BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Bagian ini menjelaskan beberapa penelitian sebelumnya yang terdiri dari sejumlah tulisan hasil riset yang telah dilakukan sebelumnya, dan digunakan sebagai acuan dalam menyusun penelitian ini.

Tabel 2.1 Penelitian Terdahulu

Penelitian Terdahulu 1	
Judul	*Effect of Distance Metrics in Determining* K-*Value in* K-Means *Clustering Using Elbow and Silhouette Method*
Nama Penulis	Danny Matthew Saputra, Daniel Saputra, *and* Liniyanti D. Oswari
Sumber Jurnal	Atlantis Press, 172 (2) [22]
Tahun	2020
Permasalahan	Kinerja dari permasalahan ini bahwa K-Means menjadi salah satu algoritma pengelompokan berbasis partisional yang sangat populer adalah K-Means. K-Means adalah algoritma untuk mengelompokkan data menjadi K kelompok dan berdasarkan jarak mereka ke *centroid*nya. Sebab, partisional, beberapa faktor yang harus ditentukan sebelum menggunakan K-Means adalah nilai K. Menentukan nilai K merupakan masalah besar karena tidak ada cara *universal* untuk menemukan nilai K.
Framework	Rstudio
Pembahasan	Pengelompokan K-Means adalah salah satu metode pengelompokan yang paling umum digunakan. Ini dapat diterapkan dalam banyak situasi yang berbeda dengan hasil yang cukup baik. Dalam makalah ini, membahas masalah periodisitas dalam dataset dan menyajikan algoritma K-Means periodik yang memodifikasi pendekatan asli. Dengan metode *Elbow* dan *Silhouette* menentukan nilai K, langkah lain seperti menganalisis ukuran eksternal harus dilakukan untuk menentukan apakah *cluster* tersebut baik atau tidak.
Penelitian Terdahulu 2	
Judul	Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means *Clustering*
Nama Penulis	Aditia Yudhistira dan Rio Andika
Sumber Jurnal	*Journal of Artificial Intelligence and Technology Information* (JAITI) [23]
Tahun	2023
Permasalahan	Permasalahan yang muncul dalam mengelola data adalah memperoleh data siswa selama pembelajaran sebagai informasi yang relevan dari basis data. Setiap tahun, pertumbuhan data siswa meningkat seiring dengan peningkatan jumlah siswa, mengakibatkan akumulasi data yang besar yang belum diolah secara optimal. Bahwa, dalam proses menganalisa menggunakan metode algoritma K-Means dalam mengolah data menjadi sebuah informasi dan pengetahuan menjadi bahan pertimbangan pihak sekolah.
Framework	Python dan Google Colab

Pembahasan	Pembahasan pada analisis pengelompokan data nilai siswa menggunakan metode K-Means *clustering* menunjukkan bahwa berdasarkan *clustering* data siswa menggunakan *dataset* siswa selama satu semester. Hasil pengujian menggunakan *silhouette coefficient* maka jumlah *cluster* yang baik yang digunakan dari nilai *silhouette coefficient cluster* lainnya.
Penelitian Terdahulu 3	
Judul	K-Means *Clustering* Video *Trending* di Youtube Amerika Serikat
Nama Penulis	Kevin Widjaja dan Raymond Sunardi Oetama
Sumber Jurnal	Ultima InfoSys: Jurnal Ilmu Sistem Informasi,9 (2) [24]
Tahun	2020
Permasalahan	YouTube bukan hanya sebuah situs untuk berbagi video yang dimaksudkan untuk hiburan semata, namun juga merupakan platform yang memiliki potensi untuk menjadi alat dalam meningkatkan penjualan suatu perusahaan. Dengan jumlah video yang diunggah dan ditonton setiap hari yang begitu banyak, menjadi sulit bagi suatu video untuk menonjol dan menjadi viral atau tren. Bahwa, pengelompokan video *trending* di Amerika Serikat dan mencari tahu faktor-faktor apa saja yang bisa memengaruhi setiap *cluster* agar menjadi viral dengan menggunakan algoritma K-Means.
Framework	Rstudio
Pembahasan	Berdasarkan hasil pembahasan, memberikan informasi tentang pengelompokan video yang sedang tren di YouTube Amerika Serikat. Diketahui bahwa terdapat 3 *cluster* dengan karakteristik masing-masing. *Cluster* dengan jumlah tayangan, suka, dan tidak suka paling sedikit adalah *cluster* dengan anggota terbanyak, sementara *cluster* dengan jumlah tayangan, suka, dan tidak suka terbanyak memiliki anggota paling sedikit. Secara umum, semakin banyak orang yang menyukai suatu video, cenderung semakin populer video tersebut.
Penelitian Terdahulu 4	
Judul	Pola *Cluster Geospatial* Eksplorasi Kejahatan Narkoba di DKI Jakarta
Nama Penulis	Raymond Sunardi Oetama, Tan Thing Heng, dan David Tjahjana
Sumber Jurnal	Ultima InfoSys: Jurnal Ilmu Sistem Informasi, 9 (1) [25]
Tahun	2020
Permasalahan	Kemajuan hasil pembangunan tersebut dapat terhambat jika tingkat kejahatan masih tinggi. Dari tahun 2000 hingga 2017, setiap kejahatan yang terjadi di Indonesia rata-rata terjadi setiap 1 menit 33 detik, menunjukkan tingkat kejahatan yang masih sangat tinggi. Bahwa, pengelompokan dari masalah kejahatan tetap menjadi perhatian utama di Indonesia.
Framework	Tableau Workbooks
Pembahasan	Berdasarkan hasil, Penelitian ini menemukan pola eksplorasi *cluster* geospasial kejahatan narkoba di DKI Jakarta. Berdasarkan jumlah total kejahatan di wilayah DKI Jakarta, *cluster* wilayah dengan tingkat kejahatan tinggi meliputi Jakarta Utara, Jakarta Pusat, dan Kepulauan Seribu. Sekilas, *cluster* wilayah dengan tingkat kejahatan menengah adalah Jakarta Barat dan Jakarta Selatan. Jakarta Timur termasuk dalam *cluster* wilayah dengan tingkat kejahatan rendah. Kejahatan narkoba merupakan jenis kejahatan terbesar di DKI Jakarta. Polda Metro Jaya berhasil menangkap jumlah narkoba terbanyak, terutama jenis shabu. Pelaku kejahatan narkoba paling banyak ditemukan pada usia lebih dari 29 tahun. Metode K-Means yang distribusi narkoba paling umum adalah melalui pos, diikuti oleh transportasi udara, serta distribusi melalui darat dan udara kurang diminati oleh pelaku kejahatan narkoba.

