BAB II

LANDASAN TEORI

2.1 Data *Mining*

Data *Mining* merupakkan suatu proses pengumpulan dan analisis data untuk menemukan keteraturan, pola, maupun hubungan pada kumpulan data yang luas. Beberapa jenis pendekatan yang digunakan dalam data *mining* diantaranya prediksi, perbandingan, klasifikasi, *cluster*, dan perkiraan (Ikhwan & Aslami, 2020). Data *Mining* juga merupakan proses statistik, metematika, kecerdasan buatan, dan *machine learning* untuk *ekstraksi* dan identifikasi suatu informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Parjito & Permata, 2021).

Sari & Chotijah, (2022) menyebutkan terdapat tiga elemen penting dalam data *mining* diantaranya adalah 1) Data *Mining* merupakan prosedur mekanis pada data yang ada, 2) Data yang hendak diproses merupakan data yang luas 3) Poin dari data *mining* yaitu untuk mendapatkan pola yang dapat memberikan informasi yang bermanfaat. Data *Mining* memiliki manfaat untuk mengelola data mentah, dimensi data yang tinggi dan data campuran yang sifatnya berbeda menjadi kumpulan informasi yang dapat dijadikan sebagai penunjang keputusan secara efektif (Sari et al., 2020).

2.2 *Clustering*

Clustering merupakan suatu metode untuk pengelompokan data yang berkarateristik sama antara data satu dengan yang lainnya. Serliani et al., (2020) menyatakan bahwa *clustering* membagi data dalam satu himpunan ke dalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data pada kelompok lain dan data yang mempunyai kesamaan kriteria akan dikelompokkan ke dalam data yang berbeda. Partisi data dilakukan dengan mencari nilai jarak terdekat antara data dengan nilai *centroid* yang telah ditetapkan baik secara acak

maupun dengan *Initial Set of Centroids* untuk menentukan nilai *centroid* dengan objek yang berurutan (Nursia et al., 2022). Rahmayani et al., (2021) menyatakan bahwasannya *clustering* berpotensi untuk mengetahui struktur pada data yang bisa digunakan lebih lanjut pada berbagai macam aplikasi seperti klasifikasi, pengolahan gambar dan sosialisasi pola. *Clustering* memiliki sifat tidak punya data latih dimana karakteristik tiap *cluster* tidak ditentukan sebelumnya melainkan berdasarkan kemiripan atribut-atribut dari suatu kelompok (Fammaldo & Hakim, 2018).

Hidayat (2022) menyebutkan bahwa tujuan *clustering* data dibedakan menjadi dua yaitu pengelompokan untuk pemahaman struktur alami data dan pengelompokan untuk penggunaan data. Pengelompokan untuk pemahaman struktur alami data bertujuan untuk proses awal kemudian dilanjutkan dengan pekerjaan inti seperti peringkasan atau *summarization* (rata-rata standar *deviasi*) sedangkan pengelompokan untuk penggunaan data bertujuan untuk mencari kelompok yang paling *representatif* terhadap data dan memberikan abstraksi dari setiap objek data.

2.3 Algoritma *K-Means*

K-Means merupakan salah satu algoritma pengelompokan data yang mempartisi data ke dalam dua atau lebih kelompok. Algoritma ini pertama kali diusulkan oleh MacQueen (1996) dan dikembangkan oleh Hartigan dan Wong tahun 1975. *K-Means* mempunyai kemampuan dalam mengelompokan data dengan jumlah yang cukup besar dan waktu komputasi yang cepat dan efesien (Missa et al., 2021). Algoritma *K-Means* memiliki ketelitian yang cukup tinggi terhadap ukuran objek (Febrianti & Lubis, 2022).

Suhartini & Yuliani, (2021) menyatakan bahwa Algoritma *K-Means* dapat merubah data yang ingin di *cluster* menuju beberapa titik yang akan digunakan sebagai acuan dalam pengklasteran data. Data yang di klasterisasi harus numerik. Algoritma *K-Means* hanya mengambil sebagian dari banyaknya dari komponen yang didapatkan kemudian dijadikan pusat *cluster* awal, pada penentuan pusat *cluster* ini dipilih secara acak dari populasi data.

