BAB 2
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Pada tahapan ini, kajian pustaka dilakukan guna mengumpulkan informasi serta data yang berkaitan dengan judul penelitian. Kajian pustaka dilakukan untuk memberikan landasan teori mengenai metode yang akan dilakukan dalam penelitian ini. Kajian pustaka merupakan suatu rangkuman yang diambil dari suatu sumber bacaan terkait topik penelitian. Ada beberapa alasan mengapa kajian pustaka diperlukan untuk mendapatkan hasil yang relevan dengan variabel dalam penelitian, seperti untuk memperkirakan keberhasilan penelitian, memperkuat alasan mengenai pentingnya pelaksanaan penelitian, serta memahami jenis metode dan teknik yang digunakan dalam penelitian terdahulu [9]. Setelah itu informasi serta data yang telah dikumpulkan tersebut dijadikan sebagai informasi pendukung serta pembanding dalam penelitian yang lagi dicoba. Penelitian ini sendiri melakukan kajian literatur pada sepuluh jurnal yang terkait dengan judul dari penelitian ini. Adapun sepuluh jurnal yang terdiri dari enam jurnal nasional serta empat jurnal internasional dengan jurnal sangat terkini yakni pada tahun 2022 serta jurnal yang terlama yaitu pada tahun 2018. Penelitian sebelumnya yang menjadi rujukan dalam penelitian ini adalah penelitian [10] sebagai jurnal acuan terkait metode tersebut. Sudah banyak dilakukan dan diterapkan penelitian yang bertujuan untuk membantu mengolah dan mengelompokkan data dalam berbagai bidang. Metode *K-Means Clustering* merupakan salah satu teknik dalam bidang data mining yang sering digunakan untuk mengelompokkan sejumlah besar data menjadi beberapa kelompok yang berbeda.

Selanjutnya, sepuluh jurnal tersebut akan disusun ringkasannya dengan menggunakan kerangka 3C2S, yaitu *Compare* untuk mencari kesamaan, *Contrast* untuk mencari perbedaan, *Criticize* untuk memberikan kritik, *Synthesize* untuk menghasilkan ide baru, dan *Summarize* untuk meringkas. Dengan mengetahui kesamaan, perbedaan, dan kekurangan dari penelitian sebelumnya, peneliti dapat menghasilkan ide baru dan meningkatkan kualitas penelitian selanjutnya. Selain itu,

dengan menggunakan kerangka 3C2S, peneliti dapat lebih mudah memahami isi dari sepuluh jurnal penelitian terdahulu dan mengidentifikasi informasi penting yang harus diambil sebagai referensi dalam penelitian yang sedang dilakukan. Ringkasan tersebut akan dimasukkan ke dalam sebuah tabel yang memuat informasi dari sepuluh jurnal penelitian sebelumnya. Dengan adanya tabel yang memuat informasi dari sepuluh jurnal penelitian sebelumnya, akan memudahkan peneliti dalam mengumpulkan data dengan cepat dan terorganisir. Oleh karena itu, penggunaan kerangka 3C2S dan tabel yang memuat referensi dari penelitian terdahulu sangat penting dalam melakukan penelitian yang berkualitas. Referensi sepuluh penelitian terdahulu yang digunakan dapat dilihat pada Tabel 2.1 yang disajikan.

Tabel 2. 1 Penelitian Sebelumnya

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
1	Analisis Mutu Pendidikan Sekolah Menengah Atas Program Ilmu Alam di Jawa Tengah dengan *Algoritme K-Means* Terorganisir [11]	Penelitian ini menggunakan Algoritma *K-Means* sebagai metode pengelompokkan data. Sama seperti penelitian yang dilakukan.	Dalam penelitian ini, digunakan kombinasi metode *K-Means* dan Hirarki untuk menganalisis data nilai Ujian Akhir Nasional (UAN) program studi Ilmu Alam. Tujuannya adalah untuk mengevaluasi kualitas pendidikan di tingkat Sekolah Menengah Atas di wilayah Jawa Tengah. Sedangkan penelitian yang dilakukan hanya menggunakan *K-Means Clustering* yang digunakan untuk pengelompokan destinasi wisata Kabupaten Tegal.	Hanya dilakukan pada Sekolah Menengah Atas Program Ilmu Alam di Jawa Tengah.	Penelitian yang dilakukan untuk pengelompokkan data sama seperti penelitian [11] dengan tujuan penelitian untuk mengevaluasi.	Dalam penelitian ini, digunakan metode pengelompokan menggunakan *algoritme K-Means* yang dikombinasikan dengan hirarki dan menentukan jumlah *cluster* optimal menggunakan kriteria *Bayesian information Criterion (BIC)*, dan diperoleh 5 kelompok yang optimal.