Penelitian Terdahulu 5	
Judul	*Crime Data Analysis in* Python *using* K - Means *Clustering*
Nama Penulis	Md Abu Saleh
Sumber Jurnal	*International Journal for Research in Applied Science & Engineering Technology* (IJRASET) [26]
Tahun	2019
Permasalahan	kejahatan dapat didefinisikan sebagai tindak pidana terhadap siapa pun atau sebuah organisasi dengan maksud untuk merugikan mereka secara langsung atau tidak langsung yang ilegal dan dapat dihukum sesuai hukum negara. Sebab, kejahatan-kejahatan ini semakin meningkat, Bahwa, ada kebutuhan untuk mengendalikannya dan ini menciptakan tekanan besar pada departemen penyelidikan. Harus ada sistem yang dapat menganalisis kejahatan dan departemen kepolisian dapat memanfaatkan teknologi ini yang dapat membuat tugas mereka lebih mudah untuk menyelidiki kasus berdasarkan tren yang berbeda selama bertahun-tahun. Analisis kejahatan adalah teknik penegakan hukum yang melibatkan analisis sistematis untuk identifikasi dan analisis tren dan pola. Informasi tentang tren dapat membantu agensi keamanan lebih efektif dan efisien dalam mengalokasikan sumber daya. Bahwa, membuat prediksi kejahatan serta menganalisis dan memvisualisasikan pola kejahatan di kota Chicago berdasarkan *dataset*. Untuk tujuan ini, algoritma pengelompokan K-Means yang merupakan teknik pemodelan prediktif, digunakan untuk klasifikasi.
Framework	Python
Pembahasan	Berdasarkan hasil, menyajikan analisis kejahatan di Kota Chicago yang diimplementasikan menggunakan Python dan pengelompokan k-Means. Beberapa pra-pemrosesan diterapkan pada *dataset* untuk membuatnya lebih akurat agar dapat bekerja lebih cepat dan mudah. Kejahatan telah dianalisis dengan bantuan *cluster*. Hasilnya ditemukan baik dan akurat. Secara keseluruhan, teknik ini terbukti layak dan modelnya diterapkan dengan cepat dan efisien.
Penelitian Terdahulu 6	
Judul	*Prediction of the number of* COVID-19 *confirmed cases based on* K-Means-LSTM
Nama Penulis	Shashank Reddy Vadyala, Sai Nethra Betgeri, Eric A. Sherer, *and* Amod Amritphale
Sumber Jurnal	*Science direct* [27]
Tahun	2021
Permasalahan	Pada Desember 2019, pneumonia viral yang tidak diketahui muncul di Wuhan, China. Pada februari 2020, WHO mengidentifikasinya sebagai *coronavirus* dan menamainya COVID-19, yang menyebabkan sindrom pernapasan akut dan merupakan penyakit yang sangat mudah menular. Meskipun langkah-langkah yang diimplementasikan di China selama fase awal penyebaran, beberapa pusat penyebaran telah muncul di seluruh dunia, terutama di beberapa negara Eropa dan di Amerika Serikat (AS). AS memiliki 1,6 juta kasus terkonfirmasi dan 96.662 kematian, yang paling tinggi di dunia, pada tanggal tersebut menurut pembaruan virus corona didunia. Di AS, Negara Bagian Louisiana memiliki 25.739 kasus terkonfirmasi (peringkat ke-9 tertinggi) dan 1.540 kematian (peringkat ke-8 tertinggi) pada tanggal tersebut. Rata-rata usia mereka yang meninggal di Louisiana adalah 70 tahun, dan banyak pasien memiliki diabetes 36,65% dan penyakit kardiovaskular 20,92%. Pada 23 Maret 2020. Bahwa, keputusan ini diambil menggunakan *cluster* K-Means untuk mengingat tingkat infeksi harian secara konsisten antara 1,5% dan 3,5%, dengan rata-rata keseluruhan 2,6% mulai 9 Maret hingga 13 Mei 2020,

