

## **BAB II**

### **LANDASAN TEORI**

Pada bab ini akan menjelaskan mengenai penelitian sebelumnya dan dasar teori yang akan dijadikan acuan atau landasan dalam pengerjaan Tugas Akhir ini.

#### **2.1 Tinjauan Pustaka**

Penelitian sebelumnya oleh Sartika & Jumadi (2019) dengan judul penelitian “*Clustering* Penilaian Kinerja Dosen Menggunakan *Algoritma K-Means* (Studi Kasus: Universitas Dehasen Bengkulu)”. Penilaian kinerja (*performance appraisal*) adalah suatu proses yang digunakan pimpinan untuk menentukan apakah seorang dosen melakukan pekerjaannya sesuai dengan tugas dan tanggung jawabnya. Penilaian kinerja juga berlaku dalam perguruan tinggi, baik itu berupa universitas, institut, maupun sekolah tinggi. Penilaian kinerja ini diperlukan dalam kerangka penjamin mutu internal dalam bentuk *feed back* (umpan balik) bagi dosen dan pimpinan universitas atau perguruan tinggi serta untuk mengetahui pengaruh pengajaran dosen terhadap mahasiswa. Pada penelitian ini diperoleh bahwa hasil pengelompokan kinerja dosen menggunakan K-Means terbentuk data kelompok dosen sangat baik terdiri dari 12 anggota dengan total nilai centroid 48.550, data kelompok dosen baik terdiri 29 anggota dengan total nilai centroid 40.340, data kelompok dosen cukup baik 10 anggota dengan total nilai centroid 37.963 dan kelompok dosen kurang baik terdiri dari 9 anggota dengan total nilai centroid 37.033 (Sartika & Jumadi, 2019).

Fikri, Agustin & Mintarsih (2017) melakukan penelitian dengan judul “Pengelompokan Kualitas Kerja Pegawai Menggunakan *Algoritma K-Means++* Dan *COP-Kmeans* Untuk Merencanakan Program Pemeliharaan Kesehatan Pegawai Di PT. PLN P2B JB Depok”. Penelitian dilakukan bahwa pengelompokan kualitas kerja pegawai menggunakan *Algoritma K-Means* dalam pemeliharaan kesehatan pegawai terdiri atas 5 kelompok kerja dengan pegawai yang berbeda dan penyakit yang berbeda, dimana penyakit dan jumlah penyakit dalam kelompok tersebut akan digunakan sebagai bahan pertimbangan Bu Fatma untuk

merencanakan program kesehatan pegawai di semester berikutnya (Fikri, Agustin, & Mintarsih, 2017).

Panggi Oktara, Liza Yulianti, & Jhoanne Fredricka (2021) dengan judul penelitian “Analisis Kinerja Pegawai Menggunakan *Algoritma K-Means* pada Dinas Pendidikan dan Kebudayaan Kabupaten Bengkulu Tengah” dimana Penilaian Kinerja dilakukan dengan form yang sudah ditetapkan. Setiap tahunnya akan dilakukan evaluasi kinerja pegawai untuk mengambil keputusan dan bahan pertimbangan dalam mengetahui kinerja yang diperoleh pegawai tersebut. Adapun kriteria penilaian kinerja pegawai pada Dinas Pendidikan dan Kebudayaan Kabupaten Bengkulu Tengah ini yaitu Orientasi Pelayanan, Integritas, Komitmen, Disiplin, Kerjasama, dan Kepemimpinan. Hasil pengelompokan terhadap data penilaian kinerja pegawai, didapatkan bahwa jumlah pegawai yang berada pada *cluster* tinggi (*Cluster I*) sejumlah 22%, dan jumlah pegawai yang berada pada *cluster* rendah (*Cluster II*) sejumlah 82% dari 27 pegawai. Fungsi dari aplikasi sudah berjalan dengan baik dan aplikasi dapat membantu pihak Dinas Pendidikan dan Kebudayaan Kabupaten Bengkulu Tengah dalam mengetahui kinerja para pegawainya berdasarkan 2 kelompok yaitu *cluster* tinggi dan *cluster* rendah (Oktara, Yulianti, & Fredricka, 2021).

Penelitian yang dilakukan oleh Hardika Khusnuliawati (2018) yang berjudul “Algoritma Pengelompokan Menggunakan *Self-Organizing Map* dan *K-Means* pada Data Sumber Daya Manusia Provinsi Indonesia” dimana metode pengelompokan data yang merupakan gabungan dari SOM dan *K-Means* dapat diimplementasikan pada permasalahan dunia nyata untuk mengelompokkan kondisi suatu wilayah berdasarkan informasi demografinya. Informasi demografis yang digunakan yaitu kualitas sumber daya manusia sebagai indikator wilayah tertinggal dari 33 provinsi di Indonesia. Algoritma SOM merupakan algoritma yang diimplementasikan pada tahap pertama untuk memperoleh visualisasi dari hasil pengelompokan provinsi di Indonesia. Sedangkan tahap kedua diujikan *algoritma K-Means* untuk memperjelas hasil pengelompokan data yang kemudian dievaluasi menggunakan *algoritma Silhouette*. Dari uji coba yang dilakukan menunjukkan hasil pengelompokan terbaik dari 33 provinsi di

