# BAB II

# TINJAUAN PUSTAKA

## 2.1 Penelitian Terkait

Penelitian yang merujuk pada Aditia Yudhistira dan Rio Andika pada tahun 2023 dengan judul "Pengelompokan Data Nilai Siswa Menggunakan Metode *K-Means Clustering*" meneliti tentang pengelompokan data nilai siswa dari 3 kategori penilaian yaitu nilai siswa, disiplin, serta sikap. Dataset yang digunakan sebagai uji coba adalah data nilai semester siswa SMA XYZ kelas x tahun ajaran 2022 sebanyak 155 data peserta didik yang diklasterisasi menggunakan *phyton* dan menghasilkan 3 *cluster*. Dari hasil pengujian menggunakan *silhouette coefficient* dimana jumlah *cluster* yang terbaik ditunjukan dengan nilai *Silhoutte* yang semakin mendekati 1, didapatkan nilai *silhouette coefficient* dari 3 *cluster* adalah 0.489, dan lebih baik dari nilai *silhouette coefficient* lainnya.

Penelitian yang merujuk pada Dwiki Aulia Fakhri, Sarjon Defit, dan Sumijan pada tahun 2021 dengan judul "Optimalisasi Pelayanan Perpustakaan terhadap Minat Baca Menggunakan Metode *K-Means Clustering*" meneliti tentang pelayanan perpustakaan terhadap minat baca dengan kriteria jumlah ketersediaan buku, buku yang di pinjam, dan lama buku dipinjam. Data yang digunakan adalah data dari 40 sampel buku, dari 40 buku ini terdapat jumlah pinjaman sebanyak 166 pinjaman dan waktu peminjaman 434 hari. Data buku di *clustering* menjadi 3 yaitu sangan diminati, diminati, dan kurang diminati. Hasil dari klasterisasi ini bisa dijadikan sebagai acuan rekomendasi untuk pengoptimalisasian pelayanan perpustakaan baik untuk tata letak maupun pengadaan buku dengan memprioritaskan jenis buku yang sangat diminati.

Penelitian yang merujuk pada Taufik Hidayat pada tarhun 2022 dengan judul "Klasifikasi Data Jamaah Umroh Menggunakan Metode *K-Means Clustering*" meneliti tentang pengelompokan data jamaah umroh

menggunakan 4 variabel yaitu, nama, jenis kelamin, usia, dan daerah asal. *Cluster* yang digunakan ada 3 *cluster*, yaitu sangat disukai, disukai, and kurang disukai. Data yang digunakan berjumlah 170 data dan menghasilkan perhitungan serta pengetahuan baru yang dapat memudahkan manager pada AET Travel Indonesia.

**Tabel 2.1** Penelitian Terdahulu

| No. | Judul Penelitian | Metode | Tujuan | Hasil |
|---|---|---|---|---|
| 1 | Pengelompokan Data Nilai Siswa Menggunakan Metode *K-Means Clustering* | *K-Means Clustering* | Tujuan penelitian ini yaitu mengetahui dan membentuk clus ter data siswa berdasarkan nilai akademik, nilai sikap, dan nilai disiplin sehingga menjadi sebuah *cluster* sehingga hasil *cluster* siswa dapat menjadi acuan dalam meningkatkan nilai siswa dalam proses pembelajaran selanjutnya | Hasil pembahasan pengelompoka n data nilai siswa menggunakan metode K-Means *clustering* menunjukan bahwa berdasarkan hasil *cluster* data siswa menggunakan dataset siswa dalam satu semester, maka didapatkan *cluster* 0 berjumlah 59 siswa, *cluster* 1 berjumlah 94 siswa, dan *cluster* 2 |