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
2	Teknik *Data Mining* Untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data *Principal Component Analysis (PCA)* [12]	Melakukan penelitian dengan pengelompokkan data sama seperti penelitian yang dilakukan.	Penelitian ini menggunakan tiga metode Pengklasifikasian antara lain : *ANN, SVM,* dan *Decision Tree.* Sedangkan penelitian yang akan dilakukan hanya menggunakan *K-Means Clustering.*	Dalam penelitian ini masih terdapat penulisan persamaan namun tidak terdapat keterangan dari persamaan tersebut serta tidak adanya contoh perhitungan dari persamaan yang ada.	Penelitian yang dilakukan untuk pengelompokan data sama seperti penelitian [12] agar selaras dengan tujuan penelitian pengelompokan data.	Dari tiga metode data *mining* yang digunakan menunjukkan tingkat akurasi, Metode *SVM-PCA* memiliki akurasi sebesar 91,50%, diikuti oleh metode *ANN-PCA* dengan akurasi sebesar 89,46%, dan metode *Decision-PCA* dengan akurasi sebesar 88,78%.
3	Implementasi *Data mining* Dalam Menentukan Destinasi Unggulan Berdasarkan *Online Reviews Tripadvisor* Menggunakan Algoritma *K-Means* [13]	Penelitian ini melakukan metodologi yang sama yaitu *Clustering* dan *K-Means.* Sama seperti penelitian yang dilakukan.	Penelitian ini dilakukan untuk menentukan destinasi unggulan berdasarkan online *reviews tripadvisor* menggunakan aplikasi *RapidMiner* sebagai alat bantu pemrosesan data. Sedangkan penelitian ini dilakukan untuk mengetahui tingkat	Penilitian ini hanya dilakukan berdasarkan *online reviews* pada satu aplikasi.	Penelitian yang dilakukan mengenai penerapan metode *K-Means Clustering* sama seperti penelitian [13] agar selaras dengan perhitungan data.	Hasil perhitungan *K-Means Clusering* diketahui *cluster* 1 memiliki rata-rata lebih tinggi daripada *cluster 2.* Didapatkan hasil destinasi terbaik di Kawasan Asia Timur yaitu *picnic/parks spot, religion institution, beach, resorts dan theaters.*

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
			popularitas destinasi wisata dengan menggunakan aplikasi *WEKA* sebagai alat bantu pemrosesan data.			
4	Analisis Pengelompokan Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Provinsi Jambi Berbasis *Data Mining* [10]	Penelitian ini mengunakan metode *K-Means Clustering* untuk pengelompokkan data. Sama seperti penelitian yang akan dilakukan.	Penelitian ini dilakukan untuk menganalisa pengelompokan kunjungan wisatawan ke objek penelitian menggunakan aplikasi *RapidMiner*. Sedangkan penelitian yang dilakukan untuk pengelompokan destinasi wisata menggunakan aplikasi *WEKA.*	Penelitian ini tidak melakukan validasi sehingga hasil yang didapatkan belum dipastikan sesuai dengan persepsi pengunjung.	Penelitian yang dilakukan mengenai penerapan metode *K-Means Clustering* sama seperti penelitian [10] agar selaras dengan penerapan metode penelitian.	Menghasilkan 3 *cluster*, yaitu Dari 174 tempat wisata unggulan di provinsi Jambi, 32 di antaranya termasuk dalam kategori rendah (C0) dan 142 di antaranya termasuk dalam kategori sedang (C1). Setelah dilakukan analisis, ditemukan 10 objek wisata unggulan dengan jumlah pengunjung tertinggi (C2).
5	Penerapan Metode *Clustering* Untuk Pengelompokan Potensi Wisata Di	Melakukan penelitian menggunakan Metode K-*Means Clustering.* Sama	Penelitian ini melakukan pengelompokan potensi wisata di Kabupaten	Penelitian ini menyantumkan beberapa persamaan namun tidak terdapat	Penelitian yang dilakukan mengenai pengelompokan destinasi wisata sama seperti penelitian [14]	Penelitian ini dilakukan untuk membangun sistem kelompok prioritas objek wisata yang

