BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Luthfi dan Wijayanto (2021) telah melakukan penelitian menggunakan metode *Hierarchical*, *K-Means*, dan *K-Medoids Clustering* yang kemudian dibandingkan dengan evaluasi *Dunn Index, Davies Bouldin Index (DBI),* dan *Calinski Harabasz Index (CHI)*. Penelitian tersebut menghasilkan model terbaik yaitu menggunakan *K-Medoids* yang lebih baik dilihat dari perbandingan rasio simpangan baku yang diaplikasikan dengan analisis sentiment wilayah kabupaten / kota di wilayah Indonesia berdasarkan angka IPM masing-masing wilayahnya sehingga didapatkan wilayah dengan angka IPM tertinggi hingga angka IPM terendah pada tahun 2019.

Khairati et al. (2019) menggunakan algoritma *K-Means, K-Means Enhanced*, dan *K-Means Maximum Minimum Criterion Algorithm* untuk melakukan penelitiannya dengan *Silhouette Index, Davies Bouldin Index (DBI), Dunn Index,* dan *Calinski Harabasz* sebagai evaluasinya. Algoritma tersebut diimplementasikan untuk data *benchmark* yaitu data *Iris*, *Ruspini*, *Seeds*, dan *Wine*. Dari hasil simulasi keempat evaluasi dapat diprediksi jumlah *cluster* optimal jika menggunakan evaluasi *Dunn*.

Ramadanti dan Muslih (2021) menyatakan bahwa penelitian mengenai analisis persebaran kasus Covid-19 di Jawa Barat menggunakan metode *K-Means Clustering* yang dibagi menjadi 3 *cluster*, yaitu tinggi, sedang, rendah. Hasil penelitian ini didapatkan *cluster* tinggi terdapat 2 kabupaten / kota, *cluster* sedang terdapat 6 kabupaten / kota, *cluster* rendah 19 kabupaten / kota di Jawa Barat.

Aiman Ayadi et al. (2020) melakukan penelitian yang menghasilkan perbandingan tingkat performa metode *K-Means* dan *Hierarchical clustering* pada sistem rekomendasi pemilihan kost dan dioptimalkan dengan *Naïve Bayes*. Didapatkan hasil akhir penelitian ini yaitu nilai akurasi *K-Means* dan

Naïve Bayes lebih tinggi dengan akurasi 90,82%, presisi 90,56%, *recall* 90,68%, dan waktu lebih lama yaitu 10 detik, sedangkan untuk nilai *Hierarchical dan Naïve Bayes* mendapatkan nilai akurasi 88,02%, presisi 87,82%, *recall* 88,00%, dan waktu lebih cepat yaitu 7,6 detik.

Sari dan Yunita (2021) menyatakan dalam penelitiannya yang membahas tingkat keparahan dan risiko penyebaran Covid-19 di Indonesia dengan menggunakan *K-Means Clustering*. Hasil dari penelitian ini bahwa DKI Jakarta dan Jawa Timur mempunyai tingkat keparahan terhadap Covid-19 tertinggi di Indonesia.

Ariawan et al. (2020) telah melakukan penelitian tentang *clustering* data remunerasi PNS dengan menggunakan metode *Local Outlier Factor* dan *K-Means Clustering* serta mengimprovisasi pada tahap *pre-processing* dan penentuan jumlah *cluster* optimal. Metode *Local Outlier Factor* dengan nilai MinPts 150 dapat mendeteksi data *outlier* paling banyak dengan jumlah data terdeteksi *outlier* sebanyak 162 data atau sebesar 22,98%. Jumlah *cluster* optimal dengan metode *elbow* sejumlah 4 *cluster* dengan nilai *Silhouette Index* sebesar 0,542, *Dunn Index* sebesar 0,040, dan *Purity* sebesar 0,89.

