BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian ini, digunakan artikel-artikel penelitian yang digunakan sebagai bahan referensi yang terdiri dari penelitian mengenai pemilihan program studi, komparasi *k-means clustering* berdasarkan nilai K, komparasi kinerja K-NN berdasarkan nilai K, dan komparasi kineja K-NN berdasarkan pengukuran jarak*euclidean* dan *manhattan*.

2.1.1 Penelitian Pemilihan Program Studi

Penelitian rekomendasi penerimaan mahasiswa baru dengan menggunakan algoritma *k-means clustering* menunjukkan bahwa algoritma ini mampu mengelompokkan 8741 data mahasiswa baru ke dalam tiga *cluster*: C1, C2, dan C3 (Muttaqien et al., 2019).

Penelitian penentuan jurusan di SMKSiti Banun dengan menggunakan algoritma *k-nearest neighbors* menunjukkan bahwa algoritma ini mampu mengklasifikasikan60 data calon siswa SMK ke dalam empat kelas jurusan, dengan nilai akurasi sebesar 84%, presisi sebesar 81%, dan *recall* sebesar 84%(Sinaga et al., 2021).

Penelitian seleksi mahasiswa baru pada program studi Teknik Informatika dengan menggunakan algoritma *k-nearest neighbors* menunjukkan bahwa algoritma ini mampu melakukan seleksi terhadap 171 data calon mahasiswa baru

yang memilih program studi Teknik Informatika, dengan nilai akurasi sebesar 57,25% (Saifudin, 2018).

Penelitian klasifikasi peminatan program studi pada Universitas Satya Negara Indonesia dengan menggunakan algoritma *k-means clustering* menunjukkan bahwa algoritma ini mampu melakukan klasifikasi terhadap 312 data calon mahasiswa baru untuk direkomendasikan memilih program studi Teknik Informatika (C1), Sistem Informasi (C2), dan Manajemen Informatika (C3), dengan jumlah data tiap *cluster*: C1 = 78, C2 = 17, dan C3 = 60(Zulkifli, 2018).

Penelitian rekomendasi program studi untuk siswa SMA dengan menggunakan algoritma *k-means clustering* menunjukkan bahwa algoritma ini mampu melakukan rekomendasi kepada 210 siswa SMA Negeri 2 Kota Jambi untuk direkomendasikan memilih program studi bidang Kesehatan/Kedokteran (C1), bidang Agama (C2), bidang Teknik (C3), bidang Pendidikan dan Seni (C4), serta bidang Olahraga (C5), dengan jumlah data tiap *cluster*: C1 = 62, C2 = 28, C3 = 30, C4 = 30, dan C5 = 60 (Alam Jusia et al., 2019).

2.1.2 Komparasi *K-Means Clustering* Berdasarkan Nilai K

Penelitian pengelompokan dokumen skripsi berdasarkan kata-kata yang terkandung di dalam masing-masing dokumen dengan variasi nilai K antara 2 sampai 6, menunjukkan bahwa kinerja terbaik algoritma *k-means clustering* diperoleh pada nilai K = 2 dengan nilai *silhoutte coefficient* (SC) sebesar 0,12 (Adhe et al., 2020).

Penelitian pengelompokan data menggunakan dataset Iris dan Wine dengan variasi nilai K antara 2 sampai 9, menunjukkan bahwa kinerja terbaik algoritma *k-means clustering* diperoleh pada nilai K = 3 dengan jumlah iterasi terendah sebesar 6 (Selvida et al., 2020).

Penelitian pengelompokan data perawatan kesehatan dengan variasi nilai K antara 2 sampai 5, menunjukkan bahwa kinerja terbaik algoritma *k-means clustering* diperoleh pada nilai K = 2 dengan nilai *mean squared error* (MSE) sebesar 0.92(Ambigavathi & Sridharan, 2020).

Penelitian pengelompokan respon masyarakat terhadap kampanye pemilu presiden tahun 2014 yang dirangkum di dalam dataset Sentipol dengan variasi nilai K antara 3 sampai 7, menunjukkan bahwa kualitas terbaik algoritma *k-means clustering* diperoleh pada nilai K = 7 dengan nilai *purity* sebesar 0.47 (Cahyanto et al., 2020).