	dan proporsi pasien yang dirawat di unit perawatan intensif secara konsisten antara 14 dan 16% yang aktif terinfeksi.
Framework	Python
Pembahasan	Berdasarkan hasil, Peramalan kasus COVID-19 yang akurat merupakan masalah yang signifikan bagi otoritas kesehatan masyarakat untuk secara efisien dan tepat waktu mengkoordinasikan perawatan pasien dan layanan lain yang diperlukan untuk menyelesaikan epidemi ini. Dalam penelitian ini, mengusulkan jaringan saraf K-Means-LSTM untuk mengatasi masalah variasi dan ketepatan dalam memprediksi jumlah kasus yang dilaporkan dalam model SEIR tradisional. Temuan dari studi ini akan membantu kebijakan dan layanan kesehatan untuk mempersiapkan dan memberikan layanan secara efisien dalam menghadapi situasi di negara-negara ini dalam beberapa hari dan minggu ke depan, termasuk perawat, tempat tidur, dan fasilitas perawatan intensif. Data harus diperbarui secara *real-time* untuk perbandingan yang lebih tepat dan prospek masa depan.
Penelitian Terdahulu 7	
Judul	*Discovering the Optimal Number* of *Crime Cluster Using Elbow, Silhouette, Gap Statistics, and* NbClust *Methods*
Nama Penulis	Noviyanti T. M. Sagala *and* Alexander Agung Santoso Gunawan
Sumber Jurnal	ComTech: *Computer, Mathematics and Engineering Applications* 13 (1) [28]
Tahun	2022
Permasalahan	Dalam permasalahan ini, beberapa tahun terakhir kejahatan menjadi penting untuk dianalisis dan dilacak untuk mengidentifikasi tren dan hubungan dengan pola dan aktivitas kejahatan. Secara umum, analisis dilakukan untuk menemukan daerah atau lokasi di mana kejahatan tinggi atau rendah dengan menggunakan berbagai metode pengelompokan, termasuk pengelompokan K-Means. Meskipun jika tidak menggunakan algoritma K-Means dalam teknik ini maka sulit menemukan jumlah *cluster* yang optimal.
Framework	Rstudio dan Azure Studio
Pembahasan	Dalam pembahasan, mengatasi masalah mengestimasi jumlah *cluster* dalam *domain* kejahatan tanpa campur tangan manusia, penelitian dilakukan metode *Elbow, Silhouette, Gap Statistics*, dan NbClust pada dataset Indikator Kejahatan Besar (MCI) tahun 2014-2019. Beberapa tahapan dilakukan untuk memproses dataset kejahatan: pemahaman data, persiapan data, pemodelan *cluster*, dan validasi *cluster*. Dua tahap pertama dilakukan di lingkungan R Studio dan dua tahap terakhir di Azure Studio. Dari hasil eksperimental, metode *Elbow, Silhouette*, dan NbClust menyarankan jumlah *cluster* optimum yang serupa, yaitu dua. Setelah memvalidasi hasil menggunakan metode *Silhouette* rata-rata, penelitian menganggap dua *cluster* sebagai *cluster* terbaik untuk *dataset*. Hasil visualisasi metode *Silhouette* menampilkan nilai 0,73. Kemudian, observasi data terkelompok dengan baik. Ini ditempatkan dalam kelompok yang benar.
Penelitian Terdahulu 8	
Judul	*Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood Affected Areas in* Jakarta
Nama Penulis	Ilham Firman Ashari, Eko Dwi Nugroho, Randi Baraku, Ilham Novri Yanda, dan Ridho Liwardana
Sumber Jurnal	*Journal of Applied Informatics and Computing* (JAIC) Vol.7, No.1 [29]
Tahun	2023