(Sinambela et al., 2020) melakukan penelitian mengenai perbandingan konseptual dan kinerja *K-Means* dan algoritma *Fuzzy C Means* untuk pengelompokkan data mining segmentasi konsumen. Atribut yang digunakan untuk proses *mining* pada segmentasi konsumen yaitu data konsumen, produk, demografi, perilaku konsumen, transaksi, RFMDC, *RFM (Referency, Frequency, Monetery)*, dan *LTV (Life Time Value).* Algoritma *clustering* dengan algoritma *classification, association*, dan CPV kemudian digabungkan untuk mendapatkan nilai potensial tiap *cluster*.

B. Landasan Teori

1. Data Mining

Data mining terdiri dari dua suku kata. Data yaitu sekumpulan bahan, alat, teks yang belum mempunyai arti yang bersifat tunggal. *Mining* yaitu proses penggalian informasi. *Data mining* merupakan proses yang

menggunakan teknik statistik, kecerdasan buatan, dan *machine learning* untuk mengidentifikasi dan mengekstrasi informasi yang terkait dengan *database* besar yang membantu dalam pengambilan keputusan (Anas 2020).

Data mining dapat dilakukan dengan mengikuti langkah-langkah sebagai berikut (Yuli Mardi, 2019) :

a. *Data Reduction*

   Data reduction atau pemilihan data perlu dilakukan sebelum perhitungan dilanjutkan ke tahap selanjutnya. Data yang sudah diseleksi kemudian dipisahkan dari data mentah atau data asli.

b. *Data Cleaning*

   Data yang sudah melalui proses data reduction atau pemilihan data, kemudian data diproses pada *data cleaning* yang bertujuan untuk memperbaiki penulisan data. Penulisan data yang tidak tepat meliputi penulisan tanda baca yang tidak sesuai, penulisan ganda, serta format penulisan yang tidak sesuai.

c. *Data Transformation*

   Data transformation yaitu proses mengubah data dengan mengubah skala atau pengkodean data dalam bentuk lain untuk menyetarakan isi data. Proses *transformasi* data dilakukan dengan melakukan normalisasi menggunakan *Min-Max Normalization.*

2. Split Data

Split data merupakan teknik membagi data asli menjadi dua bagian secara acak, yang terdiri dari data *training* dan data *testing*. Data *training* merupakan data yang akan dipakai dalam pembelajaran, sedangkan data *testing* merupakan data yang akan digunakan dalam perhitungan penelitian, atau sebagai pengujian kebenaran atau keakurasian hasil pembelajaran (Muningsih, 2022).

3. Clustering

Clustering atau pengelompokkan merupakan proses mengelompokkan data atau *record* kedalam *cluster* atau kelompok yang telah ditentukan. Suatu data dikelompokkan dalam *cluster* berdasarkan tingkat kemiripan atribut yang dimiliki, dan berbeda dengan *record* atau data di *cluster* lain (Hidayati et al. 2021).

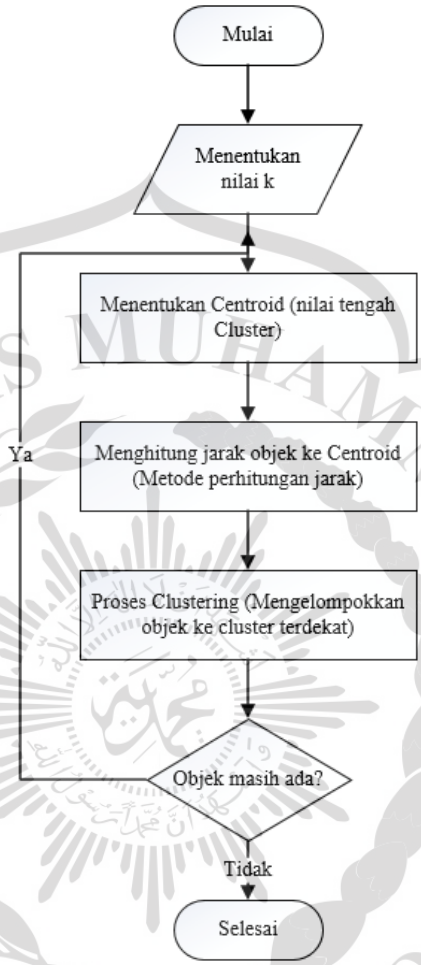
Metode *clustering* terdiri dari dua bagian, *hirarki* dan *patritional*. Metode *clustering hirarki* yaitu sebuah metode pengelompokkan data dengan pusat *cluster* yang telah ditentukan secara berurutan, dengan tujuan menampilkan data yang memiliki kemiripan, sedangkan *partitional* atau partisi merupakan metode menentukan *cluster* dalam waktu tertentu untuk mengelompokkan data menjadi beberapa kelompok yang lebih spesifik, dimana setiap *cluster* memiliki *centroid* dan menghitung setiap data yang memiliki kemiripan atau kedekatan dengan pusat *cluster* yang sudah ditentukan (Madhulatha, 2012).

