BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian yang dilakukan Agusnady (2021), menerapkan metode *Principal Component Analysis* (PCA) dalam metode K-Means *clustering* sebagai pemberian bobot terhadap atribut data di *dataset water quality* dengan pengukuran menggunkan *Sum of Square Error* (SSE) menghasilkan nilai derajat *error* yang lebih rendah sebesar 11.80 sedangkan nilai derajat *error* tanpa PCA (metode K-Means dengan penentuan nilai *centroid* secara acak) menghasilkan nilai sebesar 13.57.

Penelitian yang dilakukan Panggabean (2021), menerapkan metode *Rank Order Centroid* (ROC) sebagai pembobotan di metode *Simple Addivite Weighting* (SAW) dalam pemberian *reward* bagi pegawai honorer menghasilkan perangkingan yang lebih efektif dan menjadi lebih objektif dalam pemberian *reward* pegawai honorer.

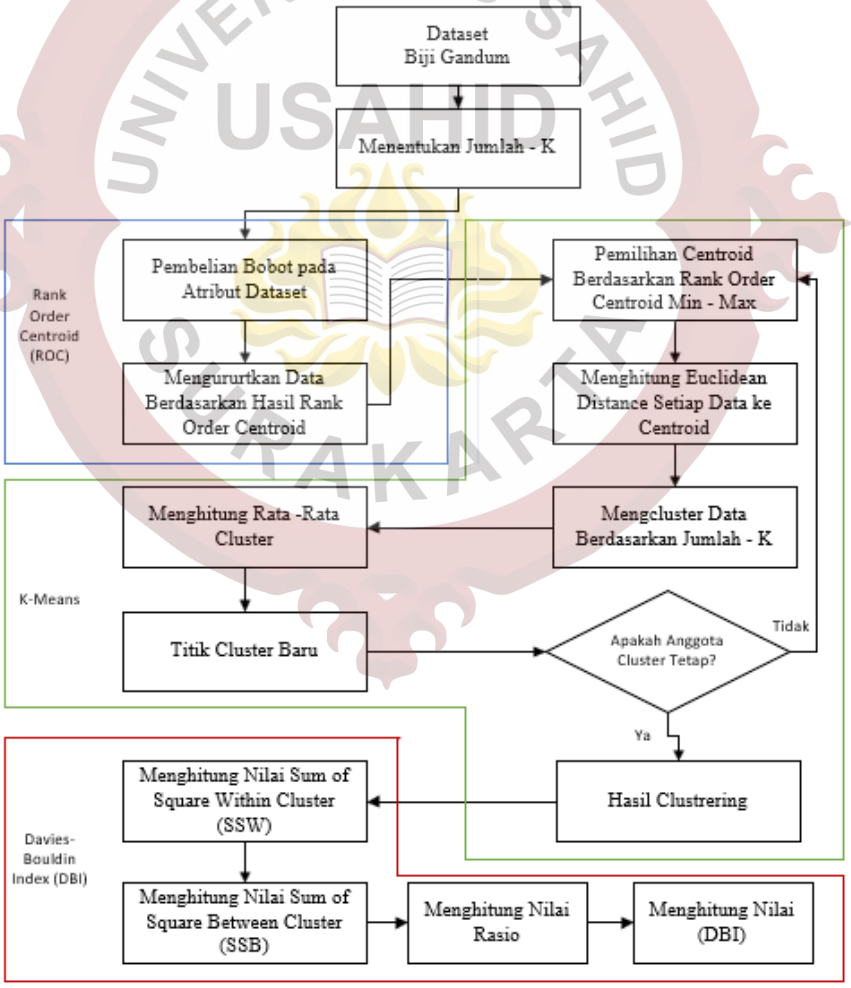
Penelitian yang dilakukan Fitriyah (2020), menerapkan metode K-Means sebagai metode analisis kedisiplinan Pegawai Negeri Sipil (PNS) Kecamatan Tingkir dari tingkat presensi menggunakan aplikasi Rapidminier Studio dengan pengujian parameter k=2 sampai k=10 dan diperoleh hasil pengujian yang terbaik dengan parameter k=6. Analisis pengujian menggunakan parameter k=6 menghasilkan pegawai dengan tingkat kedisiplinan sangat kurang baik di *cluster* 0 dengan jumlah data 178, pegawai dengan tingkat kedisiplinan sangat baik di *cluster* 1 dengan jumlah data 204, pegawai dengan tingkat kedisiplinan perlu ditingkatkan di *cluster* 2 dengan jumlah data 186, pegawai dengan tingkat kedisiplinan baik di *cluster* 3 dengan jumlah data 96, pegawai dengan tingkat kedisiplinan perlu ditingkatkan di *cluster* 4 dengan jumlah data 10, pegawai dengan tingkat kedisiplinan sangat baik di *cluster* 5 dengan jumlah data 238.

