

BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

Dalam Bab 2, peneliti akan menjelaskan landasan teori yang digunakan sebagai dasar dalam penelitian ini, serta mengulas penelitian-penelitian sebelumnya yang telah dilakukan oleh para peneliti lain.

2.1 Tinjauan Pustaka

Hasil penelitian terdahulu ditunjukkan pada tabel 2.1 berikut:

Tabel 2.1 Hasil penelitian terdahulu

No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
1	Klasterisasi Penjualan Alat-Alat Bangunan menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan)	Alat-alat bangunan	K-Means	Penerapan metode k-mean menggunakan bantuan software rapidminer dalam penelitian ini menghasilkan data klasterisasi item barang yang laris dan yang kurang laris sehingga pihak manajemen dapat menentukan prioritas pembelian stok barang pada barang yang laris (Siregar, 2018).
2	Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online	Ulasan produk toko online	Support Vector Machine (SVM)	Data ulasan produk yang dikumpulkan kemudian dibagi menjadi tiga data yaitu ulasan_1, ulasan_2, dan ulasan_3. Ketiga data ulasan tersebut memiliki tingkat ketidak seimbangan kelas yang berbeda yaitu 1, 36; 2, 44 dan 3,10 (Siringoringo dan Jamaludin, 2019).
3	Klasterisasi provinsi di	Perkembangan kasus Covid-	K-Medoids	Berdasarkan hasil analisis, bahwasanya

No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
	Indonesia berbasis perkembangan kasus Covid-19 menggunakan metode K-Medoids	19 di Indonesia		provinsi yang terkonfirmasi positif corona terbagi menjadi 3 kelompok, 2 provinsi masuk ke kelompok 1, 10 provinsi masuk ke kelompok 2, dan 22 provinsi masuk ke kelompok 3. Pada Kelompok 1 (kategori tinggi) dengan nilai rata rata kasus terkonfirmasi sebesar 12259 kasus, kasus meninggal sebesar 793, dan rata-rata kasus sembuh sebesar 5691. Pada Kelompok 2 (kategori sedang) dengan nilai rata rata kasus terkonfirmasi sebesar 2632 kasus, kasus meninggal sebesar 108, dan rata-rata kasus sembuh sebesar 1077. Pada Kelompok 3 (kategori rendah) dengan nilai rata rata kasus terkonfirmasi sebesar 388 kasus, kasus meninggal sebesar 14.6, dan rata-rata kasus sembuh sebesar 210 (Gunawan, dkk, 2020).
4	Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means	Perkembangan Covid-19 di DKI Jakarta	K-Means	Berdasarkan hasil pengujian dan uji coba terhadap penerapan metode k-means untuk melakukan pengelompokan atau klasterisasi kasus Covid-19 di DKI Jakarta, dapat disimpulkan bahwa metode kmeans dapat

No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
				<p>digunakan dengan baik. Adapun jumlah klaster yang direkomendasikan berdasarkan perhitungan nilai SSE adalah 9 klaster. Namun demikian pada prakteknya, jumlah klaster tersebut dapat ditentukan sendiri oleh pengguna aplikasi berdasarkan pertimbangan-pertimbangan tertentu. Jika jumlah 9 klaster terlalu banyak, penelitian ini merekomendasikan penggunaan 3 (tiga) klaster karena jumlah tersebut memiliki kualitas yang cukup baik berdasarkan hasil perhitungan metode Elbow. Pada klasterisasi dengan 3 (tiga) klaster, dihasilkan pengelompokan C1 sebanyak 19 kecamatan, C2 sebanyak 23 kecamatan, dan C3 sebanyak 2 kecamatan (Solichin dan Khairunnisa, 2020).</p>
5	Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen	Klasterisasi tingkat tridarma pengajaran dosen	K-Means Algoritma	K-Means dapat melakukan klasterisasi terhadap tingkat tridarma pengajaran dosen dan memiliki akurasi sebesar 53.33% terhadap 15 data yang digunakan (Muliono dan Sembiring, 2019).
6	Analisis	Penilaian	Fuzzy C-	Penerapan metode FCM

