dapat digunakan panitia penerimaan mahasiswa baru dalam pengambilan keputusan sekolah yang akan diadakan promosi

Selanjutnya penelitian yang dilakukan (Handayanto et al., 2019) dengan judul analisis dan penerapan algoritma support vector machine dalam data mining untuk menunjang strategi promosi bertujuan untuk mengetahui strategi promosi yang lebih efisien dan efektif. Sehingga dapat menunjang strategi promosi pada calon mahasiswa baru

Selanjutnya penelitian yang dilakukan (Wira et al., 2019) dengan judul implementasi metode k-medoids clustering untuk mengetahui pola pemilihan program studi mahasiswa baru tahun 2018 di universitas kanjuruhan malang bertujuan agar dapat mengetahui mahasiswa baru tersebut benar – benar memilih program studi yang masih linear dengan bidang keilmuan yang dipelajari. Sehingga dapat diketahui pola pemilihan program studi bagi mahasiswa baru.

Selanjutnya penelitian yang dilakukan (Yunita, 2018) dengan judul penerapan data mining menggunakan algoritma k-means clustering pada penerimaan mahasiswa baru (studi kasus : Universitas Islam Indragiri) bertujuan untuk mempromosikan masing-masing program studi yang ada di universitas islam indragiri. Sehingga hasil cluster dapat dilihat dari jurusan atau program studi yang di minati di masing - masing sekolah.

Selanjutnya penelitian yang dilakukan (Wirta & Erlin, 2016) dengan judul implementasi metode k - means cluster analysis untuk memilih strategi promosi penerimaan mahasiswa baru bertujuan agar dapat membantu dalam pemilihan strategi promosi penerimaan mahasiswa baru pada STMIK Amik Riau. Untuk

membuat perencanaan dan perancangan strategi promosi yang lebih baik dan tepat.

Selanjutnya penelitian yang dilakukan (Nasari & Darma, 2015) dengan judul penerapan k-means clustering pada data penerimaan mahasiswa baru (studi kasus : Universitas Potensi Utama) bertujuan untuk mengetahui jurusan yang diminati jika asal sekolah adalah SMA maka rata - rata jurusan yang diambil adalah sistem informasi dan jika asal sekolah SMK maka rata - rata jurusan yang diambil adalah teknik informatika. Dari data tersebut maka diketahui jurusan apa yang nantinya akan diambil oleh mahasiswa baru ditahun depan.

Table 2.1 Table Tinjauan Pustaka

No	Tahun	Peneliti	Judul	Metode	Yang Diukur	Kelebihan	Kekurangan	Hasil
1	2019	Vernanda	Penerapan Fuzzy C Means untuk menentukan target promosi penerimaan mahasiswa baru	Metode Fuzzy C Means	Jumlah siswa melanjutkan perguruan tinggi, Rata-rata nilai UN, Jarak sekolah ke POLSUB	Melakukan clustering lebih dari banyaknya huruf dengan cara bersamaan	Membutuhkan banyak suku serta matriks tim kelompoknya yang ditetapkan	Digunakan oleh panitia penerimaan mahasiswa baru dalam pengambilan keputusan sekolah yang akan diadakan promosi
2	2019	Handayanto	Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi	Metode Support Vector Machine (SVM)	jenis kelamin, agama, asal sekolah, jurusan, asal kota/kabupaten dan daftar ulang	SVM dapat diimplementasikan relatif mudah, karena proses penentuan support vector dapat dirumuskan dalam QP	Sulit dipakai dalam problem berskala besar. Skala besar dimaksudkan dengan jumlah sample yang diolah	Diperlukan strategi promosi yang lebih efisien dan efektif.