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
	Kabupaten Sumedang [14]	seperti penelitian yang dilakukan.	Sumedang. Sedangkan penelitian yang dilakukan melakukan penengelompokan destinasi wisata untuk mengetahui tingkat popularitas di Kabupaten Tegal.	keterangan dari persamaan tersebut.	agar selaras dengan tujuan penelitian mengelompokkan destinasi wisata.	akan dikembangkan. Dari perhitungan *clustering* yang dilakukan mendapatkan 3 *cluser, cluster 1* sebanyak 12 tempat wisata, *cluster 2* sebanyak 18 tempat wisata, dan *cluster 3* sebanyak 7 tempat wisata.
6	*Data Mining in Tourism Data Analysis : Inbound Visitors to Japan* [15]	Penelitian ini dilakukan berkaitan dengan pariwisata. Sama seperti penelitan yang dilakukan.	Penelitian ini menggunakan metode pohon keputusan dan metode statistik lanjutan dari *data mining.* Sedangkan penelitian yang dilakukan menggunakan metode *K-Means Clustering.*	Dalam penelitian ini hanya mengguanakan satu data dari perusahaan *Japan Travel Bureau (JTB) Foundation* dan data yang digunakan yang lama yaitu tahun 2010	Penelitian yang dilakukan mengenai pariwisata, sama seperti penelitian [15] agar selaras dengan pemahaasan tentang pariwisata	Hasil dari pohon keputusan menunjukkan bahwa terdapat dua kelompok berbeda yaitu wisatawan Asia dan *non-Asia.*
7	*Data Mining Using K-Means Clustering Algorithm for Grouping Countries of Origin of Foreign Tourist* [16]	Penelitian ini dilakukan menggunakan metode *K-Means* dan *Clustering.* Sama seperti	Penelitian ini dilakukan untuk Pengelompokan Negara Asal Wisatawan Asing yang datang ke	Penelitian ini hanya menggunakan data yang sudah disediakan oleh Badan Pusat Statistik	Penelitian yang dilakukan mengenai pengelompokan data sama seperti penelitian [16] agar selaras dengan tujuan	Hasil dari klasterisasi yaitu *relative* 1 terdiri dari negara-negara yang memiliki tingkat kunjungan

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
		penelitian yang dilakukan.	Indonesia. Sedangkan penelitian yang dilakukan untuk pengelompokan destinasi wisata di Kabupaten Tegal.	Pemerintahan Indonesia, namun tidak terdapat sampel data yang digunakan.	penelitian mengelompokan data.	wisatawan ke Indonesia yang *relative* rendah dengan 206 anggota, dan *relative* 2 terdiri dari negara-negara yang memiliki kunjungan wisata ke Indonesia *relative* tinggi dengan 6 anggota
8	*Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K -Nearest Neighbor Data Mining Method and Decision Tree C4.5* [17]	Dalam penelitian ini menggunakan metode *data mining.* Sama dengan penelitian yang dilakukan.	Melakukan penelitian dengan metode algoritma *K-Nearest Neighbor (KNN) and Decision Tree C4.5.* Sedangkan penelitian yang dilakukan menggunakan metode *K-Means Clustering.*	Penelitian ini tidak memberikan detail alur perhitungan, tidak hanya memberikan hasil yang sudah didapatkan dari perhitungan.	Penelitian yang dilakukan mengenai *data mining* sama seperti penelitian [17] yang selaras dengan pembahasan *data mining.*	Penelitian ini menghasil tingkat akurasi dari pembahasan, perhitungan, pengujian dan perbandingan disimpulkan bahwa nilai akurasi terbaik dari perbandingan 2 metode algortima *K-Nearest Neighbor* dengan tingkat akurasi sebesar 59,32%
9	*Performance Comparison of Data Mining Algorithms*	Melakukan penelitian yang berkaitan dengan metode *data*	Penelitian ini melakukan perbandingan performa algoritma	Penjelasan mengenai metode *Decision Tree, SVM,* dan teori	Penelitian yang dilakukan mengenai *data mining* sama seperti penelitian [18]	Kesimpulan yang diperoleh dari penelitian ini dilihat dari nilai *AUC, SVM*