Penelitian yang dilakukan Hamid (2019), menerapkan *Davies-Bouldin Index* sebagai evaluasi hasil c*lustering* dalam penentuan karyawan

tetap di PT Pyojoon Mold Indonesia dengan menggunakan metode algoritma K-Means menghasilkan kategori yang sangat baik dengan angka DBI sebesar 0.086 dengan jumlah *cluster* 2 yaitu c*luster* 0 dengan 340 data karyawan sebagai karyawan yang lulus dan *cluster* 1 dengan 164 data karyawan sebagai karyawan yang tidak lulus.

2.2 Kerangka Pemikiran

Pada penelitian ada urutan kerangka pemikiran yang harus diikuti, urutan kerangka pemikiran ini merupakan gambaran dari langkah–langkah yang harus dilalui agar penelitian ini bisa berjalan dengan baik. Kerangka pemikiran yang harus diikuti bisa dilihat pada Gambar 2.1 :



Gambar 2.1 Diagram Kerangka Pemikiran

Gambar 2.1 menunjukkan bahwa proses awal dengan memasukan *dataset* biji gandum, setelah *dataset* dimasukan proses selanjutnya menentukan jumlah k atau c*luster*. Kemudian masuk kedalam metode *Rank Order Centroid* dalam menentukan nilai *centroid* awal, proses *Rank Order Centroid* awalnya dengan memberi nilai bobot pada atribut *dataset* kemudian mengurutkan data berdasarkan hasil *Rank Order Centroid*. Setelah nilai *Rank Order Centroid* didapatkan proses kemudian memilih nilai *centroid* awal berdasarkan tertinggi - terendah *Rank Order Centroid* pada proses sebelumnya. Setelah nilai *ce*ntroid awal didapatkan kemudian menghitung jarak antara objek dengan *centroid* menggunakan rumus *euclidean distance* . Kemudian mengelompokkan data berdasarkan kedekatan data dengan *centroid* awal. Menghitung rata-rata dari data yang berada pada *centroid* yang sama untuk menentukan titik *centroid* baru. Jika titik *centroid* baru memiliki anggota *cluster* yang sama maka hasil pengelompokan sudah ditemukan jika tidak maka akan mengulangi dari proses menentukan *centroid* baru dengan menggunakan persamaan (2,1). Setelah mendapatkan hasil pengelompokan proses selanjutnya adalah menghitung nilai akurasi pengelompokan menggunakan *Davies-Bouldin Index*, dengan proses pertama menghitung nilai *Sum of Square Within Cluster* (SSW). Proses kedua menghitung nilai *Sum of Square Between Cluster* (SSB) dan kemudian menghitung nilai Rasio, setelah nilai SSW, SSB dan Rasio didapatkan kemudian menghitung nilai DBI menggunakan persamaan (2,7).

2.3 Landasan Teori

2.3.1 Data *Mining*

Menurut Pradana (2019), data *mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam basis data. Data *mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.

Menurut Murti (2017) data *mining* adalah adalah bagian integral dari *Knowledge Discovery in Databases* (KDD) sebuah langkah dalam proses mencari pola-pola yang terdapat dalam setiap informasi. Data *mining* adalah proses mencari pola pola dari sebuah KDD untuk setiap informasi.