Arifin et al., (2022) menyatakan *K-Means* ialah metode klasterisasi yang paling banyak digunakan diberbagai bidang. Hal tersebut dikarenakan metode *K-Means* bersifat sederhana dan memiliki kemampuan untuk melakukan *cluster* data yang berjumlah besar dalam waktu yang singkat.

Adapun langkah-langkah metode *K-Means Clustering* menurut Nasution et al., (2020) adalah sebagai berikut:

1. Menentukan jumlah *cluster*
2. Melakukan inisialisasi cara yang sering digunakan adalah cara random atau acak.
3. Mengalokasikan semua data ke *cluster* terdekat. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster* dengan menggunakan teori jarak *Euclidean* atau *Manhattan*.
4. Kembali ke step 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*.

2.4 Algoritma *Pillar*

Perbedaan antara Algoritma *K-Means* dan Algoritma *Pillar* terletak pada pemilihan centroid awal. Pada Algoritma *K-Means* pemilihan centroid dilakukan secara acak sebanyak nilai K yang ditentukan, namun pada Algoritma *Pillar* penempatan *centroid* awal ditentukan di dalam ruang fitur dengan menempatkan inisialisasi masing-masing *centroid* awal yang memiliki akumulasi jarak terjauh (Primandana et al., 2019). Algoritma ini terinspirasi pada penempatan pilar di setiap sudut bangunan yang jaraknya jauh sehingga masa bangunan terkonsentrasi pada setiap pilar. Langkah-langkah metode Algoritma *Pillar* menurut (Seputra & Wijaya, 2020).

1. Menentukan k=jumlah klaster.
2. Menghitung rata-rata semua data =Max semua data
3. Hitung jarak data dengan pusat klaster D[n] = dis(X, m) seperti pada persamaan 2.1.

$$dis(X, m) = |\sum_{J=1}^{N} Xj - mj| \quad (2.1)$$

1. Menghitung nmin dan nbdis $= a\frac{n}{k}$ semua data dibagi *cluster*

   Ket : Nmin (Alpha = 0,1)

   Nbids (Beta = 0,9)

2. Pengecekan keterpenuhan *range* data dengan hasil jarak Nmin dan Nbdis =IF(AND(jika memenuhi 'v', jika tidak memenuhi 'false)

   Ket : *False* = tidak memenuhi *range*

   V = memenuhi *range*

3. Pengurutan kandidat berdasarkan nilai jarak terkecil hingga keterbesar
4. Pemilihan data yang memenuhi rata-rata dari jarak (langkah ke 6) dengan jarak *centroid* seperti contoh :
   - Jika jumlah *centroid* yang terpilih adalah 4 maka dilakukan pemilahan jumlah data dan yang terpilih ialah data no 1 dan 3

2.5 Jarak *Euclidean*

Menurut Nishom (2019) jarak *Euclidean* merupakan metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam *Euclidean space* (meliputi bidang *Euclidean* dua dimensi, tiga dimensi, atau bahkan lebih). Adapun rumus dalam pengukuran jarak *Euclidean* seperti pada persamaan 2.2.

$$d(x_{i,}\, y_i) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2.2)$$

Keterangan :

d : Jarak antara *x* dan y
x : Data pusat *cluster*
y : Data pada atribut
i : Setiap data
n : Jumlah data,
xi : Data pada pusat *cluster* ke *i*
yi : Data pada setiap data ke *i*

2.6 *Davies Bouldin Index (DBI)*

Davies bouldin index (DBI) adalah *metric* untuk mengevaluasi atau mempertimbangkan hasil Algoritma *Clustering*. Pertama kali diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Dengan menggunakan *DBI* suatu *cluster* akan dianggap memiliki skema *clustering* yang optimal adalah yang memiliki *DBI* minimal (Butsianto & Saepudin, 2020).

Adapun langkah-langkah perhitungan *Davies Bouldin Index* adalah sebagai berikut.