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
	Which Occupy the Top:C4.5 and SVM [18]	*mining.* Sama dengan penelitian yang dilakukan.	*data mining* antara metode *Decision Tree C4.5* dengan metode *Support Vector Modelling (SVM).* Sedangkan penelitian yang dilakukan mengelompokkan data menggunakan metode *K-Means Clustering.*	lainnya masih sangat singkat sehingga masih sulit untuk dipahami	yang selaras dengan pembahasan *data mining.*	lebih unggul dengan nilai 100 dibandingkan C4.5 dengan 0,867, maka akurasi dan *balance error*, dari nilai presisi C4.5 lebih baik yaitu dengan nilai 75,00%, namun dari nilai *recall*, *SVM* lebih baik dengan nilai 100%.
10	Penerapan *Data Mining* Pengelompokan Menu Makanan dan Minuman Berdasarkan Tingkat Penjualan Menggunakan Metode *K-Means* [19]	Penelitian ini membahas *data mining* dengan metode algoritma *K-Means Clustering* untuk mengelompokkan data. Sama seperti penelitian yang dilakukan.	Dalam penelitian ini menggunakan objek kuliner untuk mengelompokkan menu makanan dan minuman di Dpom *Coffee.* Sedangkan pada penelitian yang dilakukan untuk mengelompokan destinasi wisata di Kabupaten Tegal.	Penelitian ini tidak memberikan contoh perhitungan yang dilakukan.	Penelitian yang dilakukan mengenai pengelompokan data sama dengan penelitian [19] yang selaras dengan tujuan penelitian untuk melakukan pengelompokkan data menggunakan *K-Means Clustering* sebagai metode perhitungan.	Menghasilkan 3 *cluster*. *Cluster* 1 menu dengan tingkat penjualan rendah sebanyak 8 menu. *Cluster* 2 dengan menu yang memiliki tingkat penjualan sedang sebanyak 40 menu. *Cluster* 3 dengan menu tingkat penjualan tinggi sebanyak 7.

Pada penelitian [11], melakukan perbandingan menggunakan metode *K-Means* dan hirarki. Diperoleh 5 kelompok terbaik berdasarkan nilai *silhoute*, di mana algoritma *complete K-Means* memiliki nilai 0,4537. Berdasarkan analisis data yang dilakukan, bahwa kota Semarang memiliki sekolah yang paling unggul dengan proporsi sebesar 12,76%, sedangkan sekolah yang terendah terdapat di Boyolali dengan proporsi 9,03%.

Pada penelitian [12], melakukan klasifikasi data ulasan destinasi wisata. Hasil pengolahan data dengan menggunakan tiga metode *data mining* yang berbeda dengan *PCA*, didapatkan bahwa metode *SVM-PCA* memiliki akurasi yang paling baik yaitu mencapai 91,50%. Metode ini diikuti oleh metode *ANN-PCA* dengan akurasi sebesar 89,46% dan metode *Decision-PCA* dengan akurasi sebesar 88,78%.

Pada penelitian [13], menentukan destinasi unggulan menggunakan metode *K-Means Clustering* dan hasil yang didapatkan dalam peneletian ini berupa informasi perjalanan yang lebih baik terkait destinasi berdasarkan nilai *cluster* yang diperoleh dalam implementasi di simulator *WEKA.*

Pada penelitian [10], melakukan analisis pengelompokan kunjungan wisatawan menggunakan algoritma *K-Means Clustering*. Hasil yang didapatkan sebesar 3 *cluster*, yaitu 32 di antaranya termasuk dalam kategori rendah (C0) dan 142 di antaranya termasuk dalam kategori sedang (C1). Setelah dilakukan analisis, ditemukan 10 objek wisata unggulan dengan jumlah pengunjung tertinggi (C2).