No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
	Klasterisasi Penilaian Kinerja Pegawai Menggunakan Metode Fuzzy C-Means (Studi Kasus : Pengadilan Tinggi Agama bandar lampung)	kinerja pegawai	Means	ini menghasilkan lima klaster yaitu klaster Klaster sangat baik yang mendapatkan reward (hadiah), Klaster Baik yang akan mendapat Rekomendasi, Klaster Cukup yang tidak mendapatkan reward, Klaster Kurang yang perlu di beri evaluasi , dan Klaster Buruk yang perlu di beri punishmen (hukuman) pada Pengadilan Tinggi Agama Bandar Lampung yang menghasilkan pola bahwa apabila nilai komiten nya baik maka nilai kedisiplinan nya juga baik, pola ini juga dapat menjadi bahan tolak ukur sebagai penentu pegawai mana yang akan mendapatkan reward (Yudhistiraa, dkk, 2022).
7	Penerapan Klasterisasi K-Means Untuk Identifikasi Sebaran Budidaya Udang Vanname	Identifikasi sebaran budidaya udang vanname	K-Means	Hasil perhitungan menggunakan K-means dengan $k=3$ menunjukkan hasil bahwa klaster 2 (C1) merupakan klaster yang paling dominan dari dataset dengan persentase mencapai 42%, sedangkan klaster 1 (C0) merupakan klaster yang memiliki pengaruh paling kecil dibandingkan klaster lainnya. Hal tersebut dilihat dari persentase

No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
8	Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-means Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara	Pemetaan lahan pertanian	Fuzzy C-Means dan K-Means	yang hanya mencapai 23% (Bakri dan Wakhidah, 2018). Berdasarkan hasil implementasi dan analisis, maka penelitian ini memperoleh kesimpulan, algoritma FCM dan KM yang diimplementasikan pada aplikasi dapat berjalan dengan baik dalam mengklasterkan daerah-daerah lahan pertanian sesuai komoditi berdasarkan pada jenis atribut yang digunakan. Dari berbagai data yang ada, data dalam bentuk angka dipilih untuk digunakan sebagai data pengujian. Hal ini dikarenakan data angka dapat dihitung nilai mean atau rata-ratanya. Selain itu pada penelitian FCM dan KM hanya dapat diimplementasikan untuk memproses data dalam bentuk angka (Tamaela, Sedyono dan Setiawan, 2017).
9	Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa	Memprediksi waktu kelulusan mahasiswa	K-Means Clustering	Setelah dilakukan pengujian dan analisis program, maka dapat diperoleh kesimpulan bahwa pada kasus ini implementasi algoritma k-means dalam data mining sudah berhasil, dan bisa menampilkan informasi prediksi kelulusan mahasiswa,