Indonesia berdasarkan informasi sumber daya manusia yaitu sejumlah dua hingga tiga kelompok dengan rata-rata nilai *silhouette value* yang dihasilkan yaitu 0.6238 dan 0.6116 (Khusnuliawati, 2018)

## 2.2 Kerangka Pemikiran

Berikut ini merupakan kerangka pemikiran yang digunakan dalam penelitian penilaian kinerja pejabat struktural Kabupaten Karanganyar yang dijelaskan oleh Gambar 2.1. Kerangka pemikiran terdiri dari:

### 1. Latar Belakang Masalah

Belum adanya metode untuk mengelompokkan penilaian kinerja bagi pejabat struktural di Pemerintah Kabupaten Karanganyar sebagai masukan bagi Bupati dalam hal ini tim Baperjakat untuk mengetahui kualitas kerja dari pejabat struktural serta sebagai contoh bagi para bawahan dalam melaksanakan tugas sehari-hari.

### 2. Perumusan Masalah

Masalah yang akan diselesaikan dalam penelitian adalah “Bagaimana mengetahui kualitas kinerja pejabat struktural di Pemerintah Kabupaten Karanganyar Menggunakan *Algoritma K-Means*?”.

### 3. Pendekatan

Pengumpulan data yang dilakukan yaitu dengan tanya jawab atau *interview*, *observasi*, maupun studi literatur tentang penelitian terdahulu serta literatur di perpustakaan. Selanjutnya, dalam tahapan ini, data yang sudah ada akan diproses untuk dikelompokkan sesuai dengan kriteria-kriteria yang ditentukan menggunakan aplikasi pengolahan data supaya memperoleh hasil yang akurat.

Data dalam penelitian ini adalah data primer yang diambil langsung pada Badan Kepegawaian dan Pengembangan Sumber Daya Manusia Kabupaten Karanganyar melalui *aplikasi* Sistem Informasi Pegawai (*SIMPeg*), yang di dalamnya ada menu *e-SKP* (Sasaran Kinerja Pegawai).