Pada penelitian [14], melakukan pengelompokkan potensi wisata di Kabupaten Sumedang menggunakan algortima *K-Means Clustering.* Hasil yang didapatkan sebanyak 3 *cluser, cluster* 1 terdapat 12 destinasi wisata yang mempunyai prioritas tinggi untuk dikembangkan, *cluster* 2 terdapat sebanyak 18 tempat wisata yang mempunyai potensi untuk diadakan *event* untuk meningkatkan jumlah kunjungan dan popularitas, dan *cluster* 3 terdapat sebanyak 7 tempat wisata yang memiliki potensi dan jumlah kunjungan yang tinggi.

Pada penelitian [15], melakukan pengelompokkan pengunjung wisatawan yang masuk ke Jepang dengan menggunakan pohon keputusan (*Decission Tree C.45*). Hasil yang didapatkan dalam penelitian ini menujukkan bahwa terdapat 2 kelompok berbeda yaitu wisatawan Asia dan wisatawan *Non-Asia*. Wisatawan Asia

(62%), seperti Korea (19,51%), Taiwan (18,10%), dan China daratan (14,16%). Kelompok pengunjung terbesar kedua adalah dari Amerika Serikat (10,65%). Dari China daratan, dua kelompok terbesar berasal dari Beijing dan Shanghai. Selain itu, 57,9% responden bepergian untuk pariwisata dan rekreasi, sementara 25% bepergian untuk berpartisipasi dalam pelatihan bisnis, konferensi, atau pameran dagang.

Pada penelitian [16], melakukan proses perhitungan dengan menerapkan metode *K-Means Clustering* didapatkan mendapatkan hasil dari klasterisasi yaitu *relative* 1 terdiri dari negara-negara yang memiliki tingkat kunjungan wisatawan ke Indonesia yang *relative* rendah dengan 206 anggota, dan *relativ* 2 terdiri dari negara-negara yang memiliki kunjungan wisata ke Indonesia *relative* tinggi dengan 6 anggota.

Pada penelitian [17], melakukan penelitian menggunakan metode *K-Nearest Neighbor (KNN)* dan *Decission Tree C4.5* menghasil tingkat akurasi dari pembahasan, perhitungan, pengujian dan perbandingan disimpulkan bahwa nilai akurasi terbaik dari perbandingan 2 metode algortima *K-Nearest Neighbor* dengan tingkat akurasi sebesar 59,32%.

Pada penelitian [18], Penelitian ini memperoleh hasil yang menggambarkan bahwa secara umum terdapat kelebihan dan kekurangan pada kedua algoritma tersebut. Dilihat dari nilai *AUC, SVM* lebih unggul dengan nilai 100 dibandingkan *Decission Tree C4.5* dengan 0,867, maka akurasi dan *balance error*, dari nilai presisi *Decission Tree C4.5* lebih baik yaitu dengan nilai 75,00%, namun dari nilai *recall, SVM* lebih baik dengan nilai 100%.

Pada penelitian [19], memfokuskan pada penggunaan metode *K-Means Clustering* dalam mengelompokkan menu makanan dan minuman di Dpom *Coffee.* Hasil yang didapat setelah melakukan pengolahan data penjualan di Dpom *Coffee* dengan menggunakan aplikasi *WEKA,* diperoleh hasil pengelompokkan menjadi 3 *cluster. Cluster* 1 menu dengan tingkat penjualan rendah sebanyak 8 menu. *Cluster* 2 terdapat 40 menu yang memiliki tingkat penjualan sedang. *Cluster* 3 terdapat sebanyak 7 menu dengan tingkat penjualan tinggi.

Studi literatur ini dilakukan untuk menambah informasi atau menambah landasan teori sebagai bahan untuk menyelesaikan penelitian ini. Hasil dari studi literatur yang dilakukan terhadap beberapa jurnal seperti pada Tabel 2.1, terdapat beberapa perbedaan dengan penelitian ini seperti pada objek penelitian, metode yang digunakan dalam penelitian, dan aplikasi yang digunakan. Namun, terdapat beberapa jurnal yang memliki tujuan penelitian dan metode penelian yang sama dengan penelitian ini. Penelitian ini juga dilakukan untuk membantu evaluasi serta memperbaiki kualitas layanan yang ada di destinasi wisata.