No	Tahun	Peneliti	Judul	Metode	Yang Diukur	Kelebihan	Kekurangan	Hasil
3	2019	Wira	Implementasi Metode K-Medoids clustering untuk Mengetahui pola pemilihan program studi Mahasiswa baru tahun 2018 di Universitas Kanjuruhan Malang	Metode K – Medoids	Program studi berdasarkan nilai tes, asal sekolah, dan program studi	Hasil proses clustering tidak bergantung pada urutan masuk dataset	Medoids diambil secara acak tanpa melihat kemiripan antar medoids	Pengelompokan mahasiswa baru menunjukkan bahwa mahasiswa baru yang berasal dari SMA/SMK dengan nilai ujian
4	2018	Yunita	Penerapan data mining menggunakan algoritma K-Means clustering Pada penerimaan mahasiswa baru	Metode K-Means cluster	Asal sekolah, program studi dan nilai UAN	Sangat fleksibel, adaptasi yang mudah untuk dilakukan	Sebelum algoritma di jalankan, titik K diinisialisasi secara random sehingga pengelompokan data yang di dapatkan bisa berbeda	Dapat dilihat jurusan atau program studi yang di minati di masing-masing sekolah
5	2016	Wirta Dan Erlin	Implementasi Metode K-Means Cluster Analysis untuk Memilih Strategi Promosi Penerimaan Mahasiswa Baru	Metode K-Means cluster	Media online, spanduk, brosur dan secara lisan	Sangat umum penggunaannya	Diinisialisasi secara random sehingga pengelompokan data yang di dapatkan bisa berbeda-beda.	Sehingga dapat dilihat strategi promosi mana yang lebih tepat dan terencana sebagai panduan bagian promosi untuk memilih strategi promosi yang sangat efektif

No	Tahun	Peneliti	Judul	Metode	Yang Diukur	Kelebihan	Kekurangan	Hasil
6	2015	Nasari Dan Darma	Penerapan K-Means Clustering pada data penerimaan mahasiswa baru	Metode K-Means cluster	Asal sekolah SMA atau SMK	Mudah dilakukan saat pengimpel ementasia n dan di jalankan	Sebelum algoritma di jalankan, titik K diinisialisasi secara random sehingga pengelompokan data yang di dapatkan bisa berbeda-beda.	Pertama jika asal sekolah adalah SMA maka rata-rata jurusan yang diambil adalah Sistem Informasi dan kedua jika asal Sekolahnya adalah SMK rata-rata jurusan yang diambil adalah Teknik Informatika

2.2 Landasan Teori

Dalam landasan teori ini akan dijelaskan secara garis besar teori – teori yang menjadi dasar atau acuan dalam penelitian :

2.2.1 Definisi Data mining

Data mining adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaanya. Kata mining berasal dari kiasan bahasa inggris, mine. Jika mine berarti menambang sumber daya yang tersembunyi di dalam tanah, maka Data Mining merupakan penggalian makna yang tersembunyi dari kumpulan data yang sangat besar. Karena itu data mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basis data (Wicaksana et al., 2013).

Data mining merupakan proses menemukan pola yang menarik dari jumlah besar data. Sebagai proses penemuan pengetahuan, biasanya melibatkan pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola, dan presentasi pengetahuan. Dalam data mining terdapat

beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar (Larose & Larose, n.d.). Data mining dapat dilakukan pada semua jenis data selama data tersebut bermakna untuk aplikasi target, seperti data basis data, data gudang data, transaksional data, dan tipe data lanjutan. Tipe data lanjutan termasuk yang terkait dengan waktu atau urutan data, aliran data, data spasial dan spasial, teks dan data multimedia, grafik dan data jaringan, atau data Web.

2.2.2 Pengelompokan Teknik Data Mining

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larose & Larose, n.d.):

A. Estimation

Dalam estimasi, kami memperkirakan nilai variabel target numerik menggunakan satu set variabel prediktor numerik atau kategoris. Model dibangun menggunakan "lengkap" catatan, yang menyediakan nilai variabel target, serta prediktor. Kemudian, untuk pengamatan baru, estimasi nilai dari variabel target dibuat, berdasarkan nilai-nilai prediktor.

B. Prediction

Prediksi mirip dengan klasifikasi dan estimasi, kecuali untuk prediksi, hasilnya ada di masa depan. Tugas prediksi dalam bisnis dan penelitian termasuk :

1. Memprediksi harga saham 3 bulan ke depan
2. Memprediksi peningkatan persentase kematian lalu lintas tahun depan jika batas kecepatan meningkat
3. Memprediksi pemenang World Series musim gugur ini, berdasarkan perbandingan statistik tim

C. Clustering

Digunakan untuk menganalisis pengelompokan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokan belum didefinisikan sebelum dijalankannya tool data mining.