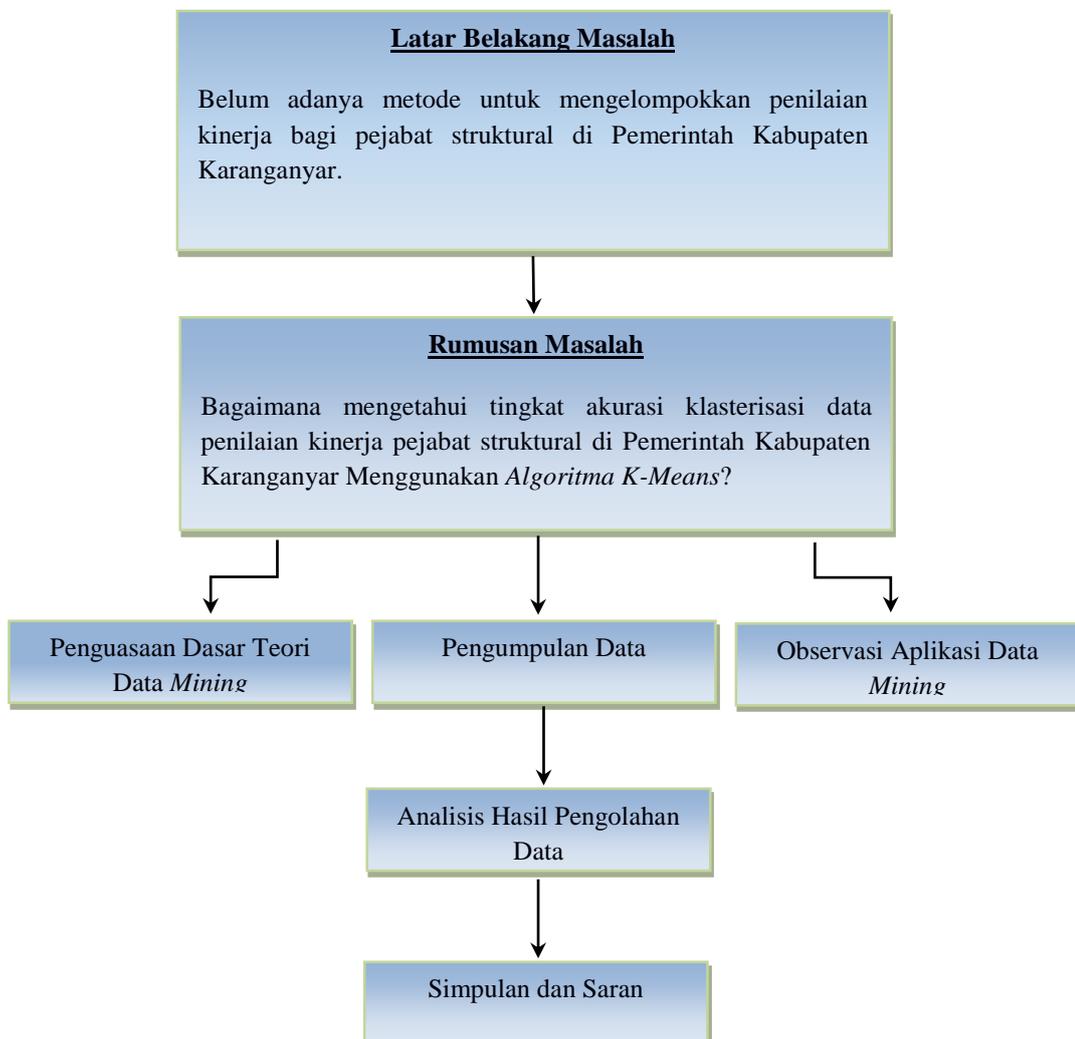
### 4. Analisa Hasil Pengujian

Pada tahap ini dilakukan analisis kepada hasil pengolahan data berupa data penilaian kinerja dari pejabat struktural Pemerintah Kabupaten Karanganyar

sehingga dari hasil yang diperoleh bisa ditentukan keputusan yang tepat untuk mengelompokkan pejabat struktural berdasarkan kriteria sangat baik, baik, dan cukup baik.

## 5. Hasil Penelitian

Pada tahap kelima ini, akan ditarik simpulan akan hasil pengolahan data sekaligus akan menjadi jawaban atas rumusan masalah penelitian. Namun juga tidak menutup kemungkinan bahwa diberikan saran dan masukan untuk penelitian sejenis selanjutnya.



**Gambar 2.1 Kerangka Pemikiran**

## 2.3 Teori-Teori Pendukung

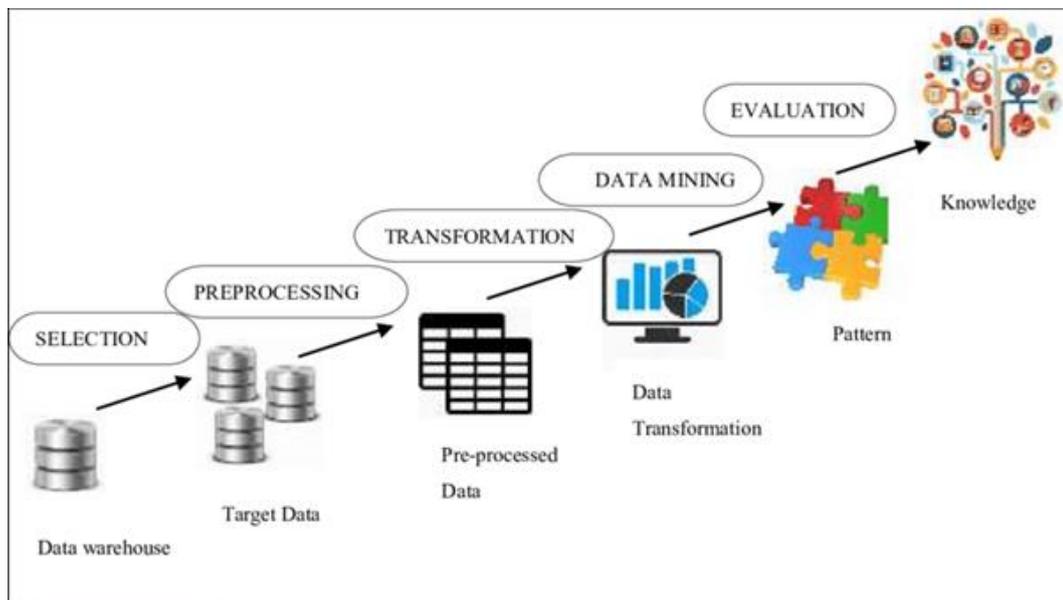
Penyusunan Tugas Akhir memerlukan suatu referensi pendukung yang digunakan sebagai landasan teori agar penelitian dapat berjalan dengan benar dan tidak meyimpang dari kaedah ilmu pengetahuan yang ada. Landasan teori diperoleh dari berbagai sumber dan literatur yang mempublikasikan pendapat beberapa ilmuwan yang digunakan sebagai pendukung pembahasan masalah dalam penelitian Tugas Akhir.

### 2.3.1 Data Mining

#### A. Pengertian Data Mining Dalam Berbagai Disiplin Ilmu

Data *mining* adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Istilah lain yang sering digunakan diantaranya *knowledge discovery (mining) in databases* (KDD), *knowledge extraction*, data atau *pattern analysis*, *data archeology*, *data dredging*, *information harvesting*, dan *bussiness intelligence*. Teknik data *mining* digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Tidak semua pekerjaan pencarian informasi dinyatakan sebagai data *mining*. Sebagai contoh, pencarian *record individual* menggunakan *database management system* atau pencarian halaman *web* tertentu melalui kueri ke semua *search engine* adalah pekerjaan pencarian informasi yang erat kaitannya dengan *information retrieval*. Teknik-teknik data *mining* dapat digunakan untuk meningkatkan kemampuan sistem-sistem *information retrieval* (Iqbal, 2020).