Permasalahan	Dalam permasalahan ini, provinsi DKI Jakarta terdiri dari beberapa kota/kabupaten yang memiliki beberapa wilayah yang sering mengalami banjir. Penelitian ini mengumpulkan data rinci berdasarkan wilayah kelurahan di Provinsi DKI Jakarta yang merupakan titik dengan frekuensi kejadian banjir tertinggi dalam 3 bulan terakhir. Menurut data BMKG, terdapat setidaknya 93 titik yang rentan terhadap banjir di wilayah DKI Jakarta, dengan ketinggian air minimal 10 cm hingga 80 cm. Data yang diproses adalah data wilayah yang tidak memiliki label tingkat keparahan banjir. Peningkatan populasi di suatu wilayah menyebabkan peningkatan kepadatan penduduk di wilayah tersebut. Kepadatan penduduk yang tinggi telah membuat wilayah ini semakin padat dan sulit diimbangi dengan kapasitas daerah resapan air yang dibutuhkannya. Akibatnya, wilayah ini telah menjadi salah satu daerah di Indonesia dengan kasus bencana banjir terbesar.
Framework	Python
Pembahasan	Pembahasan dari penelitian ini, bertujuan untuk menentukan klasifikasi daerah terkena banjir di Jakarta antara yang parah, sedang, dan rendah. Metode yang digunakan adalah metode siku (*elbow*), *Silhouette*, Davidson-Bouldin, dan Calinski-Harabasz pada algoritma K-Means, serta indeks Rand untuk evaluasi. Pengelompokan dengan 3 dan 6 kelompok adalah nilai pengelompokan terbaik berdasarkan Calinski-Harabasz. Dengan menggunakan indeks davies bouldin dari pengamatan, nilai K dengan nilai 6 memiliki nilai Davies-Bouldin terkecil dengan nilai 0,2737. Dengan menggunakan metode siluet, hasil eksperimen mendapatkan nilai terbaik secara berurutan, yaitu K=2, K=3, dan K=6 dengan nilai siluet masing-masing 0,866, 0,854, dan 0,803. Dalam eksperimen ini, berdasarkan metode siku, ditemukan bahwa nilai K terbaik adalah K=3. Hal ini didapatkan karena berdasarkan pengamatan pada penampilan data SSE dibandingkan dengan nilai K.
Penelitian Terdahulu 9	
Judul	Optimisasi Klasterisasi Nilai Ujian Nasional dengan Pendekatan Algoritma K-Means, *Elbow*, dan *Silhouette*
Nama Penulis	Allbila Rahajeng Lashiyanti, Ibnu Rasyid Munthe, dan Fitri Aini Nasution
Sumber Jurnal	Jurnal Ilmiah Wahana Pendidikan Vol. 6 no. 1 (14) [30]
Tahun	2023
Permasalahan	Permasalahan ini, masih terjadi penyajian yang acak dan tidak terstruktur, sehingga informasi yang disampaikan menjadi kurang teratur. Untuk mengatasi kekacauan tersebut, dilakukan pengelompokan menggunakan metode *data mining*.
Framework	Rstudio
Pembahasan	Dari pembahasan ini, metode yang digunakan melibatkan pengelompokan menggunakan algoritma K-Means untuk menemukan nilai K optimal melalui metode optimasi *elbow* dan *silhouette*, dengan ukuran dataset sebanyak 1536 rekaman. Melalui optimasi *elbow* dan *silhouette*, nilai K optimal yang diperoleh adalah K=3 untuk metode optimasi *elbow* dan K=2 untuk metode optimasi *silhouette*. Penelitian ini fokus pada analisis nilai rata-rata Ujian Nasional di Provinsi Jawa Tengah untuk Sekolah Menengah Kejuruan, dengan mempertimbangkan atribut seperti Bahasa Indonesia, Bahasa Inggris, Matematika, dan Kompetensi. Harapannya, penelitian ini dapat menambah serta memperjelas informasi yang mendukung proses pengambilan keputusan.

Penelitian Terdahulu 10	
Judul	Perbandingan Evaluasi Metode Davies Bouldin, *Elbow* dan *Silhouette* pada Model *Clustering* dengan Menggunakan Algoritma K-Means
Nama Penulis	Muhammad Sholeh dan Khurotul Aeni
Sumber Jurnal	STRING (Satuan Tulisan Riset dan Inovasi Teknologi) Vol. 8 No. 1 [31]
Tahun	2023
Permasalahan	Permasalahan ini, pembuatan kelompok konsumsi produk kosmetik, pengelompokan data kasus Covid, pembayaran transaksi, dan komoditas perikanan. Dalam penelitian Amanda, dia mengelompokkan nama-nama produk untuk menganalisis penjualan yang sukses dan yang kurang sukses di pasaran. Sementara itu, Hardiani mengelompokkan wilayah provinsi menjadi beberapa *cluster* berdasarkan penyebaran Covid-19.
Framework	Python
Pembahasan	Pembahsan pada penelitian ini, data akan dikelompokkan menggunakan algoritma K-Means, dan hasil pengelompokan akan dievaluasi dengan membandingkan beberapa metode evaluasi. Evaluasi ini penting untuk menentukan jumlah kelompok yang optimal. Pengujian dilakukan mulai dari dua hingga empat belas kelompok untuk mendapatkan hasil terbaik. Metode evaluasi yang digunakan meliputi Davies Bouldin, *Elbow*, dan *Silhouette*. Hasil penelitian menunjukkan bahwa semua metode evaluasi merekomendasikan pengelompokan dua kelompok sebagai yang terbaik.