D. Association

Digunakan untuk mengenali kelakuan dari kejadian - kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian.

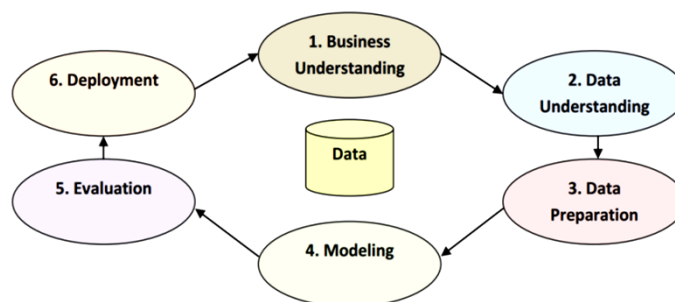
E. Classification

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan.

2.2.3 CRISP - DM

Proses CRISP-DM adalah yang paling luas kerangka kerja yang diadopsi untuk mengembangkan solusi data mining (Chauhan & Kaur, 2014). Proses CRISP - DM ini terdiri dari enam langkah atau fase, seperti berikut (North, 2012).

:



Gambar 2.1 Gambar CRISP - DM

Refrensi (North, 2012)

1. Fase Pemahaman Bisnis / Penelitian (Business Understanding)

Pertama, Jelaskan tujuan dan persyaratan proyek dalam hal unit bisnis atau penelitian secara keseluruhan

2. Fase Pemahaman Data (Data Understanding)

Kedua, Analisis data untuk membiasakan diri dengan data tersebut dan temukan wawasan awal

3. Tahap Persiapan Data (Data Preparation)

Ketiga, Pilih case dan variabel yang ingin di analisis dan lakukan transformasi pada variabel tertentu (jika perlu)

4. Pemodelan (Modeling)

Keempat, Pilih dan terapkan pemodelan yang sesuai dan kalibrasi pengaturan model untuk mengoptimalkan hasil

5. Evaluasi (Evaluation)

Model – model ini harus di evaluasi untuk kualitas dan menentukan apakah model – model itu mencapai tujuan yang ditentukan

6. Deployment

Hasilkan laporan dari pembuatan model yang sudah dibuat

2.2.4 Clustering

Algoritma Clustering mempartisi objek data (pola, entitas, unit) menjadi sejumlah cluster (kelompok, himpunan bagian, atau kategori). Namun, tidak ada kesepakatan universal dan definisi yang tepat dari istilah cluster. Seperti dirangkum oleh Everitt (1980) dan diilustrasikan sebagai berikut:

“ Cluster adalah seperangkat entitas yang sama, dan entitas dari cluster berbeda tidak sama. ”

Sebuah cluster adalah" agregat titik di ruang uji sedemikian rupa sehingga jarak antara dua titik di cluster kurang dari jarak antara titik mana pun di cluster dan titik mana pun yang tidak ada di dalamnya. ”

" Cluster dapat digambarkan sebagai daerah kontinu dari ruang ini (ruang d fitur-dimensi) yang mengandung kerapatan titik yang relatif tinggi, dipisahkan dari daerah semacam itu oleh daerah yang memiliki kerapatan titik yang relatif rendah (Ye, 2014).

Clustering telah diterapkan di berbagai bidang, seperti diilustrasikan di bawah ini :