*Knowledge Discovery in Database Process* (KDD) adalah salah satu metode yang bisa digunakan dalam melakukan data *mining*. Fayyed et al. (1996) mendefinisikan KDD sebagai proses dari menggunakan metode data *mining* untuk mencari informasi-informasi yang berharga, pola yang ada di dalam data, yang melibatkan *algoritma* untuk mengidentifikasi pola pada data. Dunham (2003) meringkas proses KDD dari berbagai step, yaitu: seleksi data, pra-proses data, transformasi data, data *mining*, dan yang terakhir interpretasi dan evaluasi. Berikut adalah ilustrasi serta penjelasan mengenai proses KDD seperti yang ada di Gambar 2.2 (Fahlevi, 2021).



**Gambar 2.2** Proses *knowledge discovery (mining) in databases (KDD)*

Gambar 2.2 menggambarkan proses KDD dalam menghasilkan *knowledge* dan terdiri dari beberapa tahap:

a. *Data Cleaning*

Untuk menghapus data yang tidak dipakai dan data yang tidak konsisten.

b. *Data Integration*

Berbagai sumber data dapat digabungkan.

c. *Data Selection*

Data yang bersangkutan pada tugas analisis diseleksi dan diambil kembali dari database.

d. *Data Transformation*

Data diubah atau diperkuat menjadi bentuk yang seharusnya untuk diolah dengan menganalisis ringkasan atau jumlah total agregasi.

e. *Data mining*

Sebuah proses penting di mana metode intelijen diterapkan dengan tujuan untuk mengolah pola-pola data

f. *Pattern Evaluation*

Untuk mengidentifikasi pola-pola menarik yang menjelaskan mengenai ukuran dasar pengetahuan yang ada

#### g. *Knowledge Presentation*

Visualisasi dan teknik representasi *knowledge* digunakan untuk menyajikan *knowledge* yang telah diolah untuk pengguna.

### **B. Posisi Data *mining* Dalam Berbagai Disiplin Ilmu**

Para ahli berusaha menentukan posisi bidang data *mining* di antara bidang-bidang yang lain. Hal dikarenakan ada kesamaan antara sebagian bahasan data *mining* dengan bahasan di bidang lain. Memang tidak seratus persen sama, tetapi ada sejumlah kesamaan karakteristik dalam beberapa hal. Kesamaan bidang data *mining* dalam bidang statistik adalah penyampelan, estimasi, dan pengujian hipotesis (Meilisa, 2021).

### **C. Pengelompokan Data *mining***

Data *mining* merupakan serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Data *mining* mulai ada sejak 1990-an sebagai cara yang benar dan tepat untuk mengambil pola dan informasi yang digunakan untuk menemukan hubungan antara data untuk melakukan pengelompokan ke dalam satu atau lebih *Cluster* sehingga *objek-objek* yang berada dalam satu *Cluster* akan mempunyai kesamaan yang tinggi antara satu dengan lainnya. Data *mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Pramudiono, 2007).

Data *mining* adalah analisis pengamatan *database* untuk menemukan hubungan yang tidak terduga dan untuk meringkas data dengan cara atau metode baru yang dapat dimengerti dan bermanfaat kepada pemilik data (Larose, 2005)

Menurut Larose (2005), metode data *mining* dibagi menjadi enam kelompok berdasarkan tugasnya, yaitu :

#### 1. Deskripsi

Deskripsi bertujuan untuk mengidentifikasi pola yang muncul secara berulang pada suatu data dan mengubah pola tersebut menjadi aturan dan kriteria yang dapat mudah dimengerti oleh para ahli pada domain aplikasinya. Aturan yang dihasilkan harus mudah dimengerti agar dapat dengan efektif meningkatkan tingkat pengetahuan (*knowledge*) pada sistem. Tugas *deskriptif* merupakan

tugas data *mining* yang sering dibutuhkan pada teknik *postprocessing* untuk melakukan validasi dan menjelaskan hasil dari proses data *mining*. *Post-processing* merupakan proses yang digunakan untuk memastikan hanya hasil yang valid dan berguna yang dapat digunakan oleh pihak yang berkepentingan.

## 2. Prediksi

Prediksi memiliki kemiripan dengan klasifikasi, akan tetapi data diklasifikasikan berdasarkan perilaku atau nilai yang diperkirakan pada masa yang akan datang. Contoh dari tugas prediksi misalnya untuk memprediksikan adanya pengurangan jumlah pelanggan dalam waktu dekat dan prediksi harga saham dalam tiga bulan yang akan datang.

## 3. Estimasi

Estimasi hampir sama dengan prediksi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi.

## 4. Klasifikasi

Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang menjelaskan dan membedakan konsep data ke dalam kelas-kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari *objek* dan memasukkan *objek* ke dalam salah satu kelas yang sudah didefinisikan sebelumnya.

## 5. Klustering

Klustering merupakan pengelompokan data tanpa berdasarkan kelas data tertentu ke dalam kelas *objek* yang sama. Sebuah *cluster* adalah kumpulan *record* yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan *record* dalam *cluster* lain. Tujuannya adalah untuk menghasilkan pengelompokan *objek* yang mirip satu sama lain dalam

kelompok-kelompok. Semakin besar kemiripan *objek* dalam suatu *cluster* dan semakin besar perbedaan tiap *cluster* maka kualitas analisis *cluster* semakin baik.