Tabel 2.1 ini dapat disimpulkan bahwa terdapat ada beberapa persamaan dengan penelitian-penelitian sebelumnya melalui tabel di atas, yang mencakup melakukan penginputan menggunakan K-Means dari berbagainya macam *framework* lainnya. Permasalahan yang terjadi di setiap perusahaan berbeda-beda. Pada penelitian ini keunggulan dari keterkaitan penelitian ini menggunakan dua metode evaluasi, yaitu *Elbow* dan *Silhouette*, untuk memverifikasi nilai K optimal. Hal ini dapat meningkatkan mendukung kredibilitas hasil penelitian. Sedangkan pada kekurangan pada penelitian terdahulu, adalah penggunaan dua *framework* yang berbeda (RStudio untuk pemahaman dan persiapan data, dan Azure Studio untuk pemodelan) dapat menyebabkan fragmentasi proses dan potensi kesulitan dalam integrasi pada analisa data. Berdasarkan beberapa penelitian terdahulu, terdapat *research gap* dalam penelitian ini yang berkaitan dengan pemilihan algoritma untuk melakukan analisa data [22] [23] [24] [25] [29] [30] [31]. Kebaruan pada penelitian ini adalah penerapannya pada *dataset* data pelanggan dari PT XYZ dalam konteks industri sepeda motor di Indonesia. Serta dengan menggunakan metode *Elbow* dan *Silhouette* secara bersamaan untuk menentukan jumlah *cluster* yang optimal bahwa hasil semua metode evaluasi ada 4 pengelompokkan seperti tipe motor, kreditor, provinsi, dan umur. Selain itu, pada penelitian ini adanya rekomendasi algoritma K-Means

clustering yang diberikan ada metode *Elbow* dan *Silhouette* serta *Scatter plot* yang memiliki hasil terbaik.

2.2 Teori tentang Skripsi

2.2.1 Analisis Data

Analisis data merupakan rangkaian langkah pengolahan data dengan tujuan mengungkapkan informasi berharga sebagai landasan untuk pengambilan keputusan guna menyelesaikan suatu permasalahan. Proses analisis mencakup kegiatan mengelompokkan data berdasarkan ciri-cirinya, membersihkan data, mengubah format data, membangun model data, hingga menemukan informasi krusial dari kumpulan data tersebut. Analisis data yang telah melalui tahapan ini perlu disajikan secara menarik dan sederhana, seringkali menggunakan format visual seperti bagan atau grafik. Pemanfaatan teknologi saat ini memiliki dampak besar pada sebagian besar aspek kehidupan, dan keterkaitannya dengan data terus berkembang secara berlangsung seiring berjalannya waktu [33].

Analisis data mempunyai Pembahasan mengenai langkah-langkah pengolahan data dan informasi yang diperoleh selama melakukan penelitian, bertujuan untuk menghasilkan hasil penelitian yang signifikan. Dalam analisis data memiliki pemeriksaan terhadap instrumen penelitian, seperti dokumen, catatan, dan rekaman dalam suatu penelitian, merupakan suatu kegiatan evaluasi terhadap alat atau metode yang digunakan dalam riset tersebut. Teknik analisis ada 2 jenis antara lain:

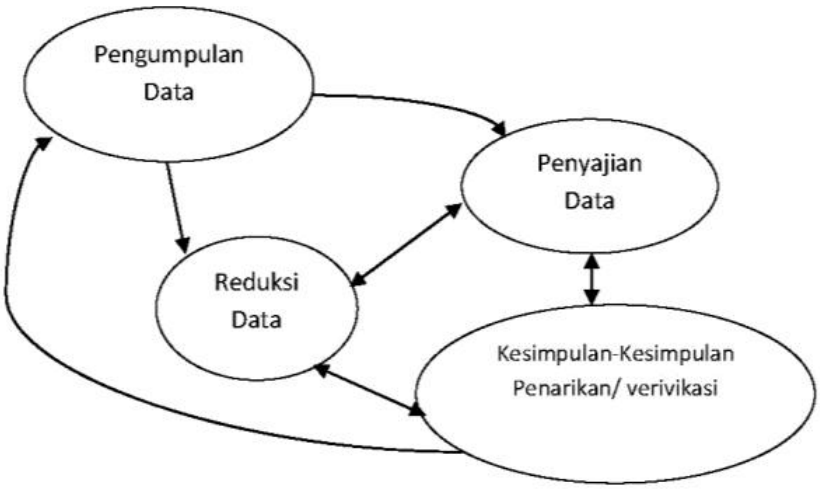
1) Analisis data kualitatif merupakan teknik analisis data yang tidak melibatkan angka atau bersifat nonnumerik. Dalam tahapannya, teknik analisis data kualitatif dapat dikelompokkan ke dalam beberapa teknik, termasuk analisis konten, naratif, dan wacana.
    1. Analisis konten merupakan merujuk pada proses analisis isi. Dalam metode analisis konten, penting bagi peneliti untuk sepenuhnya memahami dan menyelidiki informasi yang dikumpulkan selama penelitian, sehingga dapat diolah dengan

baik nantinya. Melalui pemahaman yang mendalam ini, peneliti dapat mengelompokkan informasi dari data yang ada, dimulai dari yang umum hingga yang spesifik, untuk mempermudah proses pengolahan data.

2. Analisis naratif merupakan metode yang digunakan untuk menganalisis data penelitian dengan pemberian fokus utama pada bagaimana suatu ide dapat diidentifikasi dari narasi atau data secara menyeluruh. Secara umum, teknik analisis naratif digunakan untuk melakukan interpretasi terhadap penilaian pelanggan dan proses operasional.
3. Analisis wacana merupakan metode kualitatif berfokus pada cara menganalisis penggunaan bahasa secara alami, baik dalam bentuk lisan maupun tulisan.