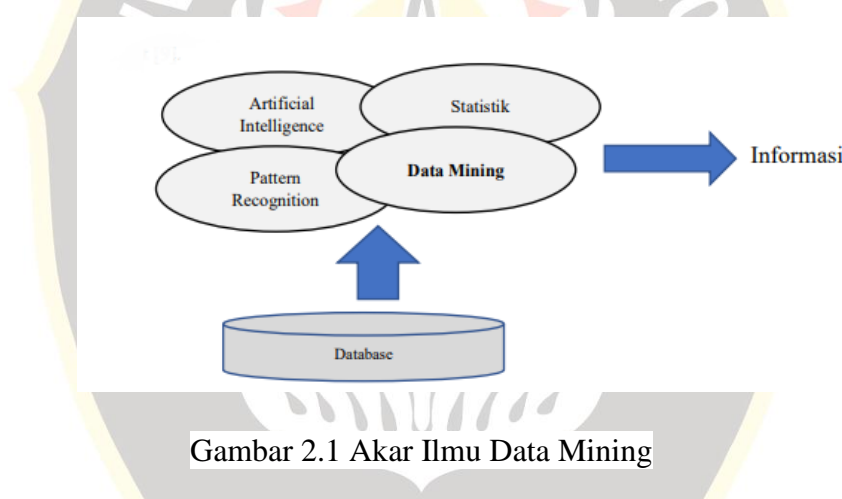
No	Judul Penelitian	Objek Penelitian	Metode / Algoritma	Hasil Penelitian
				namun untuk melihat tingkat kemampuan real k-means clustering dalam memprediksi waktu kelulusan tergantung pada mahasiswa itu sendiri (Priyatman, dkk, 2019).
10	Klasterisasi Pola Penjualan Pestisida Menggunakan Metode K-Means Clustering (Studi Kasus Di Toko Juandatani Kecamatan Hutabayu Raja)	Pola penjualan pestisida	K-Means Clustering	Hasil akhir dari 200 items penjualan pestisida pada Bulan Maret- Juni 2016 adalah Penjualan Pestisida Sangat Laku sebanyak 53 items, Penjualan Laku sebanyak 21 items dan Penjualan Tidak laku sebanyak 126 items (Tamaela, Sedyono and Setiawan, 2017).
11	Sistem Implementasi Analisis Klasterisasi Dan Sentimen Data Untuk Mengetahui Topik Yang Paling Banyak Dibahas Menggunakan Metode K-Means (Studi Kasus Data Twitter Dan Youtube Tentang Ppkm)	Topik yang paling banyak di bahas tentang PPKM	Mini Batch K-Means	Berdasarkan hasil pembentukan dan pengujian klasterisasi menggunakan metode <i>mini batch</i> K-Means terhadap data komentar di kanal Presiden Indonesia di YouTube dan data cuitan yang mengandung kata atau hastag PPKM, diperoleh nilai klaster optimal sebanyak 12 klaster. Hal tersebut diperoleh berdasarkan hasil akumulatif nilai rentang SSE yang telah dilakukan.

2.2 Dasar Teori

2.1.1 Data Mining

Menurut Ordila, dkk, (2020), Data Mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut.

Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara berbeda-beda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data. Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar. Akar ilmu data mining dapat dilihat pada Gambar 2.1.



Gambar 2.1 Akar Ilmu Data Mining

Dalam Gambar 2.1 yang menjelaskan tentang akar ilmu data mining, data mining bermula dari database kemudian dibagi menjadi artificial intelligence, statistik, pattern recognition dan data mining.

Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terikat dari berbagai database besar. Berdasarkan pengertian data mining yang telah dijelaskan di atas, maka data mining merupakan pengetahuan yang tersembunyi di dalam database yang di proses

untuk menemukan pola dan teknik statistik matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan dari database tersebut (Utomo, 2020).

2.1.2 Text Mining

Menurut definisi, Text Mining adalah proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut. Text mining mengekstrak informasi berguna dari sumber data melalui identifikasi dan eksplorasi yang tidak dalam bentuk database record, melainkan dalam data teks yang tidak terstruktur. Prapengolahan adalah tahap proses awal text mining terhadap teks untuk mempersiapkan teks menjadi data yang dapat diolah lebih lanjut. Sekumpulan karakter yang bersambungan (teks) harus dipecah-pecah menjadi unsur yang lebih berarti. Suatu dokumen dapat dipecah menjadi bab, sub-bab, paragraf, kalimat, kata dan bahkan suku kata. Metode ini melibatkan pertama menemukan istilah kunci yang mungkin mewakili isi dokumen, dan kemudian menganalisis hubungan antara dokumen menggunakan metode statistik termasuk analisis cluster, klasifikasi, dan asosiasi (Aditya, 2015).

Tahapan dalam *text mining* secara umum terdiri dari *tokenizing*, *filtering*, *stemming*, *tagging*, dan *analyzing*. *Tokenizing* digunakan untuk memisahkan setiap kata (*token*) dalam dokumen masukan. *Filtering* dilakukan untuk memilih kata-kata yang dihasilkan dari *tokenizing*, dengan menggunakan algoritma *stoplist* atau *wordlist*. Algoritma *stoplist* akan menghapus kata-kata yang dianggap tidak penting, seperti kata ganti, kata keterangan, kata sambung, kata depan, dan kata sandang. Sebaliknya, algoritma *wordlist* akan menyimpan kata-kata yang dianggap penting. Selanjutnya, proses *stemming* dilakukan untuk mencari kata dasar dari setiap kata yang telah melewati proses *filtering*. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang bernilai dari sekelompok dokumen. Oleh karena itu, sumber data yang digunakan dalam text mining adalah kumpulan teks yang memiliki format yang tidak terstruktur atau setidaknya semi terstruktur.