#### 6. Asosiasi

Tugas asosiasi dalam data *mining* adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*market basket analysis*). Tugas asosiasi berusaha untuk mengungkap aturan untuk mengukur hubungan antara dua atau lebih atribut.

### 2.3.2 Klasterisasi

*Clustering* atau klasterisasi adalah metode pengelompokan data. Menurut Tan, Steinbach, & Kumar (2007), *clustering* adalah sebuah proses untuk mengelompokkan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum (Tan, Steinbach, & Kumar, 2007)

*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 intelligence*, pengenalan pola citra, *web search*, bidang ilmu biologi, dan untuk keamanan (*security*). Di dalam *business intelligence*, *clustering* bisa mengatur banyak *customer* ke dalam banyaknya kelompok. Contohnya mengelompokkan *customer* ke dalam beberapa *cluster* dengan kesamaan karakteristik yang kuat. *Clustering* juga dikenal sebagai data segmentasi karena *Clustering* mempartisi banyak data set ke dalam banyak group berdasarkan kesamaannya. Selain itu *clustering* juga bisa sebagai *outlier detection* (Tan, Steinbach, & Kumar, 2007).

### 2.3.3 Teorema Algoritma K-Means

#### A. Algoritma K-Means

Menurut Wu & Kumar (2009), *Algoritma K-Means* merupakan algoritma pengelompokan *iterative* yang melakukan partisi set data ke dalam sejumlah *K Cluster* yang sudah ditetapkan di awal. *Algoritma K-Means* sederhana untuk diimplementasikan dan dijalankan, relative cepat, mudah beradaptasi, umum penggunaannya dalam praktek. (Wu & Kumar, 2009).

Secara historis, *K-Means* menjadi salah satu *algoritma* yang paling penting dalam bidang data *mining* (Prasetyo, 2014).

*Algoritma K-Means* adalah *algoritma* yang terbaik dalam *algoritma* *partitional clustering* dan yang paling sering digunakan di antara *Algoritma Clustering* lainnya karena kesederhanaan dan efisiensinya. Kelebihan *Algoritma K-Means* itu sendiri menurut K. Arai and A. R. Barakbah, merupakan *Algoritma Clustering* yang paling sederhana dan umum, hal ini dikarenakan *K-Means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Namun, *K-Means* mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster* (Tahta, Budi, & Ali, 2013).

*Algoritma K-Means* merupakan salah satu metode *Clustering* non-hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*. Algoritma ini mempartisi data ke dalam *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang lain. *Algoritma* ini sederhana untuk diterapkan dan dijalankan, relatif cepat, mudah diadaptasi, dan umum digunakan dalam praktek (Saragih, Sembiring, & Sayuthi, 2018).

*Algoritma K-means* merupakan algoritma paling banyak digunakan dengan aplikasi penggunaan yang berukuran kecil hingga menengah. Hal ini terjadi karena kemudahan implementasinya serta memiliki waktu komputasi yang relatif cepat dan kemampuan algoritma dalam melakukan adaptasi terhadap kasus yang berbeda. Ide dasar *algoritma K-means* sangatlah sederhana, yaitu dengan meminimalkan *Sum of Squared Error* antara objek data dengan sejumlah *k*

*centroid*. Namun kelemahan utama dari *algoritma K-means* adalah hasil yang terlalu sensitif terhadap pemilihan cluster awal dan perhitungan solusi lokal untuk mencapai kondisi optimal (Purnamasari & Dana, 2017).

*Algoritma K-means* terdiri dari beberapa langkah. Pertama, dari himpunan data yang akan dilakukan *clustering* dipilih sejumlah  $k$  objek secara acak yang akan menjadi *centroid* awal ( $k$  merupakan jumlah yang ditentukan oleh peneliti). Kedua, setiap objek yang bukan merupakan *centroid* awal dimasukkan ke dalam cluster terdekat. Ketiga, setiap *centroid* diperbarui berdasarkan rata-rata perhitungan dari objek yang ada dalam cluster tersebut. Perhitungan ini dilakukan pada setiap cluster sesuai jumlah yang ditentukan oleh peneliti. Keempat, ulangi langkah kedua dan ketiga dengan melakukan proses iterasi hingga semua *centroid* stabil. *Centroid* stabil berarti nilai *centroid* yang dihasilkan dalam iterasi yang saat ini bernilai sama dengan nilai *centroid* yang dihasilkan dalam iterasi yang sebelumnya (Setiawan, 2018).