2) Analisis data kuantitatif adalah proses penafsiran dan pemahaman data yang bersifat numerik atau dapat diukur. Metode ini melibatkan penggunaan statistik dan teknik matematika untuk mengidentifikasi pola, hubungan, atau tren dalam data numerik. Dalam tahapannya, teknik analisis data kuantitatif dapat dikelompokkan ke dalam beberapa teknik, termasuk teknik deskriptif dan inferensial.

1. Teknik deskriptif merupakan pendekatan analisis data kuantitatif yang bertujuan untuk mengevaluasi karakteristik suatu set data. Metode ini dapat digunakan ketika dihadapkan pada data dalam jumlah besar, seperti dalam kasus data sensus penduduk. Sebelum diterapkan, peneliti sebaiknya mengidentifikasi jenis data yang sedang digunakan.
2. Teknik inferensial adalah suatu pendekatan analisis yang menggunakan rumus statistik. Hasil perhitungan dari rumus ini kemudian digunakan untuk menyusun kesimpulan yang berlaku secara umum. Dengan ini, dari penjelasan tersebut, dapat disimpulkan bahwa analisis data inferensial dapat diterapkan secara umum.



Gambar 2.1 Komponen Analisis Data
Sumber: [33]

Gambar 2.1 merupakan tahapan komponen analisis data yang terdiri dari pengumpulan data, penyajian data, reduksi data, serta kesimpulan dan verifikasi berikut penjelasan pada komponen analisis data yakni:

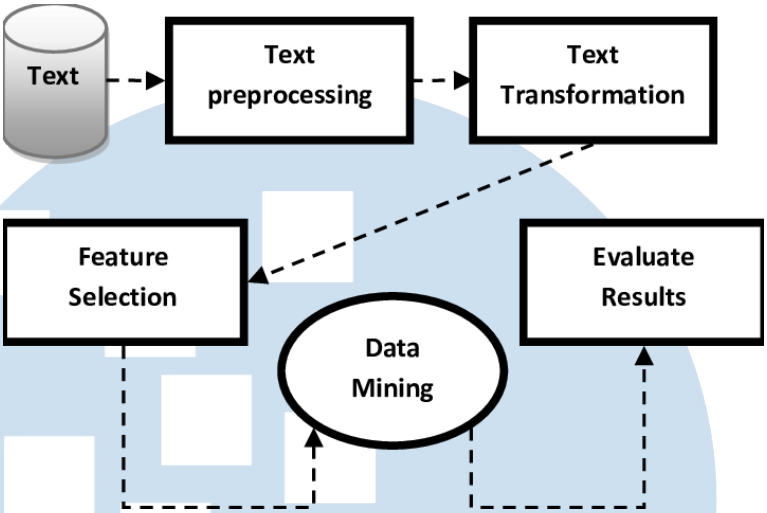
1) Pengumpulan data merupakan cara atau metode yang digunakan untuk mengumpulkan informasi atau data dari suatu penelitian atau studi seperti wawancara, kuesioner, observasi dan dokumentasi. Dengan mengatur data, menentukan elemen yang signifikan, dan menyusun kesimpulan, informasi dapat dipresentasikan dengan cara yang jelas dan dapat dipahami baik oleh diri sendiri maupun oleh orang lain. Dengan memanfaatkan data, penelitian dapat memperoleh solusi terhadap permasalahan yang telah dirumuskan. Bahwa, sangat penting untuk memilih dengan cermat teknik pengumpulan data agar data yang diperoleh memiliki validitas dan reliabilitas yang tinggi.
2) Penyajian data merupakan proses mengatur dengan sistematis sekumpulan data agar mudah dipahami, membuka peluang untuk menyimpulkan. Penyajian data kualitatif dapat berupa teks naratif seperti catatan lapangan, matriks, grafik, jaringan, atau bagan. Dengan penyajian data ini, informasi akan tersusun dalam pola hubungan, memudahkan pemahaman.

3) Reduksi data merupakan proses menyederhanakan, mengelompokkan, dan menghapus elemen yang tidak diperlukan dari data, dengan tujuan menghasilkan informasi yang relevan dan mempermudah pembuatan kesimpulan.
4) Kesimpulan dan verifikasi merupakan fase akhir dalam teknik analisis data kualitatif. Pada tahap ini, fokus tetap pada tujuan analisis yang ingin dicapai, meskipun data telah direduksi. Tujuan dari tahap ini adalah menggali makna dari data yang terkumpul dengan mencari hubungan, persamaan, atau perbedaan untuk menyimpulkan jawaban terhadap permasalahan yang ada.

2.3 Teori tentang algoritma yang digunakan

2.3.1 Text Mining

Text mining merupakan sebuah cara mengenali mengekstrak pola dari teks dan mengubahnya menjadi laporan atau data yang dapat digunakan untuk mengorganisir pengaturan menjadi serangkaian pola yang terstruktur. *Text mining* memiliki tujuan untuk mengenali kata-kata yang mencerminkan isi suatu informasi, sehingga memungkinkan analisis tentang hubungan informasi tersebut. Proses dari *text mining* sering dijelaskan sebagai suatu tahap informasi telah berlangsung penting, di mana pengguna berpartisipasi dengan berbagai dokumen-dokumen setelah menggunakan *tools* untuk dianalisis untuk jangka waktu tertentu [34]. Proses *text mining* terutama dipengaruhi oleh konsep *data mining*.