Beberapa tugas khusus dalam *text mining* meliputi kategorisasi teks dan pengelompokan teks (Rahmawati, dkk, 2013).

Text mining dapat memberikan solusi untuk mengatasi berbagai masalah terkait pemrosesan, pengorganisasian, dan analisis data tidak terstruktur dalam skala besar. Dalam menghadirkan solusi, *text mining* mengadopsi dan mengembangkan berbagai teknik dari disiplin ilmu lain, seperti *Data Mining*, *Information Retrieval*, Statistik dan Matematika, *Machine Learning*, Linguistik, Pengolahan Bahasa Alami, dan Visualisasi (Nurhuda dan Sihwi, 2014).

2.1.3 *Term Frequency – Inverse Dokumen Frequency (TF-IDF)*

Term Frequency (TF) adalah frekuensi kemunculan istilah tertentu dalam dokumen tertentu. Semakin sering kata tertentu muncul dalam suatu dokumen, semakin tinggi bobot atau nilai kecocokannya.

Inverse Document Frequency (IDF) adalah pengukuran seberapa luas distribusi sebuah kata dalam kumpulan dokumen yang bersangkutan. IDF mengindikasikan ketersediaan kata tersebut di seluruh dokumen. Semakin besar nilai suatu kata, semakin sedikit dokumen yang menyertakan kata tersebut IDF-nya (Yulian, 2018).

Data Untuk diproses lebih lanjut, segala sesuatu yang telah melalui tahap prapengolahan harus diubah ke dalam format numerik. Salah satu teknik umum untuk mengubah informasi tekstual menjadi bentuk numerik adalah metode pembobotan *Term Frequency - Inverse Document Frequency* (TF-IDF) dengan menggunakan persamaan 2.1 dan persamaan 2.2.

$$TF-IDF = (tf \times idf) \quad (2.1)$$

$$W = (1) \times \left(\ln \frac{(D+1)}{(DF+1)} + 1 \right) \quad (2.2)$$

Dengan:

TF adalah suatu kata atau term terdapat dalam dokumen atau tidak.

DF adalah jumlah dokumen yang memuat kata.

In adalah fungsi inverse.

D adalah jumlah dokumen dan W adalah bobot.

Metode TF-IDF digunakan untuk mengukur tingkat hubungan antara kata (*term*) dan dokumen dengan catatan kaki untuk setiap kata. Dalam metode ini, menggabungkan dua gagasan—frekuensi munculnya kata-kata dalam dokumen (*Term Frequency*) dan kebalikan dari frekuensi dokumen yang mengandung kata tersebut (*Inverse Document Frequency*) (Herwijayanti, dkk, 2018).

Pentingnya sebuah kata dalam sebuah dokumen ditunjukkan oleh frekuensi kemunculannya di sana. Sementara itu, frekuensi dokumen memasukkan kata-kata tersebut menunjukkan seberapa sering kata tersebut muncul di semua dokumen. Oleh karena itu, jika frekuensi kata-kata tersebut tinggi dalam dokumen pendukung tetapi rendah dalam jumlah total dokumen pendukung yang mengandung kata-kata tersebut, tingkat keterkaitan kata-kata tersebut akan tinggi (Triawati, 2009).

Pada frekuensi *term*, terdapat beberapa jenis rumus yang dapat digunakan, antara lain:

1. TF biner (*binary TF*)
2. TF murni (*raw TF*)
3. TF logaritmik
4. TF normalisasi

SEKOLAH PASCASARJANA

2.1.4 *Sum of Squared Errors (SSE)*

Salah satu metode untuk memvalidasi jumlah *cluster* adalah metode *Sum of Squared Errors (SSE)*. Meskipun *Mini Batch K-Means* bekerja dengan baik dalam pengelompokan pada dataset, penting untuk menegaskan kembali bahwa kelemahan *Mini Batch K-means* adalah bahwa kita harus menentukan jumlah *K cluster*, sebelum kita tahu berapa nilai *K* yang optimal (Sari and Devianto, 2014).