1. Teknik (kecerdasan komputasi, pembelajaran mesin, pengenalan pola, teknik mesin, teknik listrik). pengelompokan dalam rentang teknik mulai dari pengenalan biometrik dan pengenalan suara, hingga analisis sinyal radar, kompresi informasi, dan penghilangan noise.
2. Ilmu komputer. Kami telah melihat semakin banyak aplikasi pengelompokan dalam penambangan web, analisis basis data spasial, pengambilan informasi, pengumpulan dokumen tekstual, dan segmentasi gambar.
3. Kehidupan dan ilmu kedokteran (genetika, biologi, mikrobiologi, paleontologi, psikiatri, klinik, filogeni, patologi). Area - area ini terdiri dari aplikasi utama pengelompokan pada tahap awal dan akan terus menjadi salah satu bidang bermain utama untuk algoritma pengelompokan. Termasuk definisi taksonomi, identifikasi fungsi gen dan protein, diagnosis dan perawatan penyakit, dan sebagainya.
4. Astronomi dan ilmu bumi (geografi, geologi, penginderaan jauh). Pengelompokan dapat digunakan untuk mengklasifikasikan bintang dan planet, menyelidiki formasi lahan, wilayah partisi dan kota, dan mempelajari sistem sungai dan gunung.
5. Ilmu sosial (sosiologi, psikologi, arkeologi, antropologi, pendidikan). Analisis pola perilaku, identifikasi hubungan antar budaya yang berbeda, konstruksi sejarah evolusi bahasa, analisis jejaring sosial, temuan arkeologis dan klasifikasi artefak, dan studi kriminal psikologi.

6. Ekonomi (pemasaran, bisnis). Aplikasi dalam karakteristik pelanggan dan pola pembelian pengakuan, pengelompokan perusahaan, dan analisis tren saham semua mendapat manfaat dari penggunaan analisis cluster.

Teknik clustering berlaku ketika tidak ada kelas yang dapat diprediksi tetapi contoh harus dibagi ke dalam kelompok alami. Cluster ini mungkin mencerminkan beberapa mekanisme yang bekerja di domain dari mana contoh diambil, sebuah mekanisme yang menyebabkan beberapa contoh memiliki kemiripan yang lebih kuat untuk masing-masing selain yang mereka lakukan untuk contoh yang tersisa. Clustering secara alami membutuhkan perbedaan teknik klasifikasi dan asosiasi metode pembelajaran yang kami telah dipertimbangkan sejauh ini. Algoritma klaster k-means di mana k titik awal dipilih untuk mewakili pusat cluster awal, semua titik data ditugaskan ke terdekat, nilai rata-rata dari titik-titik di setiap cluster dihitung untuk membentuknya pusat cluster baru, dan iterasi berlanjut sampai tidak ada perubahan dalam cluster (North, 2012).

2.2.5 K – MEANS Clustering

'K' dalam k-berarti pengelompokan adalah singkatan dari banyaknya kelompok, atau kelompok. Algoritma k-berarti menyelesaikan ini dengan sampling beberapa set pengamatan dalam set data, menghitung rata-rata, atau rata-rata, untuk masing-masing atribut untuk dilihat dalam sampel itu, dan kemudian membandingkan atribut lainnya dalam data set ke sarana sampel itu (North, 2012).

K - means clustering adalah salah satu algoritma pengelompokan yang paling sederhana dan paling umum digunakan. Dalam teknik ini, pengguna menentukan jumlah cluster (k) yang perlu dikelompokkan dalam kumpulan data. Tujuan dari k-means clustering adalah untuk menemukan titik data untuk setiap cluster. Pusat cluster dapat menjadi rata-rata dari semua objek data dalam cluster, seperti dalam k-means, atau objek data yang paling terwakili. Pembelajaran ini termasuk unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan

mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Dasar algoritma K - means adalah sebagai berikut :

1. Tentukan nilai k sebagai jumlah klaster yang ingin dibentuk
2. Bangkitkan k centroid (titik pusat klaster) awal secara random
3. Hitung jarak setiap data ke masing-masing centroid menggunakan Rumus korelasi antar dua objek yaitu Euclidean Distance
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya
5. Tentukan posisi centroid baru dengan cara menghitung nilai rata - rata dari data-data yang ada pada centroid yang sama
6. Ulangi langkah 3 sampai dengan 5 hingga konvergen atau penghentian. Hingga pusat centroid tetap tidak berubah, perhitungan berakhir

K - means dianggap sebagai pokok metode pengelompokan, karena kemudahan implementasi. Ini bekerja dengan baik untuk banyak masalah praktis, terutama ketika cluster yang dihasilkan kompak dan hiperpheris. Karena itu, K - means adalah pilihan yang baik untuk mengelompokkan set data skala besar (Ye, 2014)