Gambar 2.2 Tahapan *Text Mining*
Sumber: [34] [35]

Pada gambar 2.2 merupakan siklus-siklus dari *text mining* terdiri dari lima tahap siklus, yakni *preprocessing*, *transformation*, *feature selection*, *data mining*, dan *evaluate*.

Pada tahap pertama *preprocessing* disebut juga sebagai pemrosesan data merupakan Langkah-langkah persiapan dan pembersihan data teks sebelum melakukan analisis atau model dalam analisis lebih lanjut dimaksudkan untuk mengubah data teks belum diolah menjadi format yang bertambah terstruktur dan mudah untuk diolah. Proses ini umumnya melibatkan *case folding*, tokenisasi, serta *data cleaning* untuk menghapus simbol atau tanda baca khusus.

Dalam tahap kedua *transformation*, yang juga dikenal sebagai transformasi data, data teks yang sudah melalui tahap *preprocessing* diubah menjadi bentuk numerik untuk pengolahan lebih lanjut. Tujuan dari langkah ini adalah mengubah data teks ke format yang dapat dimengerti dan diproses oleh model atau algoritma pembelajaran mesin. Langkah ini sering mencakup perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF).

Pada tahap ketiga *feature section* yang merupakan merujuk pada proses pemilihan fitur atau kata-kata tertentu dari *set* data teks yang relevan atau signifikan untuk analisis lebih lanjut. Dalam *text mining*, fitur dapat berupa kata-kata atau istilah yang digunakan untuk mewakili atau menggambarkan suatu dokumen atau teks. Proses *feature selection* menjadi penting karena tidak semua kata atau fitur pada teks memiliki kontribusi yang sama terhadap tujuan analisis.

Beberapa kata mungkin tidak informatif atau sering muncul dan tidak memberikan banyak nilai dalam membedakan antar dokumen. Bahwa, dengan memilih fitur yang paling relevan atau bermakna, bahwa dapat meningkatkan efisiensi analisis dan meningkatkan performa model dalam *text mining*.

Pada tahap keempat *data mining* dalam kondisi *text mining* bertujuan untuk mengenali pola tren, hubungan, dan wawasan yang terselubung dalam data teks. Tahap ini biasanya melibatkan pengembangan model atau pola yang digunakan dalam proses *data prosessing*.

Pada tahap terakhir evaluasi. Evaluasi ini merupakan suatu yang memanfaatkan untuk Mengevaluasi kinerja model atau algoritma yang sudah dikembangkan dari proses *data mining*. Dengan memanfaatkan metrik evaluasi yang relevan, seperti akurasi, presisi, *recall*, *f1-score*, atau metrik spesifik lainnya, langkah ini membantu dalam menilai seberapa efektif model dalam melakukan klasifikasi, pengelompokan, atau prediksi yang akurat berdasarkan data teks [34] [35].

2.3.2 K-Means

K-Means merupakan suatu algoritma metode non-hirarki untuk *clustering data* yang bertujuan untuk membagi data ke dalam satu atau lebih kelompok. Proses ini mengelompokkan data dengan karakteristik serupa ke dalam satu kelompok, sementara data dengan karakteristik yang berbeda ditempatkan dalam kelompok yang berbeda. Tujuan utama dari *clustering* ini adalah untuk mengurangi nilai *objective function* yang telah ditetapkan, yang biasanya mencoba untuk mengurangi variasi di dalam satu kelompok dan meningkatkan variasi antar kelompok [36]. Metode K-Means berupaya mempartisi data menjadi satu atau lebih kelompok tanpa hirarki, dengan memasukkan data yang memiliki karakteristik serupa ke dalam kelompok yang sama. Beberapa faktor yang perlu dipertimbangkan ketika menggunakan metode K-Means dalam pengelompokan data termasuk sensitivitas terhadap inisialisasi *centroid* awal dan penentuan jumlah kelompok yang optimal. Saat keberadaan K-Means sebagai salah satu metode pengelompokan tanpa arahan

(*unsupervised*) memungkinkan pengelompokan data tanpa adanya label sebelumnya. Namun, hasil *clustering* dari K-Means dapat digunakan sebagai langkah pra-pemrosesan dalam pengklasifikasi dengan arahan (*supervised*) [36]. K-Means *clustering* adalah sebuah teknik pengelompokan data non-hirarki yang mengatur data ke dalam satu atau lebih kelompok. Data yang memiliki kesamaan dalam karakteristik dikelompokkan bersama dalam satu kelompok, sementara data dengan karakteristik yang berbeda ditempatkan dalam kelompok yang berbeda. Dengan demikian, setiap kelompok memiliki variasi yang minim dalam data yang terkandung di dalamnya [37].