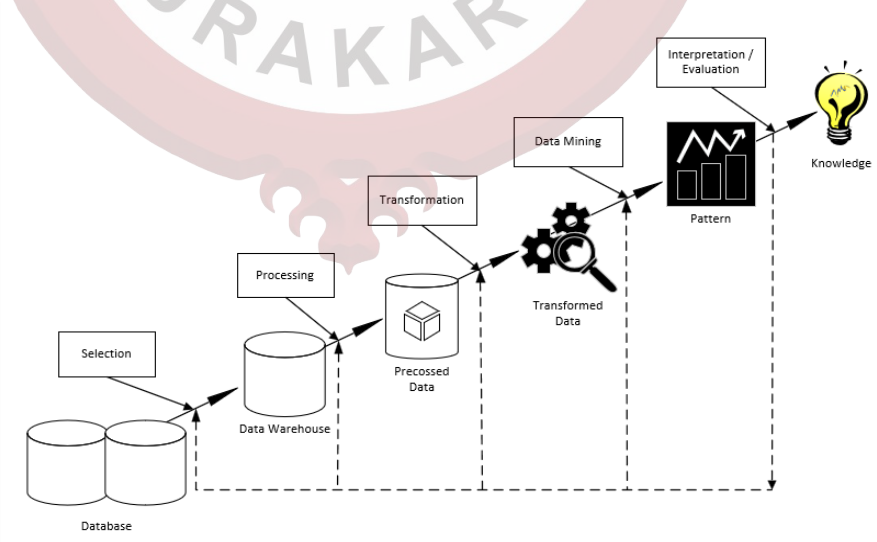
Pada Gambar 2.2 merupakan proses KDD dalam menghasilkan pengetahuan dan terdiri dari beberapa tahap:

a) *Selection*

Selection adalah pengumpulan data-data yang berkaitan dengan data *mining* yang akan diproses, data yang sudah terkumpul akan disimpan di dalam *database* atau disebut juga dengan proses data *warehouse*.

b) *Pre-Processing / Cleaning*

Pre-Processing adalah proses dimana data yang tidak sesuai bentuk/tipe dalam pemprosesan data akan dilakukan perubahan ke dalam bentuk/tipe data yang sesuai dengan algoritma data *mining*. Data dari hasil *pre-processing* memiliki sifat yang terstruktur dan efektif untuk diproses, dan data disimpan dalam *database* untuk pengolahan lebih lanjut.



Gambar 2.2 Data *Mining* sebagai proses penemuan pengetahuan

(Sumber : Mulaab, 2017)

c) *Tranformation*

Tranformation adalah proses pemilihan *coding* untuk data yang sesuai dalam proses data *mining*. Pemilihan *coding* dalam data *mining* tergantung pada jenis atau pola informasi yang akan dicari dalam data *mining*.

d) Data *Mining*

Data *mining* adalah proses perancangan metode analitis dari data yang sudah diproses.

e) *Interpretation / Evaluasi*

Interpretation adalah menampilkan hasil proses data *mining* ke dalam bentuk yang mudah dipahami/ dimengerti oleh pihak yang berkepentingan. Tahap ini juga mengevaluasi fakta atau hipotesis yang sudah ada sebelumnya dengan data hasil akhir dari proses data *mining* yang berupa pola atau informasi yang ditemukan.

2.3.2 *Clustering*

Menurut Pradana (2019), *clustering* adalah proses mengelompokkan data data menjadi beberapa kelompok, sehingga objek yang memiliki karakteristik yang hampir sama atau sama akan di kelompokan ke dalam satu kelompok dan objek yang tidak memiliki banyak persamaan atau berbeda akan di kelompokan ke kelompok lain.

Menurut Olivia (2019), *clustering* dengan pendekatan partisi (*Partition-Based Clustering*) adalah proses pengelompokan data dengan menyaring data yang sudah dianalisis ke dalam *cluster* yang ada. *Clustering* dengan pendekatan hirarki (*Hierarchical Clustering*) adalah proses pengelompokan data yang mirip ditempatkan pada hirarki yang bedekatan sedangkan yang tidak diletakan pada hirarki yang berjauhan.

2.3.3 K-Means

Menurut Olivia (2019), Metode K-Means pertama kali diperkenalkan oleh MacQueen JB pada tahun 1976. K-Means merupakan

suatu proses partisi data ke dalam bentuk satu atau lebih *cluster*, sehingga satu *cluster* memiliki data yang karateristiknya sama dan data yang memiliki karakteristik yang berbeda akan di kelompokkan di *cluster* yang berbeda.