Penelitian pengelompokan data penggunaan media *smartphone* dengan variasi nilai K antara 3 sampai 7, menunjukkan bahwa kinerja terbaik algoritma *k-means clustering* diperoleh pada nilai K = 3 dengan rata-rata jumlah iterasi terendah sebesar 10.9 (Cahyanto et al., 2020).

Penelitian analisis pengelompokan konsumsi energi di beberapa wilayah provinsi Qinhai dengan variasi nilai K antara 2 sampai 5, menunjukkan bahwa algoritma *k-means clustering*menghasilkan jarak data antar *cluster* terkecil pada nilai K = 4 dengan nilai rata-rata jarak (*separation degree*) sebesar 10826.597(Kong et al., 2021).

2.1.3 Komparasi *K-Nearest Neighbors* Berdasarkan Nilai K

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada kasus pemerataan tenaga pengajar di sekolah umum Provinsi Jawa Tengah dengan variasi nilai K dari 2 sampai 10, menunjukkan kinerja terbaik diperoleh pada nilai K = 9 dengan nilai SSE sebesar 0.617018 (Umargono et al., 2019).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada dataset Iris dengan variasi nilai K = {1, 5}, menunjukkan kinerja terbaik diperoleh pada nilai K = 1 dengan nilai akurasi sebesar 100% (Rao et al., 2022).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada masalah penentuan karir siswa SMA Kejuruan dengan variasi nilai K = {1, 3, 5, 12}, menunjukkan kinerja terbaik diperoleh pada nilai K = 3 dengan nilai akurasi sebesar 69.67% (Nunsina et al., 2020).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada dataset Iris dengan variasi nilai K = {3, 4, 6}, menunjukkan kinerja terbaik diperoleh pada nilai K = 3 dengan rata-rata nilai SSE sebesar 1.689 (M. Z. Hossain et al., 2019).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada data pengguna narkoba dengan variasi nilai K = {2, 3, 4, 5, 6}, menunjukkan kinerja terbaik diperoleh pada nilai K = 6 dengan nilai SSE sebesar 1409.6 (Winarta & Kurniawan, 2021).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada dataset Balance Scale dengan variasi nilai K = {1, 3, 5, 11, 21, 31, 41, 51}, menunjukkan kinerja terbaik diperoleh pada nilai K = 11 dan K = 21 dengan nilai

akurasi sebesar 90.08%(Paryudi, 2019).

Penelitian komparasi kinerja *k-means clustering* berdasarkan nilai K pada dataset Pima Indians dengan variasi nilai K = {5, 10, 15, 20, 25, 30, 35}, menunjukkan kinerja terbaik diperoleh pada nilai K = 10 dengan nilai akurasi sebesar 82.71% (Danil et al., 2019).

2.1.4 Komparasi *K-Nearest Neighbors* Berdasarkan Pengukuran Jarak

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada klasifikasi kelulusan siswa, menunjukkan kinerja terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *euclidean distance,* dimana diperoleh nilai akurasi sebesar 83.331% (Hidayati & Hermawan, 2021).

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada deteksi penyakit stroke, menunjukkan kinerja terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *manhattan distance,* dimana diperoleh nilai akurasi sebesar 96.03% (Iswanto et al., 2021).

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada klasifikasi data tekstual, menunjukkan kinerja terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *euclidean distance,* dimana diperoleh nilai akurasi sebesar 82.167% (Wahyono et al., 2020).

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran

jarak *euclidean distance* dan *manhattan distance* pada klasifikasi kredit macet, menunjukkan kinerja terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *euclidean distance,* dimana diperoleh nilai akurasi sebesar 92.04%(Margolang et al., 2022).

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada lima dataset yang berbeda dengan 10000, 20000, 30000, 40000 dan 50000 jumlah data, menunjukkan performa terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *manhattan distance,* dimana diperoleh nilai rata-rata *running time* sebesar 5737893.594 ms(Ghazal et al., 2021).

Penelitian hasil ulasan komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada 28 dataset UCL Machine Learning, menunjukkan performa terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *manhattan distance*, dimana diperoleh nilai rata-rata akurasi sebesar 81.13%.594 ms(Abu Alfeilat et al., 2019).

Penelitian komparasi kinerja *k-means clustering* berdasarkan pengukuran jarak *euclidean distance* dan *manhattan distance* pada klasifikasi gambar radiologi, menunjukkan kinerja terbaik *k-means clustering* diperoleh dengan menggunakan pengukuran jarak *manhattan distance,* dimana diperoleh nilai akurasi sebesar 91% (Arora et al., 2022).

