BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Data mining dalam penelitian dan penerapannya sudah banyak dilakukan pada berbagai organisasi, instansi pemerintah dan pada bidang pendidikan. Penelitian-penelitian terdahulu yang memiliki keterkaitan dengan penelitian yang dilakukan menjadi sebuah rujukan dan referensi sehingga bisa menjadi landasan teori.[3]

Pada judul 'Penerapan *naïve bayes classifier* untuk pemilihan konsentrasi mata kuliah' yang ditulis oleh Fadillah dan Hardiyana , menggunakan *data mining* dengan algoritma *naïve bayes* untuk membuat sistem analisis data yang dapat membantu mahasiswa dalam memilih konsentrasi mata kuliah berdasarkan nilai mata kuliah yang memiliki keterkaitan dengan 2 pilihan konsentrasi mata kuliah yaitu rekayasa sistem informasi dan teknologi informasi. Penelitian yang akan dilakukan memiliki kesamaan dengan penelitian sebelumnya yaitu pengguna *data mining* dalam melakukan analisis data pada nilai akademik mahasiswa. Sedangkan perbedaannya terletak pada jenis analisis data dan algoritma yang terdapat pada *data mining* yang digunakan. Penelitian annisa menggunakan metode klasifikasi algoritma *naïve baiyes* sedangkan pada penelitian yang akan dilakukan menggunakan metode *clustering* dengan algoritma *K-means*.[4]

Penelitian judul 'Penerapan Metode *Clustering* Untuk Pengelompokan Potensi Wisata Di Kabupaten Sumedang', juga menggunakan *data mining* dalam

melakukan analisis data untuk pengelompokan potensi wisata di kabupaten sumedang. Metode *data mining* yang digunakan sama dengan yang dilakukan pada penelitian ini, yaitu *clustering* dengan algoritma *K-means*. Dalam penggunaan *K-means* pada pengelompokan potensi wisata di kabupaten sumedang akan memberikan pengetahuan berupa rekomendasi objek wisata yang dapat dikembangkan pada pembuat keputusan yang dapat meningkatkan jumlah wisata di kabupaten sumedang.[5]

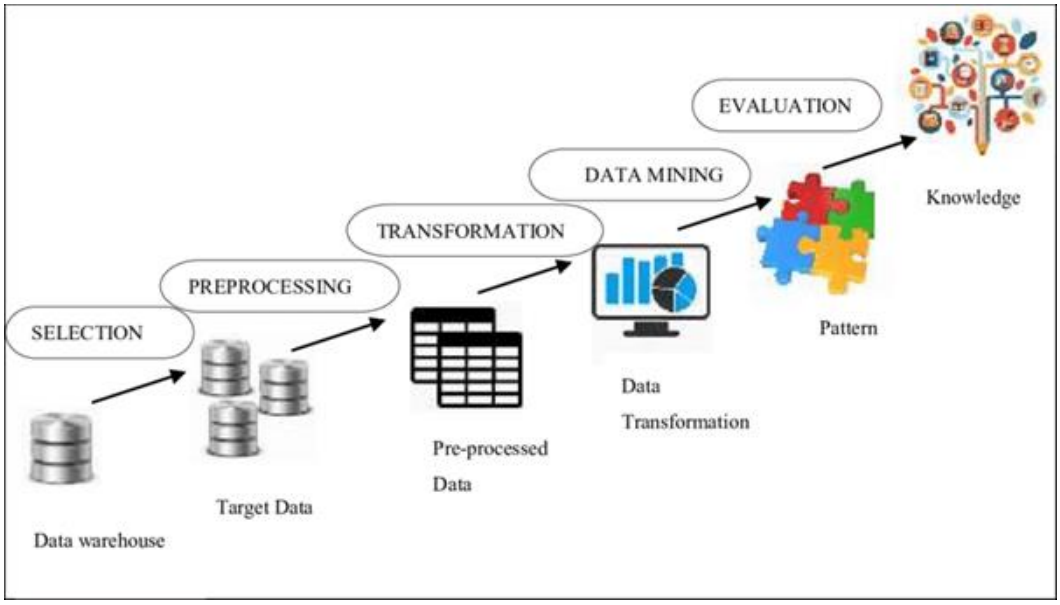
Penelitian Gustientiedina, Adiya, dan Desnelita yang berjudul 'Penerapan Algoritma *K-Means* Untuk *Clustering* Data Obat-Obatan Pada RSUD Pekanbaru' juga menggunkan algoritma *K-Means Clustering* dalam melakukan analisis data obat-obatan di RSUD Pekanbaru. Penggunaan *K-means* dilakukan untuk menganalisa penggunaan obat-obatan, perencanaan, dan pengendalian pasokan obat di rumah sakit dengan mengkluster partisi data dengan karakteristik yang sama dan karakteristik yang tidak sama di kluster yang berbeda. [6]