Oleh karena itu, ketika menggunakan *Mini Batch K-Means Clustering*, pengguna perlu beberapa cara untuk menentukan apakah mereka menggunakan jumlah *cluster* yang tepat. Adapun rumus SSE dapat dilihat pada persamaan 2.3.

$$SSE = \sum_{i=1}^k (X_i - C_k)^2 \quad (2.3)$$

K adalah jumlah *cluster*.

X_i adalah data ke-*i*.

C_k adalah *centroid cluster*.

Nilai *X_i* dalam penelitian ini merujuk pada nilai yang terdapat dalam dataset, sedangkan nilai *C_k* mengacu pada komponen data uji. Jika nilai SSE mendekati 0, hal ini menunjukkan bahwa model tersebut memiliki kesalahan acak yang sangat kecil. Hal ini berarti nilai SSE yang mendekati 0 lebih berguna dalam melakukan peramalan terhadap model yang sedang diamati. *Sum of Square Error (SSE)* mengacu pada formula yang digunakan untuk membandingkan data yang diamati dengan hasil yang diprediksi. SSE biasanya digunakan sebagai titik referensi untuk penelitian terkait saat menentukan skema klasifikasi terbaik.

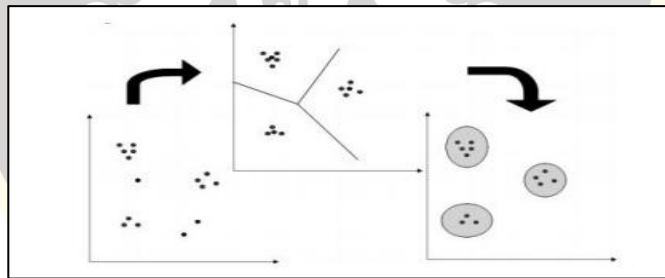
2.1.5 *Klasterisasi K-Means*

Metode Klasterisasi *K-Means* adalah sebuah metode analisis data atau metode data mining yang tidak menggunakan supervisi (*unsupervised*) dan bertujuan untuk melakukan pengelompokan data dengan menggunakan partisi sistem. Nama "K-Means" berasal dari penggunaan konstanta *K* yang menunjukkan jumlah klaster yang diinginkan, dan "*Means*" mengacu pada nilai rata-rata dari setiap kelompok data yang didefinisikan sebagai klaster. K-Means

merupakan salah satu metode data klasterisasi non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu klaster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lainnya (Rahmah, 2020).

Metode ini merupakan salah satu metode yang sering digunakan dan tidak memiliki struktur hierarkis. Metode ini termasuk dalam teknik partisi, di mana objek-objek dibagi atau dipisahkan ke dalam K bagian terpisah. Pada metode K-Means, setiap objek harus ditempatkan dalam kelompok tertentu; namun, pada tahap tertentu dalam proses, item yang ditempatkan dalam satu kelompok dapat dipindahkan ke kelompok lain pada tahap berikutnya (Rahmah, 2020).

Metode ini bertujuan untuk mengurangi perbedaan antara informasi yang termasuk dalam satu klaster dan memaksimalkan perbedaan dengan informasi yang terdapat dalam klaster lainnya. Metode K-Means upaya untuk mengkategorikan data ke dalam banyak kelompok dimana kelompok berbagi karakteristik dan data individu berbagi karakteristik dengan kelompok lain. Ilustrasi algoritma K-Means dapat dilihat pada Gambar 2.2



Gambar 2.2 Ilustrasi Algoritma K-Means (Yudi Agusta, 2007)

Dalam Gambar 2.2 menjelaskan tentang ilustrasi Algoritma K-Means, bagaimana tahapan titik-titik tanpa klaster kemudian dibagi menjadi tiga klaster yang telah dikelompokkan.