2.1.5 Tabulasi Referensi Penelitian

Berdasarkan hasil-hasil penelitian yang digunakan sebagai referensi dalam

penelitian ini, penulis melakukan tabulasi terhadap penelitian-penelitian tersebut, sebagaimana terlihat pada Tabel II.1

Tabel II.1 Tabulasi Referensi Penelitian

Penulis	Penelitian	Hasil Penelitian
(Muttaqien et al., 2019)	Rekomendasi penerimaan mahasiswa baru menggunakan *k-means clustering* pada 8741 data calon mahasiswa baru.	C1 = 1.758 data C2 = 4.928 data C3 = 2.055 data
(Sinaga et al., 2021)	Klasifikasi jurusan SMK menggunakan K-NN pada 60 calon siswa.	Akurasi = 84% Presisi = 81% Recall = 84%
(Saifudin, 2018)	Seleksi terhadap 171 calon mahasiswa baru program studi Teknik Informatika menggunakan K-NN.	Akurasi = 57,25%
(Zulkifli, 2018)	Klasifikasi 312 calon mahasiswa baru dalam memilih program studi Teknik Informatika (C1), Sistem Informasi (C2), dan Manajemen Informatika (C3) menggunakan *K-means clustering*	C1 = 78 data C2 = 17data C3 = 60 data
(Alam Jusia et al., 2019)	Klasifikasi 210 siswa SMA dalam rekomendasi program studi bidang Kesehatan (C1), Agama (C2), Teknik (C3), Pendidikan (C4) dan Olahraga (C5) menggunakan *K-means clustering*	C1 = 62 data C2 = 28data C3 = 30 data C4 = 30data C5 = 60 data

(Adhe et al., 2020)	Komparasi pengelompokan dokumen skripsi menggunakan *K-means clustering*, dengan variasi nilai K = 2 sampai K = 6 berdasarkan nilai *silhoutte coefficient* (SC)	K = 2, SC = 0,12 K = 3, SC = -0,02 K = 4, SC = -0,06 K = 5, SC = -0,08 K = 6, SC = -0,012
(Selvida et al., 2020)	Komparasi pengelompokan data Iris dan Wine skripsi menggunakan *K-means clustering*, dengan variasi nilai K = 2 sampai K = 9 berdasarkan jumlah iterasi yang digunakan	K = 2, Iterasi = 10 K = 3, Iterasi = 6 K = 4, Iterasi = 8 K = 5, Iterasi = 9 K = 6, Iterasi = 11 K = 7, Iterasi = 10 K = 8, Iterasi = 15 K = 9, Iterasi = 13
(Ambigavathi & Sridharan, 2020)	Komparasi pengelompokan data perawatan kesehatan menggunakan *K-means clustering*, dengan variasi nilai K = 2 sampai K = 5 berdasarkan nilai *mean squared error* (MSE)	K = 2, MSE = 0.92 K = 3, MSE = 13.81 K = 4, MSE = 8.607 K = 5, MSE = 12.118
(Cahyanto et al., 2020)	Komparasi pengelompokan respon masyarakat terhadap kampanye pemilu 2019 menggunakan *K-means clustering*, dengan variasi nilai K = 3 sampai K = 7 berdasarkan nilai *purity*	K = 3, *Purity* = 0.425 K = 4, *Purity* = 0.433 K = 5, *Purity* = 0.435 K = 6, *Purity* = 0.438 K = 7, *Purity* = 0.47