4. K-Means

K-Means merupakan salah satu algoritma *clustering* yang termasuk dalam *unsupervised learning* dan berfungsi untuk mengelompokkan data kedalam beberapa *cluster* dengan sistem partisi. *Unsupervised learning* merupakan algoritma *data mining* yang digunakan untuk mencari pola dari variabel atau atribut, variabel tersebut tidak memiliki label (Zulfa et al. 2021). Sari et al. (2018) menyatakan bahwa algoritma ini mampu meminimalkan jarak antara data ke *cluster*-nya. Ada 2 aturan yang ada pada algoritma *k-means* sebagai berikut:

1. Banyaknya *cluster* yang perlu dimasukkan.
2. Hanya memiliki atribut bertipe *numeric*

Bentuk diagram *Flowchart* alur proses algoritma *K-Means* terdapat pada gambar 2.1 (Primartha, 2018):



Gambar 2. 1 *Flowchart* diagram algoritma *K-Means*

Tahapan-tahapan dalam algoritma *k-means* dijelaskan sebagai berikut:

1. Menentukan jumlah *cluster* k dari dataset yang akan dibagi.
2. Menentukan data k yang menjadi titik pusat atau *centroid* awal lokasi klaster.
3. Mengelompokkan data ke dalam k *cluster* sesuai titik *centroid* terdekat yang sudah ditentukan.
4. Memperbarui nilai titik *centroid* dan mengulangi langkah ke 3 sampai *centroid konvergen* atau tidak berubah.

5. Davies Bouldin Index (DBI)

Menurut Jollyta et al. (2019) *Davies Bouldin Index (DBI)* yaitu fungsi rasio dari jumlah *cluster scatter* sampai *cluster separation*. Pengukuran evaluasi DBI bertujuan untuk memaksimalkan jarak antar *cluster*. Dengan menggunakan perhitungan DBI, sebuah *cluster* dinyatakan optimal jika memiliki nilai minimal atau mendekati angka 0 (Religia dan Sunge 2019).

6. Silhouette Index

Secara umum, evaluasi *Silhouette* menghitung rata-rata nilai setiap titik pada himpunan data, atau bisa dijelaskan bahwa *Silhouette index* merupakan perhitungan nilai setiap titik merupakan selisih dari *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya. Jumlah *cluster* dikatakan optimum jika hasil mendekati angka 1 (Khairati et al., 2019).

7. Calinski Harabasz Index (CH)

Evaluasi *Calinski Harabasz* menghitung perbandingan antara nilai *SSB (Sum of Square Between cluster)* sebagai separation dan nilai *SSW (Sum of Square Within cluster)* sebagai *compactness* yang dikalikan dengan *factor* normalisasi, yaitu selisih jumlah data dengan *cluster* dibagi jumlah *cluster* dikurang satu. *Cluster* dikatakan baik atau optimum ditunjukkan dengan semakin besar nilai CH (Khairati et al., 2019).

8. Python

Python merupakan Bahasa pemrograman yang di kenalkan pertama kali oleh Guido van Rossum di *Scitchting Mathematisch Centrum Belanda* pada tahun 1991. Kelebihan python beraneka ragam pada aspek *readability*, multifungsi, interoperabilitas dan dukungan komunitas yang memadai. Versi dari *python* menggunakan lisensi *GFL-Compatible* yang bersifat *open source* (Wahyono, 2018).