Algoritma dari metode K-Means sebelumnya menunjukkan bahwa setiap data dari himpunan data yang diberikan akan diperiksa satu per satu terhadap setiap *centroid* yang telah ditentukan. Misalkan terdapat 500 data yang akan dikelompokkan menjadi 10 klaster, maka setiap data akan diperiksa pada masing -

masing 10 kluster tersebut yang dilakukan secara iteratif dalam jumlah tertentu. Hal ini jelas akan membutuhkan waktu komputasi yang lama terutama ketika berhadapan dengan data berukuran besar. Oleh karena itu, penelitian terhadap metode K-Means berfokus pada memodifikasi metode agar dapat mengefisienkan waktu komputasi untuk pengolahan data yang sangat besar. Salah satu modifikasi tersebut yaitu dengan menggunakan pendekatan *Mini Batch*. Pada modifikasi ini, *centroid* diperbarui hanya dengan menggunakan sekelompok kecil data yang diambil secara acak dari himpunan data. Karena hanya menggunakan sekelompok kecil data dalam prosesnya, metode ini kemudian lebih dikenal dengan istilah *Mini Batch K-Means*.

2.1.6 *Mini Batch K-Means*

Mini Batch K-Means adalah sebuah varian dari algoritma K-Means yang menggunakan *mini-batch*, yaitu mengambil sebagian kecil sampel data secara acak, untuk mengurangi waktu komputasi yang diperlukan dalam proses klusterisasi, sementara masih berusaha untuk mengoptimalkan fungsi tujuan yang sama. *Mini Batch* ini secara drastis mengurangi jumlah perhitungan yang diperlukan untuk menyatu dengan solusi lokal. Berbeda dengan algoritma lain yang mengurangi waktu konvergensi K-Means, *Mini Batch K-Means* menghasilkan hasil yang umumnya hanya sedikit lebih buruk dari pada algoritma standar (Fitriyani dan Murfi, 2016).

Misalkan suatu himpunan data $X = \{x_1, x_2, \dots, x_n\}$, dengan x_n untuk $n = 1, 2, \dots, N$ berdimensi D , yang akan dikelompokkan pada K kluster. Selanjutnya akan diambil sebanyak b data secara acak dari himpunan X , yang di misalkan M , di mana variabel b mendeskripsikan ukuran *Mini Batch*.

Selanjutnya untuk tahap pembaruan *centroid* pada metode *Mini Batch K-Means* hanya akan menggunakan data $x_n \in M$. Penggunaan hanya sebagian kecil data ini jelas akan memberikan hasil yang lebih rendah dari pada menggunakan data dalam menyelesaikan masalah klusterisasi. Oleh karena itu, metode *Mini Batch K-Means* menggunakan metode *gradient descent* untuk memperbarui *centroid* pada setiap iterasinya. Penggunaan metode *gradient descent* ini bertujuan

agar fungsi objektif yang digunakan dapat dengan cepat konvergen menuju titik minimum dan menghasilkan model kluster yang sebanding dengan metode K-Means. Fungsi objektif yang akan dioptimasi pada metode ini menggunakan persamaan 2.4 berikut.

$$\min f(x, \mu) = \sum_{x_n \in M} \|\mu_k - x_n\| \quad (2.4)$$

Selanjutnya Algoritma *Mini Batch K-Means* menggunakan η yang berupa satu dibagi jumlah data dalam kluster tersebut. Karena metode ini berlangsung iteratif untuk setiap $x_n \in M$, maka formula *gradient descent* yang digunakan pada algoritma menggunakan persamaan 2.5 berikut.

$$\mu_{k_{i+1}} = \mu_{k_i} - \eta_k (\mu_{k_i} - x_n) \quad (2.5)$$

Menurut Fitriyani (2016) *Mini Batch* adalah sampel masukan data kecil yang dipilih secara acak dipilih selama setiap iterasi, yang secara signifikan mengurangi waktu perhitungan. Secara keseluruhan, algoritma *Mini Batch K-Means* dapat dijelaskan sebagai berikut.