2.2 Dasar Teori

Pada bagian ini, membahas terkait dasar teori yang sesuai dengan topik pada penelitian ini. Dasar teori dapat berasal dari literatur atau penelitian sebelumnya. Beberapa dasar teori yang dijadikan acuan dalam penelitian ini adalah sebagai berikut:

2.2.1 Knowledge Discovery in Database (KDD)

KDD atau *Knowledge Discovery in Databases* adalah sebuah proses untuk mendapatkan informasi yang berguna dalam basis data. Terdapat beberapa langkah proses *KDD*, dimulai dengan memahami bidang aplikasi, kemudian membuat kumpulan data target dari data yang tersimpan dalam *database*, dan selanjutnya membersihkan dan memproses data [20].

Data mining merupakan bagian dari proses *Knowledge Discovery in Database (KDD)*. Tahapan yang ada pada proses *data mining* terbagi menjadi 6 tahapan. Berikut ini penjelasan mengenai 6 tahapan *data mining* [21]:

1) *Data cleaning* (Pembersihan data)

   Data cleaning digunakan untuk menghapuskan data yang kurang lengkap dan tidak konsisten, sehingga data dapat digunakan untuk proses ulang. Sebagai contoh pada dataset penjualan produk yang memiliki nilai harga yang kosong, nilai tersebut dapat dihapus.

2) *Data integration* (Integrasi data)

   Data integration digunakan untuk menyatukan dan mengkombinasikan beberapa sumber data yang berulang dan beberapa file yang berulang kedalam

satu sumber. Sebagai contoh menggabungkan data harga produk pada *excel* dengan jumlah produk yang ada pada *CSV*.

3) *Data selection* (Seleksi data)

   Data selection merujuk pada proses memilih atau mengambil data yang relevan untuk dianalisis dari suatu *database*. Sebagai contoh memilih hanya kolom-kolom tertentu dari tabel yang akan digunakan dalam analisis, misalnya hanya memilih kolom tanggal, produk, dan jumlah penjualan dari tabel penjualan.

4) *Data transformation* (Transformasi data)

   Data transformasi digunakan untuk mengubah format atau menggabungkan data agar dapat diproses dengan lebih tepat dan mudah. Sebagai contoh mengubah format data yang tadi menggunakan format *xlsx* menjadi format *CSV*.

5) *Data mining* (Proses *mining)*

   Data mining digunakan untuk menemukan informasi yang tidak terlihat secara langsung dari data yang berguna. Sebagai contoh menggunakan algoritma *clustering* untuk mengelompokkan data penjualan berdasarkan pola pembelian konsumen.

6) *Evaluation pattern* (Evaluasi pola)

   Evaluation pattern digunakan untuk mengidentifikasi pola yang dapat dimasukkan ke dalam *knowledge base* yang telah ditemukan. Evaluasi pola dilakukan untuk mengevaluasi hasil data mining yang sudah diperoleh. Sebagai contoh mengidentifikasi pola pembelian konsumen yang signifikan dan mengevaluasi pengaruhnya terhadap penjualan.

2.2.2 Data

Data adalah kumpulan informasi yang diperoleh melalui pengamatan, yang dapat berupa angka, simbol atau karakteristik. Data dapat memberikan pandangan tentang situasi atau masalah tertentu [22]. Data merupakan elemen awal dari informasi yang perlu diolah dan diproses lebih lanjut untuk mendapatkan informasi yang berarti dan berguna. Berdasarkan definisi ini, disimpulkan bahwa data yang akurat dapat menghasilkan informasi yang berkualitas setelah diolah atau diproses menjadi suatu angka atau bentuk lainnya[23]. Data memegang peran penting dalam penelitian. Data ini digunakan sebagai sumber analisis dan dasar untuk menarik