| | | | | |
|---|---|---|---|---|
| | | | | berjumlah 1 siswa |
| 2 | Optimalisasi Pelayanan Perpustakaan terhadap Minat Baca Menggunakan Metode *K-Means Clustering* | *K-Means Clustering* | Penelitian ini bertujuan untuk mengoptimalisasikan penempatan buku sehingga siswa bisa dengan cepat mencari buku sesuai dengan minat bacanya dengan lebih efektif dan bisa tertarik dengan buku yang lain karena berada dalam satu pengelompokan. Sedangkan untuk pihak perpustakaan bisa memprioritaskan untuk pengadaan buku selanjutnya | Penelitian ini dapat dijadikan sebagai acuan rekomendasi untuk pengoptimalisasikan pelayanan perpustakaan baik untuk tata letak maupun pengadaan buku dengan memprioritaskan jenis buku yang sangat diminati. |
| 3 | Klasifikasi Data Jamaah Umroh Menggunakan Metode *K-Means Clustering* | *K-Means Clustering* | Untuk membantu pihak perusahaan dapat mengelompokkan jamaah umroh berdasarkan data yang telah diolah. Sehingga pihak manajer dapat | Di dapatkan hasil pengelompokan ini maka di dapatkan pengetahuan baru yang dapat memudahkan |

| | | | | |
|---|---|---|---|---|
| | | | membuat paket umroh yang sesuai dengan keinginan konsumen | maneger dalam strategi pemasaran pada AET Travel Indonesia |
| 4 | Penerapan *Data mining* dalam Perancangan Sistem Pendukung Keputusan Seleksi Penerimaan Beasiswa Menggunakan Naive Bayes Classifier (Studi Kasus: IIB Darmajaya) 2020 | *Naive Bayes* | Tujuan dari pelaksanaan penelitian ini adalah membangun suatu sistem yang akan digunakan guna mendukung proses pengambilan keputusan dengan menerapkan Teknik datamingung memanfaatkan algoritma Naïve Bayes Classiffier agar dapat membantu pihak IIB Darmajaya khususnya Unit Kemahasiswaan dalam penagambilan keputusan dalam penyeleksianpenerimaan beasiswa prestasi. | Hasil dari implementasi sistem ini ialah memberikan keterangan tentang informasi penerima beasiswa berdasarkan rengking yang dapat digunakan sebagai alat bantu dalam proses pengambilan keputusan. Dengan adanya sistem ini, proses perhitungan untuk |

| | | | | |
|---|---|---|---|---|
| | | | | menentukan penerima beasiswa dapat dilakukan dengan mudah, cepat dan akurat. |
| 5 | Pengelomp Okkan Penyakit Pasien Mengguna Kan Algoritma K-Means (2022) | *K-Means Clustering* | Untuk mempermud ah proses pengelolaan data yang banyak, Puskesmas Bahorok memerlukan suatu sistem dalam mengambil keputusan untuk mengetahui pengelompok kan penyakit berdasarkan usia pasien yang sering terkena penyakit pada Puskesmas | Pengelompoka n dengan metode KMeans dapat menghasilkan jumlah *cluster* yang sama dengan jumlah data yang berbeda – beda tanpa harus memiliki data yang sama. Dengan dibangunnya sistem ini untuk mempermudah kan user dalam mengelompok an |

| | | | | |
|---|---|---|---|---|
| | | | Bahorok. | penyakit pada pasien berdasarkan usia secara efektif dan efisien khususnya untuk Staff Pegawai dan Administrasi. Dengan metode KMeans sangatlah mempermudah kan user dalam mengelompok an suatu data hanya dengan memiliki karakteristik yang sama. |
| 6 | *Data mining* Dengan Algoritma Neural Network Dan Visualisasi Data Untuk Prediksi Kelulusan Mahasiswa | *Neural Network* | Algoritma neural network digunakan untuk memprediksi kelulusan mahasiswa yang sulit dilakukan | Hasil penelitian yang dilakukan dari tahap[1] |

| | | | | |
|---|---|---|---|---|
| | (2020) | | secara manual, sedangkan visualisasi digunakan untuk menampilkan data rekapitulasi secara visual sehingga lebih menarik dan mudah dipahami. | awal sampai dengan tahap pengujian menggunakan *Data mining* dengan algoritma Neural Network untuk kelulusan mahasiswa menghasilkan prediksi dengan Precision 87.80%, Recall 86.90% dan nilai akurasi sebesar 92.83%. Sedangkan dari visualisasi data dari dataset yang berjumlah 2742 record menampilkan beberapa rekapitulasi |