Dari penelitian Rahayu, Hikmah, Ningsih, dan Fauzan yang berjudul 'Penerapan *K-Means Clustering* Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa' pun melakukan *Clustering* pada data mahasiswa bidik misi menggunakan algoritma *K-means*. Tujuan dari penelitian tersebut adalah untuk membantu mengklasifikasikan mahasiswa layak atau tidak layak dalam mendapatkan beasiswa bidikmisi. *Algoritma K-means* melakukan kluster isasi dengan kluster sangat layak, layak, dipertimbangkan, dan kurang layak pada data mahasiswa berdasarkan 6 variabel, yaitu penghasilan orang tua, keadaan rumah, jumlah tanggungan orang tua dan nilai akademik.[7]

2.2 Studi Pustaka

2.2.1 Data Mining

Pengertian sederhana dari *data mining* adalah penggalian sebuah pola yang menarik atau informasi penting yang belum diketahui sebelumnya dari data yang ada dalam *database*[8]. *Data mining* dilatar belakangi oleh terjadinya ledakan data pada database organisasi-organisasi yang telah mengumpulkan atau melakukan perekaman data yang sudah bertahun-tahun seperti data transaksi, penjualan, pembelian, nasabah dan sebagainya. Agar data yang semakin lama membesar setiap bertambahnya waktu, peneliti merumuskan sebuah metode atau teknik yang dapat memanfaatkan gunungan data dengan mengeksplorasi dan menemukan pola-pola yang menarik dan tersembunyi sehingga dapat disajikan informasi yang berguna untuk pelaku bisnis untuk membuat keputusan bisnisnya. Tidak sampai di situ para peneliti juga mengembangkan metode ini sehingga dapat diterapkan pada bidang lainnya Beberapa jurnal ilmiah, data mining sering diartikan juga sebagai *Knowladge Discovery in Database* (KDD), karena *data mining* terdapat di dalam prosesnya. KDD memiliki beberapa tahapan pada prosesnya dalam mengekstrak informasi atau pola yang menarik dan akan diuraikan sebagai berikut



Gambar 2. 1 *Knowledge Discovery Of Data*

(Sumber: https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/)[9]

A. Data seleksi

Data seleksi adalah memilih atau menyeleksi data yang ada di dalam database atau dari kumpulan data untuk sesuai kebutuhan penggalian data. Data yang diseleksi akan disimpan pada berkes atau format file untuk memisahkan dengan data yang lain di dalam database[10].

B. Pembersihan data

Data yang telah terseleksi akan dicek kembali untuk melihat tidak adanya data yang duplikat, kosong, tidak konsisten, kesalahan data, seperti kesalahan cetak atau *noise* pada data. Kemudian dilakukan perbaikan atau penghapusan agar akurasi pada saat penggalian data meningkat[10].