### **B. Tujuan Clustering K-Means**

Tujuan pengelompokan (*Clustering*) data dapat dibedakan menjadi dua, yaitu pengelompokan untuk pemahaman dan pengelompokan untuk penggunaan. Jika tujuannya untuk pemahaman, kelompok yang terbentuk harus menangkap struktur alami data, biasanya proses pengelompokan dalam tujuan ini hanya sebagai proses awal untuk kemudian dilanjutkan dengan pekerjaan inti seperti peringkasan (rata – rata, standard deviasi), pelabelan kelas pada setiap kelompok untuk kemudian digunakan sebagai data latih klasifikasi, dan sebagainya (Murti, 2017)

Sementara untuk penggunaan, tujuan utama pengelompokan biasanya adalah mencari *prototype* kelompok yang paling representatif terhadap data, memberikan abstraksi dari setiap *objek* data dalam kelompok dimana sebuah data terletak di dalamnya. Contoh – contoh tujuan pengelompokan untuk pemahaman sebagai berikut :

#### a. Biologi

Seperti yang sudah banyak diketahui, bahwa hewan – hewan di alam ini dikelompokkan menurut karakter – karakter tertentu secara hierarkis, yaitu kerajaan, *filum*, kelas, *ordo*, suku, *genus*, dan spesies. Level tertinggi adalah kerajaan,

sedangkan level terendah adalah spesies. Satu jenis hewan mempunyai nama spesies sendiri. Dua hewan dengan spesies berbeda dapat mempunyai *genus* yang sama. Sejumlah hewan dengan *genus* berbeda dapat mempunyai suku yang sama. Begitu juga dengan *ordo*, *filum*, dan kerajaan. Semua hewan berada dalam kelompok yang sama (satu kelompok) di level kerajaan, yaitu hewan.

#### b. Information Retrieval

Situs web di internet berjumlah miliaran. Ketika di-*query*, mesin pencari akan memberikan hasil ribuan halaman. Teknik pengelompokan dapat digunakan untuk mengelompokkan hasil halaman yang diberikan mesin pencari ke dalam kelompok yang lebih kecil di mana setiap kelompok berisi halaman yang berkarakteristik sama atau mirip. Misalnya dengan kata kunci *query* “*movie*” dapat diberikan hasil halaman yang dibedakan dalam kategori seperti “*genre*”, “*star*”, “*theaters*”, dan sebagainya. Setiap kategori dapat dipecah kembali menjadi subkategori yang membentuk hirarki sehingga membantu pengguna mengeksplorasi hasil *query*.

#### c. Instansi Pemerintah

Instansi Pemerintah Daerah dalam hal ini Instansi Kepegawaian di suatu daerah mempunyai data informasi pegawai dalam jumlah besar, baik itu data pribadi, data riwayat kenaikan pangkat, data riwayat jabatan dan lain lain. Salah satu yang penting juga adalah data penilaian kinerja seluruh pegawai yang terdiri dari Sasaran Kinerja Pegawai dan Perilaku Kerja. Pengelompokan data penilaian kinerja ini dapat diterapkan untuk memecah pegawai ke dalam kelompok – kelompok kecil yaitu pegawai dengan kinerja sangat baik, baik dan cukup baik. Pengelompokan tersebut dapat digunakan untuk mendapatkan informasi terkait kualitas kinerja dari pejabat struktural yang dapat dijadikan masukan bagi Tim Baperjakat (Badan Pertimbangan Jabatan dan Kepangkatan) saat akan melakukan promosi jabatan. Sedangkan contoh – contoh tujuan pengelompokan untuk penggunaan sebagai berikut :

##### a. Kompresi

Data – data yang bergabung dalam setiap kelompok dapat dianggap berkarakter sama atau mirip sehingga data – data dalam kelompok yang sama dapat dikompresi dengan diwakili oleh indeks *prototype* dari setiap kelompok. Setiap *objek*

direpresentasikan dengan indeks *prototype* yang dikaitkan dengan sebuah kelompok. Teknik kompresi ini dikenal dengan kuatisasi vektor (*vector quatization*).

#### b. Pencarian Tetangga Terdekat Secara Efisien

Pada teknik *K-Nearest Neighbor*, komputasi untuk mencari tetangga terdekat akan semakin berat ketika jumlah data semakin besar. Hal tersebut tidak sebanding dengan jumlah data yang akhirnya digunakan sebagai tetangga terdekat. Dengan pengelompokan, kita dapat membuat *prototype* dimana setiap *prototype* mewakili satu kelas. Dengan cara ini, komputasi pencarian tetangga terdekat dapat digantikan dengan *prototype* terdekat. Hal ini dapat mengurangi waktu komputasi secara signifikan. Konsekuensinya adalah mungkin cara mendapatkan tetangga terdekat menjadi kurang *representative* karena diukur berdasarkan kelompok data bukan individu data sehingga ada kemungkinan distorsi hasil yang tidak diinginkan.

### C. Langkah *Clustering K-Means*

Menurut Witten & Frank (2018) adapun langkah-langkah metode *K-Means* antara lain sebagai berikut (Witten & Frank, 2018) :

- a) Pilih secara acak  $k$  buah data sebagai pusat *Cluster*.
- b) Jarak antara data dan pusat *Cluster* dihitung menggunakan *Euclidian Distance*. Untuk menghitung jarak semua data ke setiap titik pusat *Cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagai berikut :

$$d_{ij} = \sqrt{\sum_{k=1}^p \{X_{ik} - X_{jk}\}^2} \quad (2.1)$$

Keterangan Persamaan 2.1

$d_{ij}$  = Jarak objek antara objek  $i$  dan  $j$

$p$  = Dimensi data

$X_{ik}$  = Koordinat dari objek  $i$  pada dimensi  $k$

$X_{jk}$  = Koordinat dari objek  $j$  pada dimensi  $k$

- c) Data ditempatkan dalam *cluster* yang terdekat, dihitung dari tengah *cluster*.