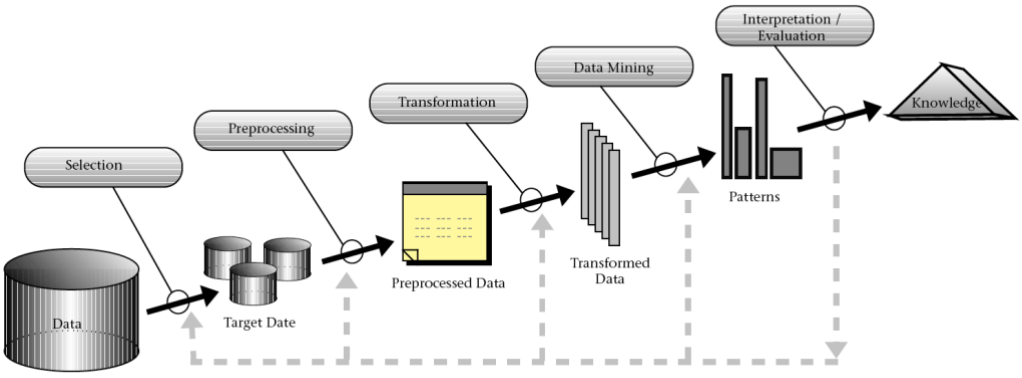
| | | | | |
|---|---|---|---|---|
| | | | | pelaporan berupa dashboard yang sangat komplit, sehingga dengan prediksi dan visualisasi data tersebut dapat membantu dalam kelulusan mahasiswa dan memberikan rekomendasi tindakan yang tepat dan harus dilakukan oleh manajemen atau pihak yang berwenang untuk mengambil keputusan. |

| | | | | |
|---|---|---|---|---|
| 7 | Komparasi Metode Apriori dan FP-Growth *Data mining* Untuk Mengetahui Pola Penjualan (2023) | KDD *(Knowledge Discovery in Database)* | Tujuan penelitian ini adalah untuk mengetahui pola penjualan produk terlaris dan untuk meningkatkan kuantitas penjualan produk parfum [2] | Penelitian ini menghasilkan pola frekuensi tinggi untuk itemsets dengan nilai minimum support 20% menghasilkan produk yang menjadi The Most Tree Items adalah Jo Malone 82,49%, Zarra 28,25%, dan Zwitsal 20,34%. Sedangkan aturan asosiasi yang terbentuk dari nilai Min. Supp 20% dan Min. Conf 80%, mendapatkan kombinasi 2 itemsets yaitu Jo Malone dan Zarra. |

| | | | | |
|---|---|---|---|---|
| | | | | Sedangkan untuk kombinasi 3 itemsets yaitu Jo Malone, Zarra dan Baccarte dengan status valid dan kuat dibuktikan dengan nilai lift lebih besar dari 1, oleh karena itu aturan asosiasi tersebut sangat tepat untuk dapat digunakan. |

## 2.2 Tahapan Dalam *Data mining*

*Knowledge Discovery in Database* (KDD) suatu teknik pembentukan pola atau rule dalam informasi. Informasi yang dihasilkan didapatkan dari suatu data yang besar atau dikenal dengan tambang data yang disimpan dalam basis data yang awalnya belum diketahui dan menghasilkan suatu data yang potensial bermanfaat. Iterasi dalam *data mining* disebut proses KDD [3]. Berikut langkah-langkah dari KDD:




**Gambar 2.1 Tahapan KDD**

1. *Selection*/Seleksi data yaitu sekumpulan data harus melalui tahap KDD. Data yang dipakai harus melalui proses *data mining* dengan penyeleksian data dari kumpulan data, kemudian data di simpan dalam *database* secara terpisah dan disimpan dalam suatu file/berkas.
2. *Preprocessing*/Pembersihan data; tahap ini dilakukan pembersihan data atau membuang data yang tidak konsisten dan memperbaiki data yang salah serta melakukan pemeriksaan data yang tidak konsisten, data reududansi, duplikasi data, sehingga menghasilkan informasi yang relevan dan berguna.
3. *Transformation*/Trnasformasi data yaitu data diubah menjadi bentuk yang sesuai dengan yang di mining. Pengkodean merupakaan proses *data mining* yang akan dicari dalam *database*.
4. *Data mining* merupakan teknik atau metode dalam pembentukan pola sehingga mengasilkan informasi yang dibutuhkan. pola yang ditemukan; dari pola yang dihasilkan maka didapatkanlah suatu informasi yang mudah dipahami oleh pengguna dan menjadi sumber pengetahuan dalam pengambilan keputusan.
5. *Interpretation/Evaluation* adalah pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*.