1. *Sum of Square Within-Cluster (SSW)*

   Untuk mengetahui kohesi dalam sebuah *cluster* ke-I salah satunya adalah dengan menghitung nilai dari *Sum of Square Within-Cluster (SSW)* seperti pada persamaan 2.3.

$$SSW_i = \frac{1}{mi} \sum_{j=i}^{ci} d(X_j, C_j) \quad (2.3)$$

   Dimana :

   mi = jumlah data dalam *cluster* ke- i

   ci = *centroid cluster* ke- i

   $d(X_j, C_j)$ = jarak setiap data ke *centroid* i yang dihitung menggunakan *Euclidean Distance*.

2. *Sum of Square Between-Cluster (SSB)*

   Perhitungan *Sum Of Square Between-Cluster (SSB)* bertujuan untuk mengetahui separasi atau jarak antar *cluster* dengan rumus perhitungan seperti pada persamaan 2.4.

$$SSB_{ij} = d(X_i, X_j) \quad (2.4)$$

   Dimana :

   $d(X_i, X_j)$ = jarak antara data ke- i dengan data ke- j di *cluster* lain.

3. *Ratio* (Rasio)

   Perhitungan rasio (Ri,j) ini bertujuan untuk mengetahui nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j untuk menghitung nilai rasio yang dimiliki oleh masing-masing *cluster*. indeks I dan j merupakan merepresentasikan jumlah *cluster*, dimana jika terdapat 4 *cluster* maka terdapat indeks sebanyak 4 yaitu i,j,k dan l. untuk menentukan nilai rasio seperti pada persamaan 2.5.

$$R_{ij\ldots..,n} = \frac{SSW_i + SSW_j + \cdots + SSW_n}{SSB_i + \cdots + SSB_{ni,nj}} \quad (2.5)$$

   Dimana :

   SSW_i = *Sum of Square Within-Cluster* pada *centroid i*

   SSB_i = *Sum of Square Between Cluster* data ke-i dengan j pada *cluster* yang berbeda

4. *Davies Bouldin Index (DBI)*

   Nilai rasio yang diperoleh dari persamaan 2.5 digunakan untuk mencari nilai *DBI* dengan menggunakan persamaan 2.6.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j,\ldots,k}) \quad (2.6)$$

   Dimana, $(R_{i,j,\ldots,k})$ merupakan *ratio* dari nilai *SSW* dan *SSB* melalui persamaan 2.4 dari persamaan 2.5 maka dapat diketahui *k* adalah jumlah *cluster*. Dari perhitungan *Davies Bouldin Index (DBI)* dapat disimpulkan bahwa jika semakin kecil nilai *Davies Bouldin Index (DBI)* yang diperoleh (*non negatif* >= 0) maka *cluster* tersebut semakin baik.

2.7 Metode *Waterfall*

Metode *waterfall* adalah salah satu model perangkat lunak yang bersifat *linier* atau berurutan, di mana setiap tahapannya harus diselesaikan secara berurutan dan tahapannya bersifat sekuensial dan tidak berulang. Setiap tahap

memiliki *input* dan *output* yang terdefinisi dengan jelas. Model ini sangat cocok digunakan untuk pengembangan perangkat lunak yang bersifat stabil, dan jelas dalam spesifikasinya (Nur 2019). Berikut alur dari metode *waterfall.*

Gambar 2.1 Ilustrasi Model *Waterfall*

2.8 *Black Box Testing*

Black Box Testing adalah salah satu teknik pengujian perangkat lunak yang dilakukan tanpa mengetahui secara detail cara kerja kode program di dalamnya. Teknik ini melihat sistem atau program sebagai sebuah kotak hitam (*black box*) di mana input diberikan dan *output* yang dihasilkan dievaluasi, tanpa memperhatikan bagaimana alur program di dalamnya (Hidayat & Muttaqin 2018).

Tujuan dari *black box testing* adalah untuk memastikan bahwa sistem atau program berfungsi dengan benar sesuai dengan persyaratan fungsional dan non-fungsional yang telah ditetapkan, serta mengidentifikasi kesalahan atau kegagalan dalam program (Wijaya & Astuti 2021).