- d) Pusat *cluster* baru akan ditentukan bila semua data telah ditetapkan dalam *cluster* terdekat.
- e) Proses penentuan pusat *cluster* dan penempatan data dalam *cluster* diulangi sampai nilai *centroid* tidak berubah lagi.

#### 2.3.4 Jenis Data Dalam Set Data

Sebuah data set dapat dipandang sebagai sebuah koleksi dari *objek – objek* data. Nama lain dari sebuah *objek* data adalah *record*, titik, vektor, pola, *event*, *case*, *sample*, observasi atau entitas. *Objek – objek* data dijelaskan oleh sejumlah atribut yang menangkap karakteristik dasar dari sebuah *objek*, seperti massa dari sebuah *objek* fisik atau waktu pada saat sebuah kejadian terjadi. Nama – nama lain untuk atribut adalah variabel, karakteristik, *field*, fitur atau dimensi. Atribut adalah sifat atau *property* atau karakteristik *objek* data yang nilainya dapat bermacam – macam dari sebuah *objek* ke *objek* lain, dari satu waktu ke waktu yang lain. Misalnya, warna kulit seseorang bisa berbeda dengan warna kulit orang lain, berat badan seseorang juga bisa berubah dari waktu ke waktu. Warna kulit bisa menjadi nilai simbolik (putih, hitam, kuning, langsung, cokelat, sawo matang), sedangkan berat badan bisa berupa nilai numerik. Atribut yang menjadi *element* setiap data mempunyai jenis yang beragam. Berat badan, pada contoh sebelumnya, mempunyai nilai *numeric* sehingga dapat dibandingkan satu sama lain, sedangkan warna kulit tidak bisa dibandingkan karena menggunakan nilai yang sifatnya kualitatif. Umumnya, type atribut ada dua, yaitu kategoris (kualitatif) dan *numeric* (kuantitatif) (Darmawan, 2022).

#### 2.3.5 Evaluasi *Davies-Bouldin Index*

*Davies-bouldin index* merupakan salah satu metode evaluasi internal yang mengukur evaluasi cluster pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap centroid dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar centroid dari clusternya.

*Sum of square within cluster* (SSW) merupakan persamaan yang digunakan untuk mengetahui matrik kohesi dalam sebuah cluster ke- $i$  yang dirumuskan sebagai berikut :

$$\underline{SSW_i = \frac{1}{m_1} \sum_{j=1}^{m_1} d(x_j, c_i)} \quad (2.2)$$

Dari persamaan tersebut,  $m$  merupakan jumlah data dalam cluster ke- $i$ ,  $c_i$  adalah centroid cluster ke- $i$ , dan  $d()$  merupakan jarak setiap data ke *centroid* yang dihitung menggunakan jarak *euclidean*.

*Sum of square between cluster* (SSB) merupakan persamaan yang digunakan untuk mengetahui separasi antar cluster yang dihitung menggunakan persamaan :

$$\underline{SSB_{i,j} = d(c_i, c_j)} \quad (2.3)$$

Setelah nilai *kohesi* dan *separasi* diperoleh, kemudian dilakukan pengukuran rasio ( $R_{ij}$ ) untuk mengetahui nilai perbandingan antara cluster ke- $i$  dan cluster ke- $j$ . Cluster yang baik adalah cluster yang memiliki nilai *kohesi* sekecil mungkin dan *separasi* yang sebesar mungkin. Nilai rasio dihitung menggunakan persamaan sebagai berikut :

$$\underline{R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{ij}}} \quad (2.4)$$

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai *davies-bouldin index* (DBI) dari persamaan berikut :

$$\underline{DBI = \frac{1}{K} \sum_{i=1}^k \max_{i \neq j} (R_{i,j})} \quad (2.5)$$

Dari persamaan tersebut,  $k$  merupakan jumlah cluster yang digunakan. Semakin kecil nilai DBI yang diperoleh (non-negatif  $\geq 0$ ), maka semakin baik cluster yang diperoleh dari pengelompokan *K-means* yang digunakan (Muhammad, 2015)