## 2.3 Landasan Teori

### 2.2.1 *Data mining*

*Data mining* adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Pengolahan data menjadi informasi/pola/pengetahuan yang berguna dibutuhkan peranan *data mining* di dalamnya. Secara umum terdapat 5 peranan dalam *data mining*, yaitu estimasi, prediksi, klasifikasi, klustering, dan asosiasi (Suntoro,2019).

**Tabel 2.2** Perbedaan Peranan *Data mining*

| Jenis | Atribut | Kelas/Label/Target | Keterangan |
|---|---|---|---|
| Estimasi | Numerik | Numerik | |
| Prediksi | Numerik | Numerik | Rentang waktu |
| Klasifikasi | Numerik/Kategorial | Numerik/Kategorial | |
| Klustering | Numerik | - | |
| Asosiasi | - | - | Hubungan antar atribut |

Dari kelima jenis didalam tabel diatas, dapat dijelaskan sebagai berikut :

**1. Deskriptif**

Deskriptif lebih kepada merujuk ke fungsi dalam pemahaman data, hal ini membuat pengolahnya bisa teliti lebih mendalam. Tujuan dari proses ini adalah untuk menemukan pola dan karakteristik yang terdapat pada data. Fungsi deskriptif dimanfaatkan sebagai *pattern* tertentu yang awalnya tidak terlihat di dalam data.

**2. Prediktif**

Fungsi terkait dengan proses yang nantinya digunakan untuk mengetahui pola khusus dari data yang digunakan. Pola ini bisa ditemukan dari beberapa variabel dalam data,

ketika sudah menemukan pola, maka pola yang digunakan dipakai untuk memperkirakan variabel lain dan masih belum diketahui nilainya karena itu disebut fungsi prediktif.

**3. Klasifikasi**

Klasifikasi dilakukan untuk sebagai cara menyimpulkan beberapa pengertian karakteristik dari suatu grup atau kelompok data. Seperti data pelanggan yang tak lagi menggunakan produk, karena menganggap produk kompetitor memberi manfaat lebih banyak dan customer *value* untuk para pelanggan.

**4. *Clustering***

*Clustering* merupakan fungsi yang mengarah pada proses identifikasi kelompok, kemudian produk atau barang yang memiliki karakteristik khusus. Biasanya digunakan dalam mengetahui kelompok-kelompok tertentu dalam penyebarannya.

**5. Asosiasi**

Fungsi asosiasi adalah fungsi *data mining* yang dapat diproses untuk melakukan identifikasi relasi atau hubungan dari setiap data yang ada. Data ini bisa merupakan data dahulu maupun data yang didapat saat ini.

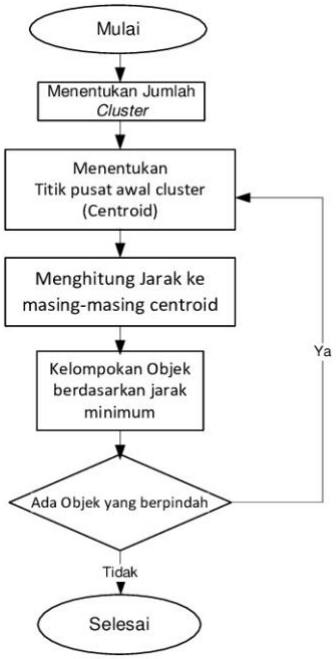
### 2.2.2 *Clustering*

*Clustering* merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*. Objek yang di dalam *cluster* memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*. Oleh karena itu, *clustering* sangat berguna dan bisa menemukan group atau kelompok yang tidak dikenal dalam data. *Clustering* banyak digunakan dalam berbagai aplikasi seperti misalnya pada business inteligence, pengenalan pola citra, *web search*, bidang ilmu biologi, dan untuk keamanan (*security*). Di dalam business inteligence, *clustering* bisa mengatur banyak customer ke dalam banyaknya kelompok (Karin Annisa, Budi Serasi Ginting, Mili Alfhi Syari, 2022). Dalam *clustering* dikenal empat tipe data. Keempat tipe data tersebut adalah:

a. Variabel berskala interval
b. Variabel biner
c. Variabel nominal, ordinal, dan rasio
d. Variabel dengan tipe lainnya

### 2.2.3 *K-Means Clustering*

Algoritma *K-Means* merupakan algoritma yang paling banyak digunakan dalam berbagai penerapan karena mudah untuk diimplementasikan. *K-Means* juga metode pengelompokan sederhana yang mengelompokkan data ke dalam k kelompok berdasar *centroid* masing-masing kelompok. Berikut adalah tahapan Algoritma *K-Means Clustering*:[4]




Gambar 2.2 Algoritma *K-Means Clustering*

1. Pilih jumlah *cluster* (k) yang ingin dibuat
2. Pilih titik data k secara acak dan letakkan tiap titik ke *cluster* sebagai *centroid* (pusat *cluster*) awal.