Menurut Wanto (2020) metode K-Means memiliki karakteristik yaitu :

1) Metode K-Means merupakan metode yang sederhana, mudah dan sangat cepat dalam proses *clustering*.
2) Metode K-Means pada jenis *dataset* tertentu tidak dapat melakukan segmentasi data dengan baik, di mana hasil segmentasinya tidak memberikan pola kelompok yang mewakili karakteristk bentuk alami data.
3) Data yang mengandung *outlier* dapat menjadi masalah ketika pengelompokkan di metode K-Means.
4) Metode K-Means sangat sensitif pada pembangkitan *centroid* awal secara acak.
5) *Clustering* yang dihasilkan oleh metode K-Means bersifat tidak unik (selalu berubah-ubah), terkadang baik, terkadang jelek.
6) Suatu *cluster* kemungkinan tidak mempunyai anggota.
7) Metode K-Means sangat sulit untuk mencapai global optimum.

Menurut Wanto (2020), Berikut ini langkah-langkah yang terdapat pada algoritma K-Means :

1. Tentukan jumlah c*luster* (k) pada *dataset*.
2. Tentukan nilai pusat (c*entroid*).

   Penentuan nilai *centroid* pada tahap awal dilakukan secara acak, sedangkan pada tahap iterasi digunakan rumus rumus sebagai berikut :

   $$V_{ij} = \frac{1}{Ni} \sum_{k=0}^{Ni} Xkj \qquad (2,1)$$

   Di mana

   V_{ij} : *Centroid* rata-rata *cluster* ke-i untuk variabel ke-j

   N_i : Jumlah anggota *cluster* ke-i

i, k : Indeks dari *cluster*

j : Indeks dari variabel

Xkj : Nilai data ke-k variabel ke-j untuk *cluster* tersebut

3. Pada masing-masing *record*, hitung jarak terdekat dengan *centroid* menggunakan rumus *Euclidean distance* .

   $De = \sqrt{(xi - si)^2 + (yi - ti)^2}$ (2,2)

   Di mana

   De : *Euclidean distance*

   i : Banyaknya objek

   (x,y) : Koordinat objek

   (s,t) : Koordinat *centroid*

4. Kelompokkan objek berdasarkan jarak ke *centroid* terdekat.
5. Ulangi langkah ke-3 hingga langkah ke-4, lakukan iterasi hingga *centroid* bernilai optimal.

Pada penerapan metode K-Means *cluster* analisis, data yang bisa diolah dalam perhitungan adalah data numerik yang berbentuk angka. Sedangkan data selain angka juga bisa diterapkan tetapi terlebih dahulu harus dilakukan pengkodean untuk mempermudah perhitungan jarak/ kesamaan karakteristik yang dimiliki dari setiap objek. Setiap objek dihitung kedekatan jaraknya berdasarkan karakter yang dimiliki dengan pusat *cluster* yang sudah ditentukan sebelumnya, jarak terkecil antara objek dengan masing-masing *cluster* merupakan anggota *cluster* yang terdekat. Setelah jumlah *cluster* ditentukan, selanjutnya dipilih sebanyak 3 objek secara acak sesuai jumlah *cluster* yang dibentuk sebagai pusat *cluster* awal untuk dihitung jarak kedekatannya terhadap semua objek yang ada.

2.3.4 *Rank Order Centroid* (ROC)

Rank Order Centroid (ROC) dapat dihasilkan melalui tingkat kepentingan atau prioritas dari kriteria. Teknik ROC memberikan bobot pada setiap kriteria sesuai dengan ranking yang akan diberikan nilai berdasarkan tingkat prioritas (Ghazali, 2021).