1. Mulailah dengan menentukan jumlah kluster yang diinginkan.
2. Inisialisasi μ_k yang merupakan *centroid* awal dari setiap kluster secara acak.
3. Ambil data sebanyak b secara acak, yaitu himpunan data M .
4. Menghitung persamaan (2.2) untuk mendapatkan *centroid* terdekat terhadap setiap data $x_n \in M$.
5. Hitung μ_k sama dengan satu dibagi jumlah data dalam kluster tersebut.
6. Hitung *centroid* baru untuk setiap kluster dengan menggunakan persamaan (2.1)
7. Ulangi langkah 5 dan 6 untuk setiap $x_n \in M$.
8. Evaluasi konvergensi parameter μ_k . Jika kriteria konvergensi belum terpenuhi, yaitu kondisi ketika tidak terdapat perpindahan dari data yang digunakan atau mencapai nilai maksimum iterasi, maka ulangi tahapan 3, 4, 5, 6, dan 7 dengan menggunakan μ_k dan himpunan data M yang baru. Keluaran yang dihasilkan berupa nilai μ_k sebagai *centroid* dari setiap kluster.

2.1.7 Word Cloud

Word cloud (awan kata) adalah kelompok kata yang paling sering muncul dalam materi tekstual yang sedang dianalisis. Kumpulan kata yang dihasilkan disebut ilustrasi awan kata karena menyerupai sayap elang. Frekuensi kemunculan suatu kata dapat ditentukan dengan menghitung panjang hurufnya. Frekuensi kemunculan sebuah kata dalam awan kata meningkat seiring dengan bertambahnya ukuran huruf. Awan kata ini menawarkan tata letak yang menarik dan mudah digunakan untuk melihat kata yang sering muncul. Namun, masalah dengan awan kata adalah awan kata tidak dapat menunjukkan seberapa sering kata muncul dalam teks yang sedang dianalisis (Imam dan Fajtriab, 2015).

Python adalah bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Python diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif. Python juga didukung oleh komunitas yang besar. Python mendukung multi paradigma pemrograman, utamanya; namun tidak dibatasi; pada pemrograman berorientasi objek, pemrograman imperatif, dan pemrograman fungsional. Salah satu fitur yang tersedia pada python adalah sebagai bahasa pemrograman dinamis yang dilengkapi dengan manajemen memori otomatis. Seperti halnya pada bahasa pemrograman dinamis lainnya, python umumnya digunakan sebagai bahasa script meski pada praktiknya penggunaan bahasa ini lebih luas mencakup konteks pemanfaatan yang umumnya tidak dilakukan dengan menggunakan Jurnal Dasar Pemograman Python STMIK Juni 2018 2 bahasa script. Python dapat digunakan untuk berbagai keperluan pengembangan perangkat lunak dan dapat berjalan di berbagai platform sistem operasi (Syahrudin dan Kurniawan, 2018).

2.1.8 Pemberlakuan Pembatasan Kegiatan Masyarakat

Sejak awal tahun 2021, Pemerintah Indonesia telah menerapkan kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) sebagai upaya dalam

menangani pandemi COVID-19. Sebelumnya, terdapat kebijakan Pembatasan Sosial Berskala Besar (PSBB) yang telah diterapkan sebelum PPKM di beberapa wilayah di Indonesia. PPKM dilaksanakan di wilayah-wilayah yang menjadi fokus penyebaran virus COVID-19, seperti Pulau Jawa dan Bali (Krisdiyanto, 2021).

Pemerintah telah mengeluarkan kebijakan publik yang disebut Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) untuk pertama kalinya pada tanggal 11 Januari - 25 Januari 2021. Kebijakan ini mencakup daerah DKI Jakarta dan 23 kabupaten/kota di enam provinsi yang memiliki tingkat risiko tinggi terkait COVID-19. Sebelumnya, pemerintah telah memperkenalkan istilah lain pada awal pandemi, yaitu Pembatasan Sosial Berskala Besar (PSBB) dan PPKM Mikro. Aturan-aturan ini mulai diberlakukan sejak tanggal 17 April 2020. Kemudian, pemerintah memperkenalkan istilah baru, yaitu PPKM, sebagai upaya untuk mengendalikan penyebaran COVID-19 di daerah-daerah dengan risiko tinggi (Rizal, 2021).