C. Transformasi data

Transformasi data adalah melakukan penggabungan data atau perubahan data pada nila, format atau bahkan regresi untuk menyesuaikan dengan algoritma *data mining* yang akan dipakai. Beberapa algoritma *data mining* memiliki sensitivitas pada inputan nya sehingga perlu dilakukan perubahan data yang akan diinputkannya. Tahapan ini tergantung dengan algoritma dan pola atau informasi yang dicari dari kumpulan data[10].

D. Data mining

Pada tahapan ini dilakukan analisis data untuk mengekstrak pola atau informasi pada himpunan data yang sebelumnya tidak disadari. Dalam *data mining* terdapat beberapa jenis dan algoritma. Sehingga di tahapan ini dilakukan implementasi algoritma yang tepat sesuai dengan tujuan penggalian data[10].

E. Interpretasi atau Evaluasi

Tahapan terakhir adalah interpretasi atau evaluasi dari hasil algoritma *data mining*. Informasi yang didapat dari algoritma sebelumnya dibuatkan output yang lebih mudah dipahami oleh pihak yang berkepentingan[10].

Data mining merupakan bidang ilmu yang tidak berdiri sendiri, melainkan sangat terkait dengan ilmu statistika, *mechine learning*, visualisasi data, pengenalan pola dan database untuk mengekstrak informasi penting dari data. Fungsi utama dari *data mining* adalah menghasilkan informasi dari kumpulan data, informasi yang dihasilkan bisa berupa deskriptif atau prediksi. Informasi deskriptif artinya adalah informasi yang menjelaskan karakteristik

dari data yang dianalisis, seperti korelasi, kluster, tren, anomali dan teritorial data. Sedangkan prediktif merupakan informasi yang berupa model-model pengetahuan untuk keperluan prediksi. Untuk lebih jelasnya *data mining* memiliki kelompok-kelompok berdasarkan fungsionalitasnya [11]:

A. Klasifikasi dan Prediksi

Klasifikasi dalam *data mining* adalah mencari model atau fungsi yang mendeskripsikan atau perbedaan konsep atau kelas data, untuk tujuan memprediksi kelas dari data yang labelnya tidak diketahui

B. Analisis Asosiasi

Analisis asosiasi merupakan fungsi *data mining* untuk mencari aturan-aturan asosiasi yang menggambarkan atau mendeskripsikan kondisi—kondisi nilai atribut yang kemunculannya relatif banyak dalam himpunan data dibandingkan nilai atribut yang lainnya. Implementasinya biasanya pada analisis tabel transaksi penjualan dengan

C. Konsep/Kelas Deskripsi

Pada jenis pertama ini, *data mining* memiliki fungsi untuk meringkas karakteristik data atau fitur dari target kelas dan membandingkannya dengan satu set kelas yang lain.

D. Analisis *Cluster*

Analisis *Cluster* adalah suatu proses dalam *data mining* yang melakukan pengelompokan suatu objek yang memiliki kesamaan, kedekatan atau kerkaitan

satu sama lain dalam suatu kelompok atau biasa disebut dengan *cluster* dan memiliki perbedaan dengan *cluster* yang satu dengan yang lainnya

2.2.2 Clustering

Clustering atau klusterisasi merupakan salah satu teknik atau metode dalam mengekstrak informasi dari himpunan data. *Clustering* bekerja dengan mengelompokkan objek data berdasarkan tingkat kemiripan antar data dan memiliki tingkat rendah dari masing-masing tiap kluster yang terbentuk[12]. Dalam prosesnya, *clustering* tidak bekerja secara manual. Melainkan menggunakan sebuah algoritma dalam melakukan partisi sebuah data sehingga menghasilkan kelompok atau grup-grup dalam data yang tidak diketahui sebelumnya. Penggunaan klasterisasi memiliki manfaat pada implementasinya adalah sebagai berikut[13]:

1. Klasterisasi data sangat bermanfaat dalam melakukan prediksi dan Analisa bisnis tertentu. Misalnya segmentasi zonasi, marketing, pasar dan lain sebagainya
2. Klasterisasi juga memiliki manfaat dalam identifikasi objek pada berbagai bidang tertentu. sebagai contoh visi komputer dan pengolahan citra.