2.9 Kemiskinan

Kemiskinan merupakan permasalahan yang kompleks yang tidak hanya terjadi di negara berkembang seperti Indonesia, tetapi juga terjadi di negara maju (Kasim et al., 2021). Adawiyah (2020) menyatakan bahwa kemiskinan merupakan suatu situasi dimana sebuah rumah tangga kesulitan untuk memenuhi kebutuhan, dimana lingkungannya kurang memberikan peluang untuk meningkatkan kesejahteraan. Sejalan dengan pendapat Oki et al., (2020) yang menyatakan bahwa kemiskinan disebabkan oleh dimensi alam, dimana alam tidak memberikan dukungan dengan kesuburan tanah sehingga menyebabkan masyarakat tidak mampu menggunakan lingkungan untuk usaha. Mampu didefinisikan memiliki harta yang dapat digunakan untuk memenuhi kebutuhan sehari-hari. Sedangkan kaya adalah suatu kondisi dimana jumlah aset/penghasilan lebih besar daripada hutang/pengeluaran (Yunus 2019).

2.10 Tinjauan Pustaka

Sebagai upaya penguatan topik penelitian, penulis melakukan analisis dari hasil riset penelitian sebelumnya yang berkaitan dengan topik penelitian. Berikut ini beberapa hasil dari penelitian sebelumnya:

a. Sari et al., (2020) melakukan penelitian yang berjudul "Penerapan Algoritma *K-Means* Untuk *Clustering* Data Kemiskinan Provinsi Banten Menggunakan *Rapidminer*" memperoleh hasil penelitian bahwa dengan penerapan Algoritma *K-Means* memperoleh hasil 3 *cluster*, yaitu *cluster* kemiskinan tingkat sedang, tingkat tinggi, dan tingkat rendah. Sehingga, dengan adanya 3 *cluster* tersebut dapat membantu pemerintah dalam memprioritaskan 3 kabupaten (kabupaten Pandeglang, kabupaten Lebak, dan kabupaten serang) dalam memberikan bantuan terutama biaya beasiswa pendidikan maupun dana sosial serta perbaikan infrastruktur lainnya demi kesejahteraan hidup penduduk Banten.
b. Rahmayani et al., (2021) melakukan penelitian yang berjudul "Analisis Algoritma *K-Means* untuk Klustering Penerima Bantuan Sosial Covid-19" memperoleh hasil penelitian perhitungan Algoritma *K-Means* dengan pengujian *tools Rapidminer* 5.3 diperoleh hasil yang sama dengan perhitungan manual dan dapat diterapkan agar lebih efektif dalam menentukan penerimaan bantuan.
c. Filki (2020) melakukan penelitian yang berjudul "Algoritma *K-Means Clustering* dalam Memprediksi Penerima Bantuan Langsung Tunai (BLT) Dana Desa" memperoleh hasil penelitian dengan menggunakan metode *K-Means* memudahkan dalam proses prediksi penerima BLT desa dan menjadi rekomendasi dalam akurasi dan kecepatan didalam pengolahan data.
d. Andrianti & Firmansyah (2020) melakukan penelitian yang berjudul "Penerapan *Clustering* Data Kurang Mampu Di Desa Situmekar Menggunakan Algoritma *K-Means*" memperoleh hasil penelitian dengan penerapan Algoritma *K-Means Clustering* untuk mengelompokan data penduduk kurang mampu lebih efektif dan efisien sehingga pihak desa bisa

lebih mudah dalam menentukan penerima bantuan dengan teknik data *mining*.