$$\mu k = \frac{1}{Nk} \sum_{q=1}^{Nk} xq \qquad (2.1)$$

Keterangan:

uk = Titik *centroid* dari *cluster* ke - K

N*k* = Banyaknya data pada *cluster* ke - K

Xq = Data ke q pada *cluster* ke – K

Dengan metode *cluster* K-Means ini, tentunya akan cukup rumit untuk menginterpretasikan koefisien. K-Means telah mengalami perkembangan yang memungkinkannya untuk memodelkan *dataset* dengan struktur khusus menggunakan teknik kernel. Terdapat sejumlah permasalahan yang harus dipertimbangkan saat menggunakan metode K-Means, termasuk variasi dalam model *clustering* serta pemilihan model yang sesuai dengan karakteristik *dataset* yang sedang dianalisis. *Unsupervised learning* adalah salah satu metode dalam *machine learning* di mana model dilatih pada *dataset* yang tidak memiliki label atau target *output*. Salah satu teknik utama dalam *unsupervised learning* adalah *clustering*. *Clustering* adalah teknik *unsupervised learning* yang bertujuan untuk mengelompokkan data ke dalam beberapa *cluster* atau kelompok berdasarkan kemiripan atau kesamaan tertentu di antara data-data tersebut. Data dalam satu

cluster akan memiliki karakteristik yang lebih mirip satu sama lain dibandingkan dengan data yang berada di *cluster* yang berbeda. Tujuannya adalah memisahkan data ke dalam kelompok yang berbeda untuk analisis lebih lanjut, misalnya segmentasi pelanggan dalam bisnis serta memahami dan mengungkap struktur atau pola tersembunyi dalam data [32]. Tujuannya adalah untuk menemukan pola atau struktur yang tersembunyi dalam data. Berikut adalah penjelasan lebih lanjut mengenai data PT XYZ:

1) Dalam data PT XYZ yang digunakan tidak memiliki training dan validasi yang ditentukan sebelumnya. Seperti, dalam data pelanggan. Data ini hanya memiliki informasi seperti tipe motor, kreditor, provinsi, dan umur.
2) Data PT XYZ memiliki atribut atau karakteristik dari data yang digunakan untuk menemukan titik *cluster*. Misalnya, dalam data penjualan motor, fitur bisa termasuk, tipe motor, kreditor, provinsi, dan umur pelanggan.
3) Pemilihan data dari *Unsupervised learning* menjadi algoritma cocok adalah K-Means. Karena dengan menggunakan algoritma K-Means yang dimana data pelanggan PT XYZ dikelompokkan berdasarkan kesamaan fitur. Tujuan utamanya adalah untuk mengelompokkan data ke dalam cluster yang berbeda.

2.4 Teori tentang Tools yang digunakan

2.4.1 Python

Python merupakan bahasa pemrograman tingkat tinggi yang diciptakan oleh Guido van Rossum. Bahasa pemrograman ini mendukung pemrograman berorientasi objek (OOP) dan menyediakan antarmuka lintas platform. Kemampuan Python dapat diperluas melalui beragam paket yang tersedia di perpustakaan Python, yang mencakup lebih dari 100.000 paket dengan berbagai fungsi. Pengembang dapat memanfaatkan paket-paket Python ini untuk mendapatkan fleksibilitas dan kenyamanan ekstra dalam proses pembuatan model. Salah satu keunggulan Python yang dikenal luas adalah kemudahannya bagi pemula untuk mempelajarinya. Kode-kode dalam Python memiliki struktur

yang lebih mudah dipahami karena menggunakan sintaks yang mirip dengan bahasa Inggris, sehingga memudahkan pemula untuk memahaminya. Python juga memiliki berbagai fungsi yang dapat dijalankan dengan mudah melalui penggunaan *library-library* yang tersedia di dalamnya. Python memiliki beragam aplikasi yang melibatkan pembuatan situs *web*, pengolahan data, dan bahkan pembuatan *game*. Python dikenal sebagai suatu bahasa pemrograman yang memiliki beragam pustaka *opensource* yang sangat komprehensif dan terstruktur. Kemampuan Python diakui dalam menangani berbagai aspek seperti pembentukan *Big Data, Data Mining, Data Science, Deep Learning*, dan yang sedang populer, yaitu *machine learning* [39]. Berikut kelebihan menggunakan *tools* Python dengan Jupyter NoteBook yakni:

1) *Cloud-based*: Jupyter Notebook tidak memiliki sistem penyimpanan *cloud*, ini dikarekan Jupyter Notebook dijalankan pada *local machine* dan disimpan ke dalam *hard disk* laptop atau komputer.
2) *File Syncing*: Hanya laptop atau komputer yang tentunya harus memiliki *file* yang sama yang bisa membuka Jupyter Notebook.
3) *File Sharing*: Jupyter Notebook tidak memiliki fitur untuk berkolaborasi.
4) *Library Install*: Dengan Jupyter Notebook, pengguna harus menginstal setiap *library* yang ingin gunakan ke perangkat pengguna menggunakan pip atau pengelola paket lainnya. Pengguna juga akan dibatasi oleh RAM, ruang disk, GPU, dan CPU yang tersedia di komputer.
5) *File View Without Install*: Jupyter Notebook berbasis *local*, jadi setiap ada *library* atau *file* baru yang digunakan harus menginstall terlebih dahulu.