Konsep dasar dari sebuah kluster isasi adalah menghasilkan kemiripan yang tinggi di dalam satu kluster dan memiliki kemiripan yang rendah antar kluster yang terbentuk atau memiliki tingkat perbedaan yang tinggi antar kluster yang terbentuk dan memiliki tingkat perbedaan yang rendah di dalam

satu kluster. Pada proses kluster isasi, untuk mendapatkan kesamaan yang dimaksud adalah dengan menghitung secara numerik pada buah dua buah objek. Jika kedua buah objek dibantingkan akan menghasilkan tingkat kemiripan yang tinggi dan begitu pun sebaliknya.

Klasterisasi memiliki dua jenis yang berbeda dalam implementasinya, yaitu klasterisasi hierarki dan klasrisasi partisi dan penjelasannya ada di bawah berikut[13]:

a. Klasterisasi Hierarki

   Klastereisasi hierarki adalah mengelompokkan data dengan menggunakan berupa grafik atau diagram yang berbentuk hierarki, dimana setiap iterasi terjadi penyatuan dua kelompok atau pemisahan pada semua set data ke dalam kluster. Step-step pada klasterisasi hierarki adalah sebagai berikut:

   1. Melekukkan identifikasi pada item dalam data set berdasarkan jarak terdekat
   2. Satukan item dalam satu kluster atau kelompok
   3. Hitung jarak setiap kluster atau kelompok
   4. Ulangi kembali step di atas sehingga semua terhubung

b. Klasterisasi Partisi

   Klasterisasi Partisi berbeda dengan klasterisasi hierarki yang mengelompokkan data berdasarkan hierarki. Klasterisasi partisi melakukan pengelompokan data ke dalam kluster- kluster berdasarkan

titik pusat yang telah dipilih dan memiliki tujuan untuk meminimalisir jarak antar data dengan titik pusat masing-masing kluster. Contoh algoritma klasterisasi partisi adalah K-Means, Fuuzzy K-Means, dan Mixture Modeling

2.2.3 K-Means

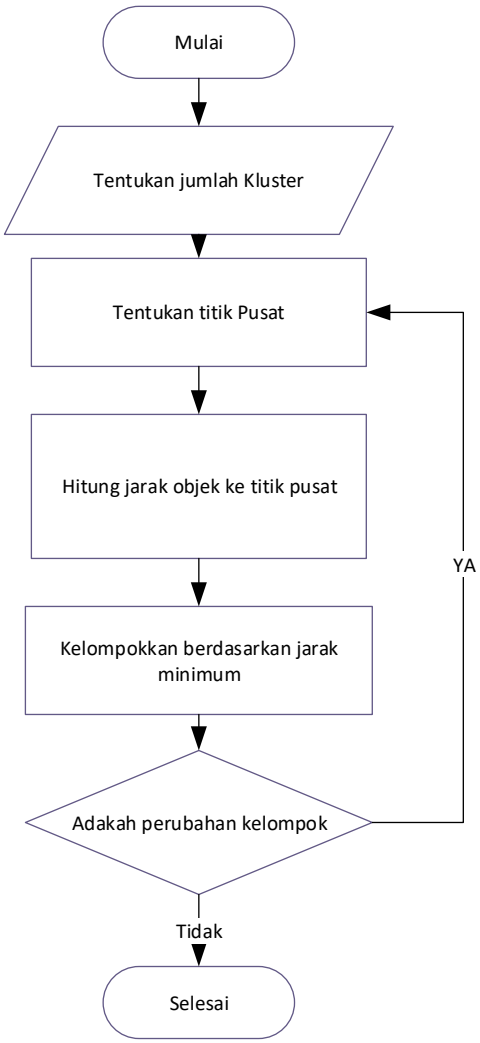
K-Means adalah algoritma *clurtering* yang sangat umum dan sering digunakan karena dalam implementasinya efisien dan mudah dipahami. *K-means* ditemukan oleh J.B. Mac Queen pada tahun 1976 [14]. K-means merupakan algoritma yang masuk ke dalam golongan *Unsupervised Learning*, *Unsupervised Learning* merupakan metode-metode dalam *data mining* yang menggali informasi secara eksplisit dengan menggunakan data yang tidak memiliki label secara implisit. Berbeda dengan *supervised learning* yang menggunakan label pada data untuk menghasilkan pengetahuan atau informasi yang prediktif. Sedangkan *unsupervised learning* menarik informasi deskriptif tanpa label data pada data set. Algoritma K-Means melakukan proses pengelompokan data tanpa mengetahui target kelasnya terlebih dahulu. Input yang dimasukkan hanya objek atau data set dan jumlah klusternya yang biasa disimbolkan dengan *k.* algoritma ini kemudian mengelompokkan objek ke dalam tiap-tiap *k* yang telah ditetapkan[14].