e. Sari et al., (2022) melakukan penelitian yang berjudul "Pengelompokan Data Penduduk Penerima BSTP (Bantuan Sosial Tunai Pandemic) Menggunakan Metode Algoritma *K-Means Clustering* (Kantor Desa Padang Brahrang)" memperoleh hasil penelitian dengan pengelompokan data penduduk penerima BSTP menggunakan data *mining* metode *clustering* sangat tepat digunakan untuk menghasilkan *knowledge* kelompok prioritas bantuan di desa Padang brahrang.
f. Arifin (2022) melakukan penelitian yang berjudul "Penerapan Metode *K-Means Cluster* Calon Penerima Kartu Jombang Sehat (KJS) Berbasis *Website*" memperoleh hasil dengan penerapan metode *K-Means clustering* dapat membantu pemerintah desa dalam menentukan penerima bantuan KJS dan data yang digunakan ialah data penduduk Desa Menturo berdasarkan 11 kriteria yang sudah ditentukan.
g. Muhidin & Baragigiratri (2019) melakukan penelitian yang berjudul "Pemetaan Penduduk Calon Penerima Bantuan Renovasi Rumah Desa Pesangkalan Menggunakan Algoritma *Clustering K-Means*" memperoleh hasil penelitian pemetaan penduduk calon penerima bantuan renovasi rumah menggunakan algoritma *K-Means* menghasilkan 3 cluster yaitu penduduk yang layak, penduduk yang kurang layak dan penduduk yang tidak layak menerima bantuan.
h. Ikhwan & Aslami (2020) melakukan penelitian yang berjudul "Implementasi Data *Mining* untuk Manajemen Bantuan Sosial Menggunakan Algoritma *K-Means*" memperoleh hasil dengan implementasi algoritma *K-Means* menghasilkan sebuah aplikasi yang dapat digunakan untuk mengidentifikasi kelompok prioritas penerima hibah PKH untuk keluarga berpenghasilan rendah di kecamatan medan tembung.
i. Parjito & Permata (2021) melakukan penelitian yang berjudul "Penerapan Data *Mining* untuk *Clustering* Data Penduduk Miskin Menggunakan

Metode *K-Means*" Memperoleh hasil dengan menggunakan metode K-*Means*, didapatkan kategori masyarakat miskin dan tidak miskin yang dapat digunakan sebagai salah satu alat dalam menentukan kelompok masyarakat yang memperoleh bantuan dengan melihat variabel kepemilikan aset. Hal tersebut dikarenakan variabel kepemilikan aset mempunyai pengaruh pembuatan *cluster* terbaik.

j. Primandana et al., (2019) melakukan penelitian yang berjudul " Optimasi Penentuan *Centroid* pada Algoritma *K-Means* Menggunakan Algoritma *Pillar* (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi Jawa Timur)" memperoleh hasil dengan penggunaan Algoritma *Pillar* dapat meningkatkan kinerja dari Algoritma *K-Means*.

k. Febrianti & Lubis (2022) melakukan penelitian yang berjudul "Identifikasi Calon Penerima Bantuan Satu Keluarga Satu Sarjana (SKSS) Menggunakan Algoritma *K-Means*" memperoleh hasil dengan metode *K-Means* mampu mengelompokkan penerima bantuan zakat bagi keluarga satu sarjana (SKSS). Sehingga dengan adanya penelitian ini mampu memudahkan kecepatan pengelolaan serta akurasi data penerima zakat menjadi tepat sasaran.

l. Suhartini & Yuliani (2021) melakukan penelitian yang berjudul "Penerapan Data *Mining* untuk Meng*cluster* Data Penduduk Miskin Menggunakan Algoritma *K-Means* di Dusun Bagik Endep Sukamulia Timur" memperoleh hasil penelitian dengan penggunaan algoritma *K-Means* sangat berpengaruh dalam pengelompokan data penduduk miskin di wilayah sukamulia timur. Pengujian dilakukan menggunakan aplikasi *Rapidminer* dan memperoleh hasil sebanyak 3 *Cluster*.

m. Nishom (2019) melakukan penelitian yang berjudul "Perbandingan Akurasi *Euclidean Distance, Minkowski Distance,* dan *Manhattan Distance* pada Algoritma *K-Means Clustering* berbasis *Chi-Square"* mendapatkan hasil bahwa metode *Euclidean* merupakan metode terbaik dalam perbandingan 3 metode untuk mengidentifikasi status disparitas Guru sekolah diseluruh Kota Tegal.

UNIVERSITAS MUHAMMADIYAH
GRESIK