### 2.3.6 RapidMiner

*Software* pendukung yang akan digunakan oleh peneliti pada penelitian ini adalah *software RapidMiner Studio Educational 10.0. RapidMiner* sebelumnya dikenal sebagai YALE (*Yet Another Learning Environment*). YALE dikembangkan pada tahun 2001 oleh sekelompok orang yang berasal dari Unit Kecerdasan Buatan Universitas Teknik Dortmund (*A.I Unit in TU Dortmund*). Kelompok tersebut terdiri dari Ralf, Klinkenberg, Ingo Mierswa serta Simon Fisher. Dilanjutkan pada tahun 2006, Ralf dan Ingo mendirikan sebuah perusahaan yang bernama *Rapid-I*. Ralf dan Ingo melanjutkan untuk membangun serta mengembangkan YALE sehingga setahun kemudian pada tahun 2007, YALE mengalami perubahan nama menjadi *RapidMiner*. Pada tahun 2013, perusahaan yang didirikan Ralf dan Ingo mengalami perubahan nama dari *Rapid-I* menjadi *RapidMiner* sesuai dengan nama aplikasi yang diciptakan oleh mereka. *RapidMiner* adalah platform perangkat lunak ilmu data yang dikembangkan oleh perusahaan yang memiliki nama yang sama. *RapidMiner* menyediakan lingkungan terintegrasi untuk melakukan persiapan data, data mining, AI, *word mining* guna mendukung penelitian, pendidikan, pelatihan, *rapid prototyping* serta pengembangan aplikasi. Aplikasi *RapidMiner* juga mendukung dalam melakukan proses pembelajaran mesin termasuk persiapan data, hasil visualisasi, validasi model dan optimasi data. Walaupun aplikasi *RapidMiner* sangat membantu dalam tugasnya, aplikasi ini masih bersifat *opensource* yang berarti model inti terbuka dan dapat didapatkan secara gratis pada website yang disediakan dengan syarat penggunaan *RapidMiner* harus bertujuan dalam membantu pendidikan. Jika digunakan secara komersil, maka akan dikenakan biaya.

*RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap data *mining*, teks *mining* dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki kurang lebih 500 operator data *mining*, termasuk operator untuk *input*, *output*, *data preprocessing* dan visualisasi. *RapidMiner* merupakan *software* yang berdiri

sendiri untuk analisis data dan sebagai mesin data *mining* yang dapat diintegrasikan pada produknya sendiri. *RapidMiner* ditulis dengan menggunakan bahasa *java* sehingga dapat bekerja di semua sistem operasi.

*RapidMiner* menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan XML (*Extensible Markup Language*) yang mendefinisikan proses analitis keinginan pengguna untuk diterapkan ke data. File ini kemudian dibaca oleh *RapidMiner* untuk menjalankan analisis secara otomatis.

### 2.3.7 Penilaian Kinerja

Penilaian kinerja (*performance appraisal*) menurut (Dessler, 2020) adalah suatu proses atau kegiatan yang dilakukan oleh perorangan atau kelompok dalam sebuah perusahaan untuk mengevaluasi dan mengomunikasikan bagaimana karyawan melakukan pekerjaan dengan cara membandingkan hasil pekerjaannya dengan seperangkat standar yang telah dibuat dalam suatu periode tertentu yang digunakan sebagai dasar pertimbangan suatu kegiatan.

Dalam Peraturan Pemerintah Nomor 30 Tahun 2019 tentang Penilaian Kinerja Pegawai Negeri Sipil, Penilaian Kinerja Pegawai Negeri Sipil bertujuan untuk menjamin *objektivitas* pembinaan Pegawai Negeri Sipil yang didasarkan pada sistem prestasi dan sistem karier. Penilaian dilakukan berdasarkan perencanaan kinerja pada tingkat individu dan tingkat unit atau organisasi, dengan memperhatikan target, capaian, hasil, dan manfaat yang dicapai, serta perilaku Pegawai Negeri Sipil.

Pasal 4 pada Peraturan Pemerintah Nomor 30 Tahun 2019 tentang Penilaian Kinerja Pegawai Negeri Sipil berbunyi Penilaian Kinerja Pegawai Negeri Sipil dilakukan berdasarkan prinsip :

- a) *Objektif*
- b) *Terukur*
- c) *Akuntabel*
- d) *Partisipatif*
- e) *Transparan*

Dalam Peraturan Pemerintah ini perencanaan kinerja terdiri atas penyusunan dan penetapan SKP (Sasaran Kinerja Pegawai) dengan memperhatikan Perilaku Kerja.

### **2.3.8 Pejabat Struktural**

Pejabat struktural adalah suatu jabatan yang secara tegas ada dalam struktur organisasi. Kedudukan jabatan struktural bertingkat-tingkat dari tingkat yang terendah (eselon IV.b) hingga yang tertinggi (eselon I.a). Contoh jabatan struktural di PNS Pusat adalah : Sekretaris Jenderal, Direktur Jenderal, Kepala Biro, dan Staf Ahli. Sedangkan contoh jabatan struktural di Pegawai Negeri Sipil Daerah adalah : Sekretaris Daerah, Kepala Dinas/Badan/Kantor, Kepala Bagian, Kepala Bidang, Kepala Seksi, Camat, Sekretaris Camat, Lurah, dan Sekretaris Lurah.

Berdasarkan Keputusan Kepala Badan Kepegawaian Negara Nomor 13 Tahun 2002 tentang Ketentuan Pelaksanaan Peraturan Pemerintah Nomor 100 Tahun 2000 tentang Pengangkatan Pegawai Negeri Sipil dalam Jabatan Struktural sebagaimana telah diubah dengan Peraturan Pemerintah Nomor 13 Tahun 2002, jabatan struktural adalah kedudukan yang menunjukkan tugas, tanggung jawab, wewenang dan hak seseorang Pegawai Negeri Sipil dalam rangka memimpin suatu satuan organisasi Negara.