(Kong et al., 2021)	Komparasi pengelompokan konsumsi energi di wilayah provinsi Qinhai tahun 2019 menggunakan *K-means clustering*, dengan variasi nilai K = 2 sampai K = 5 berdasarkan nilai *separation degree* (C)	K = 2, C = 26846.661 K = 3, C = 43537.777 K = 4, C = 10826.597 K = 5, C = 19427.901
(Awad & Hamad, 2022)	Komparasi pengelompokan data penggunaan media *smartphone* menggunakan *K-means clustering*, dengan variasi nilai K = 3 sampai K = 7 berdasarkan nilai iterasi yang digunakan.	K = 3, Iterasi = 10.9 K = 4, Iterasi = 15.2 K = 5, Iterasi = 21.6 K = 6, Iterasi = 17.8 K = 7, Iterasi = 17.3
(Umargono et al., 2019)	Komparasi klasifikasi pemerataan tenaga pengajar menggunakan K-NN, dengan variasi nilai K = 2 sampai K = 10 berdasarkan nilai SSE.	K = 2, SSE = 2.945106 K = 3, SSE = 2.00171 K = 4, SSE = 1.733812 K = 5, SSE = 1.416148 K = 6, SSE = 1.484771 K = 7, SSE = 1.249919 K = 8, SSE = 0.866852 K = 9, SSE = 0.617018 K = 10, SSE = 0.642253
(Rao et al., 2022)	Komparasi klasifikasi dataset Iris menggunakan K-NN, dengan variasi nilai K = 1dan K = 5 berdasarkan nilai akurasi.	K = 1, Akurasi = 100% K = 5, Akurasi = 96.67%
(Nunsina et al., 2020)	Komparasi penentuan karir siswa SMK menggunakan K-NN, dengan variasi	K = 1, Akurasi = 52.17% K = 3, Akurasi = 69.67%

	nilai K = 1, K = 3, K = 5, dan K = 12 berdasarkan nilai akurasi.	K = 5, Akurasi = 56.52% K = 12, Akurasi = 47.83%
(M. Z. Hossain et al., 2019)	Komparasi klasifikasi spesies bunga Iris menggunakan K-NN, dengan variasi nilai K = 3, K = 4, dan K = 6 berdasarkan nilai SSE.	K = 3, SSE = 1.689 K = 4, SSE = 5.287 K = 6, SSE = 3.543
(Winarta & Kurniawan, 2021)	Komparasi klasifikasi pengguna narkoba menggunakan K-NN, dengan variasi nilai K = 2 sampai K = 6 berdasarkan nilai SSE.	K = 2, SSE = 3585,914 K = 3, SSE = 2328,052 K = 4, SSE = 1977,802 K = 5, SSE = 1569,47 K = 6, SSE = 1409,6
(Paryudi, 2019)	Komparasi klasifikasi dataset Balance Scale menggunakan K-NN, dengan variasi nilai K = 1, K = 3, dan K = 5, K = 11, K = 21, K = 31, K = 41, dan K = 51 berdasarkan nilai akurasi.	K = 1, Akurasi = 84.8% K = 3, Akurasi = 84.8% K = 5, Akurasi = 86.56% K = 11, Akurasi = 90.08% K = 21, Akurasi = 90.08% K = 31, Akurasi = 89.28% K = 41, Akurasi = 89.44% K = 51, Akurasi = 89.6%
(Danil et al., 2019)	Komparasi klasifikasi dataset Pima Indians menggunakan K-NN, dengan variasi nilai K = 5, K = 10, dan K = 15, K = 20, K = 25, K = 30, dan K = 35 berdasarkan nilai akurasi.	K = 5, Akurasi = 79.93% K = 10, Akurasi = 82.71% K = 15, Akurasi = 75.84% K = 20, Akurasi = 79.55% K = 25, Akurasi = 75.28% K = 30, Akurasi = 76.77%

		K = 35, Akurasi = 75.09%
(Hidayati & Hermawan, 2021)	Komparasi klasifikasi kelulusan siswa menggunakan K-NN dengan *euclidean* dan *manhattan* distance berdasarkan nilai akurasi.	*Euclidean*= 83.331% *Manhattan*= 82.823%
(Iswanto et al., 2021)	Komparasi deteksi penyakit stroke menggunakan K-NN dengan *euclidean* dan *manhattan* distance berdasarkan nilai akurasi.	*Euclidean*= 95.93% *Manhattan*= 96.03%
(Wahyono et al., 2020)	Komparasi klasifikasi data tekstual menggunakan K-NN dengan *euclidean* dan *manhattan* distance berdasarkan nilai akurasi.	*Euclidean*= 82.167% *Manhattan*= 80.76%
(Margolang et al., 2022)	Komparasi klasifikasi kredit macet menggunakan K-NN dengan euclidean dan manhattan distance berdasarkan nilai akurasi.	Euclidean = 92.04% Manhattan = 87.96%
(Ghazal et al., 2021)	Komparasi klasifikasi pada 5 dataset yang berbeda menggunakan K-NN dengan euclidean dan manhattan distance berdasarkan nilai rata-rata runtime (detik).	Euclidean = 14437.1933 Manhattan = 95631.5599
(Abu Alfeilat et al., 2019)	Komparasi klasifikasi pada 28 dataset UCL Machine Learning yang berbeda menggunakan K-NN dengan euclidean	Euclidean = 80.01% Manhattan = 81.13%

	dan manhattan distance berdasarkan nilai rata-rata akurasi.	
(Arora et al., 2022)	Komparasi klasifikasi gambar radiologi menggunakan K-NN dengan euclidean dan manhattan distance berdasarkan nilai akurasi.	Euclidean = 88.64% Manhattan = 91%