Setiap kluster yang terbentuk terdapat titik pusat (centroid) dan mewakili tiap-tiap kluster. Hal pertama dalam proses algoritma K-Means adalah

menentukan nilai *centroid* pada tiap-tiap kluster dengan cara pemilihan acak atau random untuk menghasilkan pengelompokan *centroid* pertama, yang kemudian akan dilakukan penentuan *centroid* kembali untuk mengoptimalkan posisi *centroid* atau titik pusat kluster[14].

Pada dasarnya algoritma *k-means* mengambil secara random beberapa item yang terdapat pada objek atau data set untuk dijadikan titik pusat kluster awal. Kemudian *K-Means* melakukan perhitungan jarak antar setiap item pada populasi data dengan titik kluster yang telah didefinisikan sebelumnya dan memasukkan item-item pada data ke dalam kluster berdasarkan jarak terdekat dengan titik pusat masing-masing kluster. Setelah itu dilakukan perhitungan kembali untuk menentukan titik pusat baru dan menghasilkan *cluster* baru sampai menghasilkan kluster yang memiliki titik pusat yang optimal.

Berikut adalah diagram alir dari prosedur proses algoritma *Clustering K-Mean[14]s.*



Gambar 2. 2 *Flowchart K-Means*

Tahap 1

Menentukan jumlah kluster (*k*) atau banyaknya kelompok yang akan dihasilkan dari proses *K-Means.*

Tahap 2

Menentukan titik pusat atau *centroid* di setiap kluster yang telah ditentukan sebelumnya. Penentuan titik pusat pada tahap awal dilakukan

dengan cara random. Sedangkan pada tahap selanjutnya atau iterasi selanjutnya ditentukan dengan perhitungan sebagai berikut;

$\overline{v_{ij}} = \frac{1}{N}\sum_{k=0}^{N_i} x_{kj}$ …… persamaan (1)

Di mana :

$\overline{v_{ij}}$ = *centroid*/rata-rata cluster ke-I untuk variabel ke-j

N_i = jumlah data yang menjadi anggota *cluster* ke-i

i, k = indeks dari kluster

j = indeks dari variabel

x_{kj} = nilai data ke-k yang ada di dalam kluster tersebut untuk variabel ke-J

Tahap 3

Menghitung jarak terdekat titik pusat dengan tiap-tiap *record* pada data set menggunakan rumus *Euclidean Distance* sebagai berikut:

$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$ …… persamaan (2)

Dimana :

D_e = *Euclidean Distance*

i = banyak sampel

x, y = koordinat objek

s, t = koordinat *centroid*

Tahap 4

Setelah hasil perhitungan menggunakan rumus *Euclidean Distance* didapat maka selanjutnya adalah mengelompokkan data berdasarkan nilai terkecil.

Tahap 5

Mengulangi kembali pada tahap ke-2 hingga tahap ke-5 sampai titik pusat berada pada titik optimal atau tidak adanya perubahan anggota kelompok atau *cluster*.