3. Kelompokkan semua titik data sesuai dengan jarak *centroid* terdekat yang telah dibuat.
4. Hitung varians dan tempatkan *centroid* baru dari setiap *cluster*, jika data yang digunakan memiliki dimensi lebih dari satu, maka untuk menghitung jaraknya dapat menggunakan *euclidean distance*, dimana rumusnya adalah :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Keterangan :

*i :* index dari atribut

*n :* jumlah data

$x_i$ *:* atribut dari data ke-i, (i=1,2,3,…,…,n)

$y_i$ *:* atribut dari pusat *cluster* ke-i, (i=1,2,3,…,…,n)

5. Ulangi Langkah ketiga, yang berarti menetappkan Kembali semua titik data ke *centroid* terdekat baru dari setiap *cluster*.

### 2.2.4 *Fitness*

*Fitness* adalah aktivitas fisik yang dapat membuat orang menjadi lebih bugar dengan olahraga latihan angkat beban atau *weight lifting*, aerobic, dan pemenuhan nutrisi. *Fitness* adalah olahraga yang paling efektif untuk membentuk *body image* dan menurunkan berat badan agar terlihat lebih menarik. *Fitness* sendiri adalah olahraga yang difokuskan untuk membakar lemak, dan pembentukan bagian otot tubuh yang diinginkan (Laras Apriliani, Julianti Kasih, 2021). Terdapat berbagai macam manfaat *fitness*s yang bisa didapatkan bila dilakukan secara rutin.

### 2.2.5 *Personal Trainer*

*Personal Trainer* adalah seseorang yang bertugas untuk memandu atau mengajarkan seseorang yang berlangganan di tempat *fitness* sebagai sarana olahraga dalam waktu tertentu (Laras Apriliani, Julianti Kasih, 2021). Personal trainer memiliki peran untuk membuat program latihan yang paling tepat, sehingga member *gym* dapat lebih fokus

dalam mengikuti program latiihan tersebut, dengan mengikuti program latihan yang diberikan oleh *personal trainer*, member akan lebih mudah dalam mencapai target yang diinginkan dengan lebih teratur. [5]

### 2.2.6 *RapidMiner*

*RapidMiner* adalah *software* yang dapat diakses oleh siapa saja dan bersifat terbuka (*open source*). *RapidMiner* ini dijadikan sebuah solusi untuk menganalisa terhadap data *processing*. Pada *RapidMiner* ini digunakan berbagai teknik seperti teknik deskriptif dan prediksi. *RapidMiner* merupakan mesin pengolahan atau penambangan data yang dapat diintegrasikan ke dalam produknya sendiri dan tersedia sebagai perangkat lunak mandiri untuk analisis data [6].

### 2.2.7 *Tableau*

Tableau adalah sebuah *tools* yang dapat mempermudah pembuatan analisis visual interaktif dalam bentuk *dashboard*. Dapat disimpulkan bahwa *Tableau* adalah *software* yang bisa mengolah data menjadi sebuah visual yang menarik. Dengan begitu, kumpulan data tersebut akan lebih mudah dimengerti [7].

### 2.2.8 *David Bouldin Index* (DBI)

*Davies Bouldin Index* (DBI) adalah parameter yang dijumpai pada tahun 1979 oleh Donal W. Bouldin dan David L. Davies. DBI digunakan untuk menggambarkan algoritma K-Means. Semakin mendekati nilai 0 maka skema *clustering* tersebut akan semakin ideal. Pendekatan pengukuran ini bertujuan untuk memaksimalkan jarak antara klaster yang satu dengan yang lainnya dan pada waktu yang sama mencoba untuk meminimalkan jarak antara objek dalam sebuah klaster [8].