2.2 Program Studi

Program studi adalah kesatuan kegiatan pendidikan dan pembelajaran yang memiliki kurikulum dan metode pembelajaran tertentu dalam satu jenis pendidikan akademik (Jamaluddin et al., 2020).Program studi merupakan ujung tombak dari sebuah perguruan tinggi dalam menghasilkan lulusan yang berkualitas karena erat hubungannya dengan penyelenggaraan pendidikan tinggi. Untuk itulah, dalam pelaksanaannya, perlu dilakukan perencanaan dan pegembagan program studi yang strategis, sehingga pihak kampus dapat mengetahui faktor-faktor apa saja yang mempengaruhi perencanaan dan pengembangan program studi yang bersangkutan (Ary & Sanjaya, 2020).

Citra sebuah program studi merupakan salah satu alasan utama dalam pertimbangan mahasiswa memilih program studi dalam melanjutkan studinya. Citra program studi ini dapat berupa tingkat akreditasi, fasilitas pendidikan yang ditawarkan, dan prospek terhadap keterbukaan lapangan kerja setelah calon mahasiswa lulus dari program studi yang bersangkutan (Mahmudah & Faidah, 2020).

Akreditasi merupakan sebuah alat yang digunakan untuk mengukur apakah

standar mutu yang telah ditetapkan oleh BAN-PT telah diterapkan dan dilaksakan dengan baik oleh masing-masing program studi. Pada umumnya, akreditasi program studi memiliki tujuan sebagai berikut (Yusran & Nasution, 2020):

1. Menjamin mutu dari program studi pada perguruan tinggi, dengan melihat apakah standar yang sudah ditetapkan telah dijalankan dengan baik
2. Mendorong perbaikan mutu dari program studi pada perguruan tinggi secara bertahap dan berkesinambungan
3. Hasil dari akreditasi dapat digunakan sebagai alat ukur untuk alokasi dana atau bantuan dari pihak luar, sebagai bahan pertimbangan dalam transfer kredit, maupun pengakuan dari badan atau instansi yang berkepentingan.

2.3 Data Mining

Datamining merupakan sebuah proses kegiatan yang digunakan untuk mencari informasi yang berguna berupa pola-pola yang sebelumnya tidak diketahui dari sejumlah besar data. Sebagai inti dari *knowledge discovery in database* (KDD), *data mining* dapat melakukan proses pengorganisasian dan pengidentifikasian data secara otomatis serta menemukan pola-pola pada sekumpulan data yang berjumlah besar dan kompleks (Alkhairi & Situmorang, 2022).Dalam *data mining*, klasifikasi merupakan kateogori *supervised learning* dapat digunakan untuk mengelompokkan sekumpulan data, dimana data yang akan dikelompokkan harus memiliki kelas label atau target terlebih dahulu (Wahyono et al., 2020). *Data mining* juga dapat digunakan untuk mengklasifikasikan data dalam kelompok-kelompok (*cluster*) yang memiliki

kedekatan tingkat kesamaan menggunakan metode *clustering*, yang merupakan bagian dari *unsupervised learning*, tanpa adanya pemberian label kepada kelas target yang diinginkan (Suliman, 2021).

Teknik *data mining* secara umum terbagi menjadi tiga bagian, yaitu (Rumahorbo & Sekarwati, 2020):

1. *Clustering*

   Teknik *data mining* yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripannya satu dengan yang lain. Berdasarkan tingkat kemiripan ini, data di kelompokkan ke dalam kelas-kelas yang disebut dengan *cluster*.

2. *Classification*

   Teknik yang paling umum digunakan di dalam *data mining*, dimana sekumpulan besar data diproses ke dalam sebuah model yang kemudian membaginya ke dalam kelas-kelas yang sudah ditentukan sebelumnya.