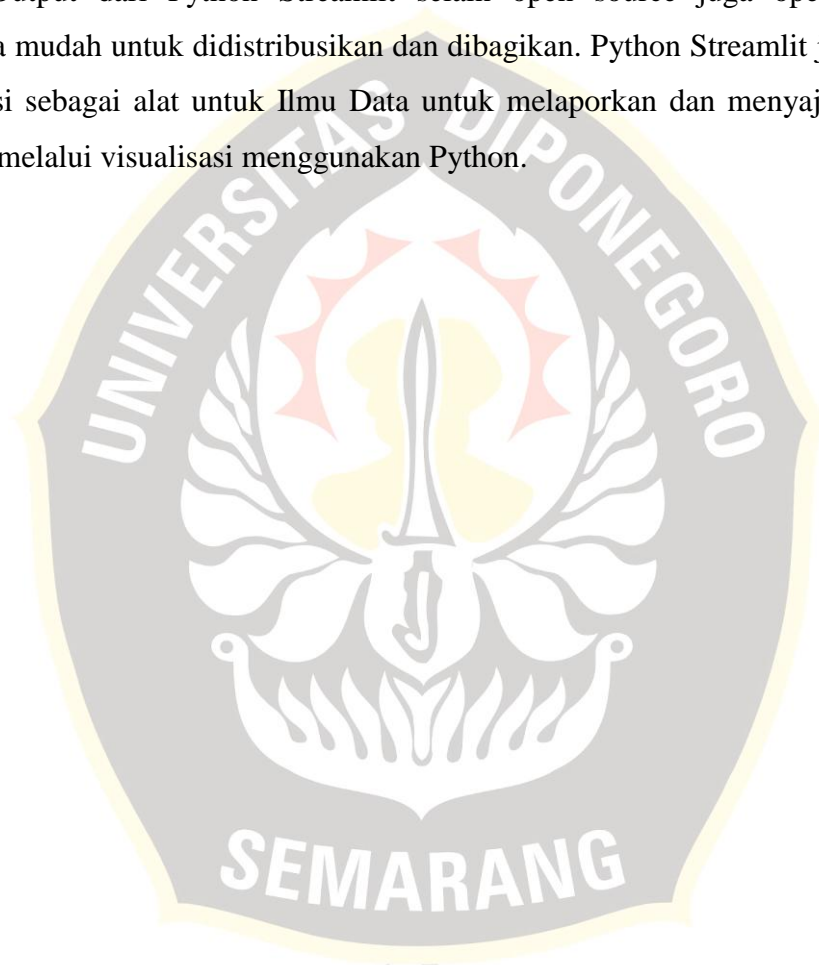
2.1.9 Kerangka Kerja Streamlit

Kerangka kerja (*Framework*) adalah sebuah kerangka kerja yang digunakan untuk mengembangkan *website*. Kerangka kerja ini dibuat untuk membantu pengembang *website* dalam menulis kode bersih. Proses penulisan kode akan lebih sederhana, cepat, dan terstruktur jika Anda menggunakan template pengkodean. Kerangka kerja Streamlit merupakan salah satu kerangka kerja yang ada pada bahasa pemrograman Python, kerangka kerja Streamlit dapat mengkonversikan tampilan program menjadi versi website sehingga hasil dari program tersebut lebih mudah dipahami dan digunakan oleh pengguna.

Streamlit adalah pustaka Python sumber terbuka yang memungkinkan pengguna mengubah skrip data menjadi aplikasi berbasis web interaktif. Ini menyediakan cara mudah untuk membuat aplikasi web interaktif dari skrip data. Selain open source, Streamlit juga mempromosikan open sharing, sehingga memudahkan pendistribusian dan berbagi aplikasi yang dibuat menggunakan library (Putranto, dkk, 2023).

Machine learning juga dapat *dideploy* dengan Python Streamlit. Kita semua tahu bahwa visualisasi data sangat penting selama proses analisis. Kita dapat membuat visualisasi data interaktif dengan Streamlit. Jadi Streamlit Python ini dapat membantu Anda membuat portofolio data yang akan menarik perhatian perekrut.

Output dari Python Streamlit selain open source juga open sharing sehingga mudah untuk didistribusikan dan dibagikan. Python Streamlit juga dapat berfungsi sebagai alat untuk Ilmu Data untuk melaporkan dan menyajikan hasil analisis melalui visualisasi menggunakan Python.



SEKOLAH PASCASARJANA