Formula untuk menentukan prioritas ROC yaitu :

Jika $c_{r^1} \geq c_{r^2} \geq c_{r^3} \geq c_{r^4} \geq \ldots \geq c_m$

Maka, $w_1 \geq w_2 \geq w_3 \geq w_4 \geq \ldots \geq w_n$

Selanjutnya, jika k merupakan banyaknya kriteria, maka:

W1 = $\frac{1+\frac{1}{2}+\frac{1}{3}\ldots+\frac{1}{k}}{k}$

W2 = $\frac{0+\frac{1}{2}+\frac{1}{3}\ldots+\frac{1}{k}}{k}$

W3 = $\frac{0+0+\frac{1}{3}\ldots+\frac{1}{k}}{k}$

Wk = $\frac{0+\ldots+0\ldots+\frac{1}{k}}{k}$

Secara umum Pembobotan ROC dapat dirumuskan sebagai berikut:

$W_k = \frac{1}{k}\sum_i^k = 1(\frac{1}{i})$ (2,3)

Dimana :

Wk : Normalisasi rasio perkiraan skala bobot tujuan

i : Total jumlah tujuan

k : *Ranking* dari i tujuan

2.3.5 *Davies-Bouldin Index* (DBI)

Davies-Bouldin Index (DBI) adalah salah satu metode pengukuran validitas *cluster* pada suatu metode pengelompokan, kedekatan data antara titik pusat *cluster* dari *cluster* yang diikuti akan dijumlah sebagai kohesi. Sedangkan jarak antar titik pusat *cluster* dari *cluster* yang diikuti disebut sebagai *seprasi*. *Davies-Bouldin Index* adalah sebuah pengukuran yang memaksimalkan jarak *inter-cluster* antara c_i dan c_j, dan mencoba meminimalkan jarak antara titik dalam sebuah *cluster* pada waktu yang sama. Jika perbedaan antar *cluster* terlihat jelas maka tingkat kesamaan karakteristik antar masing-masing *cluster* rendah, yang berarti jarak *inter-cluster* maksimal. Jika perbedaan antar *cluster* tidak terlalu jelas maka tingkat kesamaan karakteristik antar masing-masing *cluster* tinggi, yang berarti jarak *intra-cluster* minimal. (Sitompul, 2018). Tahapan dari perhitungan *Davies-Bouldin Index* sebagai berikut.

Sum of Square Within Cluster (SSW) sebagai metrik kohesi dalam sebuah *cluster* ke-i diformulasikan pada persamaan (2,4).

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \quad (2,4)$$

Dimana :

SSW : *Sum of Square Within Cluster*

m_i : Jumlah data dalam *cluster* i

$d(x, c)$: Jarak data x ke *centroid* c

Nilai d dalam persamaan (2,4) bisa menggunakan formula ketidakmiripan (jarak) yang digunakan ketika proses pengelompokkanya sehingga validasi yang diberikan juga mempunyai maksut yang sama terhadap proses pengelompokkanya.

Sementara metrik untuk separasi antara dua *cluster* digunakan formula *Sum of Square Between Cluster* (SSB) dengan mengukur jarak antar *Centroid* c_i dan c_j seperti pada persamaan (2,5).

$$SSB_{i,j} = d(c_i, c_j) \quad (2,5)$$

Dimana :

SSB : *Sum of Square Between Cluster*

$d(c_i, c_j)$: Jarak *centroid* c_i dengan *centroid* c_j

Langkah selanjutnya adalah menghitung nilai R_{ij}. R_{ij} adalah ukuran Rasio seberapa baik nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j. Nilainya didapatkan dari komponen kohesi dan separasi. *cluster* yang baik adalah yang mempunyai kohesi yang sekecil mungkin dan separasi yang sebesar mungkin. R_{ij} di formulasikan dalam persamaan (2,6).

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (2,6)$$

Dimana :

R_{ij} : Nilai Perbandingan *cluster* i dengan *cluster* j

Setelah kita mendapatkan nilai SSW, SSB dan R_{ij}. Nilai *Davies Bouldin Index* (DBI) dapat dihitung menggunakan formula dalam persamaan (2,7).

$$DBI = \frac{1}{k} \sum_{j=1}^{k} \max(R_{ij}) \quad (2,7)$$

Dimana :

K : Jumlah *Cluster* yang digunakan

Nilai DBI yang lebih kecil atau rendah memiliki akurasi yang lebih akurat (Muhima, 2023), artinya kemiripan antar data dalam satu *cluster* akan semakin mirip. DBI banyak digunakan untuk membantu *clustering* berbasis non-hierarki seperti K-Means atau K-Modes untuk menentukan berapa jumlah *cluster* yang tepat untuk digunakan.