3. *Regression*

   Teknik *data mining* yang umumnya digunakan untuk masalah prediksi, dengan cara membangun model hubungan antara satu atau lebih atribut di dalam data sehingga jika ada perubahan atribut di dalam data tersebut dapat digunakan untuk memprediksikan nilainya di waktu yang akan datang.

2.4 *Clustering*

Clustering adalah sebuah teknik multivariat yang dalam proses analisisnya memanfaatkan algoritma *clustering*, dengan tujuan utamanya untuk

mengelompokkan data berdasarkan kedekatan karakteristiknya, sehingga dapat digunakan untuk mengarahkan temuan pola dari masing-masing kelompok tersebut ke dalam data lain yang belum diketahui polanya (Nagari & Inayati, 2020).

Sejauh ini, *clustering* masih sering digunakan dalam penelitian di banyak bidang, karena adanya perkembangan dalam bidang *clustering analysisi*, seiring dengan perkembangan dalam bidang statistik, ilmu komputer, dan kecerdasan buatan. *Clustering analysis* merupakan proses analisis data yang ketat dan teliti, dimana di dalamnya terdapat beberapa tahapan seperti seleksi fitur dan transformasi, pemilihan algoritma *clustering*, dan evaluasi terhadap hasil *clustering*(Liu et al., 2020). Sebuah *cluster* dapat dikatakan baik jika anggota di dalamnya memiliki sifat homogenitas yang tinggi dan memiliki sifat heterogenitas yang tinggi terhadap *cluster* yang lainnya(Adhe et al., 2020).

Secara umum, konsep *clustering* bekerja dengan cara mengelompokan sekumpulan data ke dalam beberapa kelompok (*cluster*), yang dilakukan tanpa adanya pengetahuan yang mendalam tentang kelompok data tersebut. *Clustering* bertujuan untuk mengelompokan sekumpulan data di dalam dataset, ke dalam *cluster* yang memiliki karakteristik yang memiliki kemiripan yang sama, dimana masing-masing*cluster*tersebut mempunyai karakteristik yang berbeda-beda (Dinata et al., 2020).

Kinerja algoritma *clustering*, khususnya yang berbasis pada pengelompokan berdasarkan jarak antar data, sangat tergantung pada representasi datauntukmemfasilitasi pengelompokan dengan lebih baik. Karena termasuk ke

dalam kategori *unsupervised learning* yang tidak membutuhkan kelas label pada data, tantangan utama bagi algoritma *clustering* adalah bagaimana meminimalisir saling ketergantungan antara representasi data dan metode pengukuran jarak yang digunakan dalam proses pengelompokan (Gao et al., 2020).

2.5 *K-Means Clustering*

*K-means clustering*merupakan algoritma *clustering* sederhana dan paling sering digunakan untuk mengklasifikasikan dataset yang tidak diberikan label sebelumnya. Algoritma ini bertujuan untuk menemukan kemiripan data di dalam *cluster* yang diwakili oleh variabel K, dengan menggunakan *mean* atau *centroid* (titik pusat data) sebagai metrik ukuran untuk mengkarakterisasi cluster. *Centroid*dalam *k-means clustering*merupakan titik data yang menunjukkanpusat cluster, dan bisa jadi bukan merupakan anggota di dalam dataset. Algoritma ini membagi n titik data menjadi *clluster* berjumlah k, dimana setiap titik data n memiliki *cluster* yang sesuai dengan *centroid* yang paling mendekati.Selanjutnya, dihitung nilai euclidean distance dari setiap titik data n ke *centroid* masing-masing *cluster*.Setelah itu, titik data dalam sebuah cluster ditetapkan ke dalam *centroid*yang memiliki nilai euclidean distance terkecil. Ketika tidak ada titik data yang tersedia untuk ditetapkan, *centroid* baru "C" dihitung ulang sehingga menjadi *centroid* baru untuk iterasi berikutnya. Proses ini diulang terus-menerus hingga tidak ada perubahan posisi dari *centroid* C tersebut (Ambigavathi & Sridharan, 2020).

Secara umum, langkah-langkah dari algoritma *k-means clustering*adalah

sebagai berikut (Dinata et al., 2020):

1. Tentukan nilai k sebagai jumlah cluster yang akan di bentuk.
2. Tentukan nilai pusta *cluster* awal secara acak atau menggunakan metode inisialisasi sebanyak k.
3. Hitung jarak setiap data dengan masing–masing pusat *cluster* dengan menggunakan rumus pengukuran jarak
4. Kelompokkan setiap data berdasarkan jarak terkecil dengan pusat *cluster* masing-masing.
5. Perbaharui nilai pusat *cluster*, dengan menghitung nilai rata-rata dari anggota pada masing-masing *cluster*.
6. Ulangi langkah proses dari langkah dua hingga tidak terjadi perubahan anggota pada masing-masing *cluster*

2.6 Klasifikasi

Klasifikasi dapat diartikan sebagai sebuah proses untuk menentukan lokasi penempatan objek data kedalam kategori tertentu, dengan cara mempelajari sifat-sifat, karakteristik,dan fungsi dari data tersebut, untuk kemudian dipetakan pada kelompok kelas yang telah ditetapkan sebelumnya (Hasanah et al., 2021). Pada prosesnya, klasifikasi biasanya memiliki beberapa komponen sebagai berikut (Permana & Sahara, 2019):

1. *Class* (Kelas)

Merupakan sebuah variabel yang terdapat di dalam model klasifikasi dan berfungsi untuk mewakili label-label yang sudah ditentukan pada objek yang akan diklasifikasikan.

2. *Predictors* (Prediktor)

   Merupakan sebuah variabel yang terdapat di dalam model klasifikasi dan berfungsi untuk mewakili karakteristik atau atribut-atribut dari data yang akan diklasifikasikan.

3. *Training Dataset* (Dataset Latih)

   Merupakan kumpulan data yang digunakan untuk melatih model, sehingga model tersebut dapat menemukan pola-pola data di dalamnya.

4. *Testing Dataset* (Dataset Uji)

   Merupakan kumpulan data yang digunakan untuk menguji model, sehingga model tersebut dapat diukur kinerjanya sebelum diimplementasikan menjadi bentuk sebuah aplikasi.

2.7 *K-Nearest Neighbors*

K-nearest neighbors (K-NN) adalah algoritma yang termasuk ke dalam kategori *supervised learning*,yang melakukan klasifikasi terhadap data berdasarkan jarak terdekat dari data yang diinputkan (Asnawi et al., 2021). Algoritma K-NN menghitung jarak dari satu data ke data yang lain dalam sebuah dimensi q, dan nilainya digunakan sebagai nilai kedekatan/kemiripan antara data uji dengan data latih (Mardhyath et al., 2020).Pada fase pembelajaran (*training*), algoritma ini menyimpan vektor-vektor fitur data dan hasil klasifikasi dari data

pembelajaran, yang kemudian digunakan pada fase pengujian (*training*), untuk menghasilkanklasifikasi berupa fitur-fitur yang sebelumnya belum diketahui(Wahyono et al., 2020).Inti dari K-NN adalah bagaimana menentukan nilai K yang sesuai, sehingga diperoleh nilai K yang paling optimal untuk seluruh data yang diproses di dalam model (Lubis et al., 2020).

Untuk mengukur jarak antara data baru (*testing*) dengan data lama (*training*), pada algoritma K-NN dapat digunakan persamaan-persamaan pengukuran jarak sebagai berikut (Iswanto et al., 2021):

1. *Euclidean distance*

$$D_{Euclidean} = \|x - y\|_2 = \sqrt{\sum_{j=1}^{N} |x - y|^2} \quad \text{(II.1)}$$

2. *Minkowski distance*

$$D_{Minkowski} = \|x - y\|_\lambda = \sqrt[\lambda]{\sum_{j=1}^{N} |x - y|^\lambda} \quad \text{(II.2)}$$

3. *Manhattan distance*

$$D_{Manhattan} = \|x - y\|_1 = \sum_{j=1}^{N} |x - y| \quad \text{(II.3)}$$

4. *Chebyshev distance*

$$D_{Chebyshev} = \|x - y\|_\lambda = \log_{\lambda - \infty} \sqrt[\lambda]{\sum_{j=1}^{N} |x - y|^\lambda} \quad \text{(II.4)}$$

Dalam algoritma K-NN, selain nilai pemilihan metode pengukuran jarak, nilai K yang dipilih sebagai jumlah tetangga terdekat juga akan mempengaruhi akurasi dari klasifikasi, karena dalam proses penentuan kelas data baru, algoritma ini menggunakan sistem *voting* (suara terbanyak), berdasarkan jumlah tetangga terdekat yang dimiliki oleh masing-masing data (Iswanto et al., 2021).