kesimpulan [24]. Data dibagi menjadi empat jenis, yaitu nominal, ordinal, interval, dan rasio untuk memudahkan pengelompokan dan analisis. Data nominal merupakan metode pengukuran yang paling sederhana yang hanya memiliki fungsi untuk mengidentifikasi dan membedakan. Contoh data nominal adalah tua muda, kaya miskin, dan lainnya. Data ordinal menyatakan kategori dan juga peringkat dari konstruk yang diamati. Contoh data ordinal adalah jenis kendaraan, peringkat kejuaraan, dan lainnya. Data interval adalah hasil dari Skala pengukuran di mana jarak antara dua titik sudah diketahui sebelumnya. Contoh data interval adalah abcde, 12345, dan lainnya. Data rasio juga diperoleh dari pengukuran yang jarak antara dua titik skala sudah diketahui dan memiliki titik nol absolut. Contoh data rasio adalah jumlah penduduk, jumlah tenaga kerja, dan lainya. Dalam penelitian, data dapat dikumpulkan dengan metode wawancara dan observasi[25].

2.2.3 Data Mining

Data mining adalah proses mencari pola atau model baru yang berguna, dapat dimengerti, dan optimal dalam *database* besar[26]. Berdasarkan fungsi yang dilakukan *data mining* memiliki beberapa metode, yaitu *Classification* digunakan untuk mengategorikan item ke dalam satu kelas tertentu dari beberapa kelas yang ada. Sementara itu, *clustering* digunakan untuk menganalisis data dan menemukan kelompok produk yang memiliki kesamaan tertentu. *Association* digunakan untuk mengenali kaitan antara peristiwa-peristiwa yang berlangsung pada suatu waktu. *Sequencing* digunakan untuk mengidentifikasi pola urutan kejadian yang berbeda pada periode waktu tertentu secara berulang. *Forecasting* adalah teknik yang digunakan untuk mengestimasi nilai atau kejadian di masa depan dengan memanfaatkan pola dan tren yang terlihat pada data yang besar [27]. Berikut merupakan karakteristik dari *data mining*[28]:

a) *Data mining* melibatkan identifikasi dan penemuan pola yang mungkin tidak diketahui sebelumnya dalam data yang besar dan kompleks.
b) *Data mining* umumnya dilakukan pada *dataset* yang memiliki volume besar.

c) Penggunaan *dataset* besar dapat meningkatkan kepercayaan terhadap hasil yang dihasilkan.
d) Data *mining* dapat membantu dalam pengambilan keputusan kritis, terutama dalam strategi

2.2.4 Z-Score Normalization

Normalisasi data adalah tahap penting dalam praproses data, yaitu memperkecil skala nilai-nilai pada dataset untuk mempermudah proses pengolahan. Kebutuhan untuk melakukan normalisasi data karena sering ada rentang nilai yang berbeda-beda antar atribut pada *dataset*, yang dapat mengurangi efektifitas peran atribut pada *dataset*. *Z-Score normalization* adalah salah satu teknik statistika yang dapat digunakan dalam *big data. Z-Score* juga dikenal sebagai nilai baku atau nilai standar, dan digunakan dalam *data mining* untuk menemukan data yang merupakan nilai standar. Teknik *Z-Score* melakukan transformasi data dengan menciptakan rentang nilai baru berdasarkan rentang nilai yang ada di *dataset.* Nilai baru yang dihasilkan dengan teknik *Z-Score* didasarkan pada perbedaan antara nilai rata-rata dan standar deviasi[29]. Metode Normalisasi *Z-Score* menggunakan rata-rata dan deviasi standar setiap atribut fitur untuk mengubah skala nilai data[30]. Persamaan untuk menghitung *Z-Score Normalization* dapat ditemukan pada Persamaan berikut[29]:

$$\boldsymbol{X_{baru}} = \frac{X_{lama} - \bar{X}}{\sigma} \tag{2.1}$$

Keterangan:

$\bar{\boldsymbol{X}} = rata - rata$

$\boldsymbol{\sigma} = standar\ deviasi$

Berikut ini merupakan contoh perhitungan *Z-Score Normalization* menggunakan persamaan 2.1:

$$\boldsymbol{X_{baru}} = \frac{X_{lama} - \bar{X}}{\sigma}$$

$$\boldsymbol{X_{baru}} = \frac{15 - 3}{4}$$

$$\boldsymbol{X_{baru}} = \frac{12}{4}$$
$$\boldsymbol{X_{baru}} = 3$$

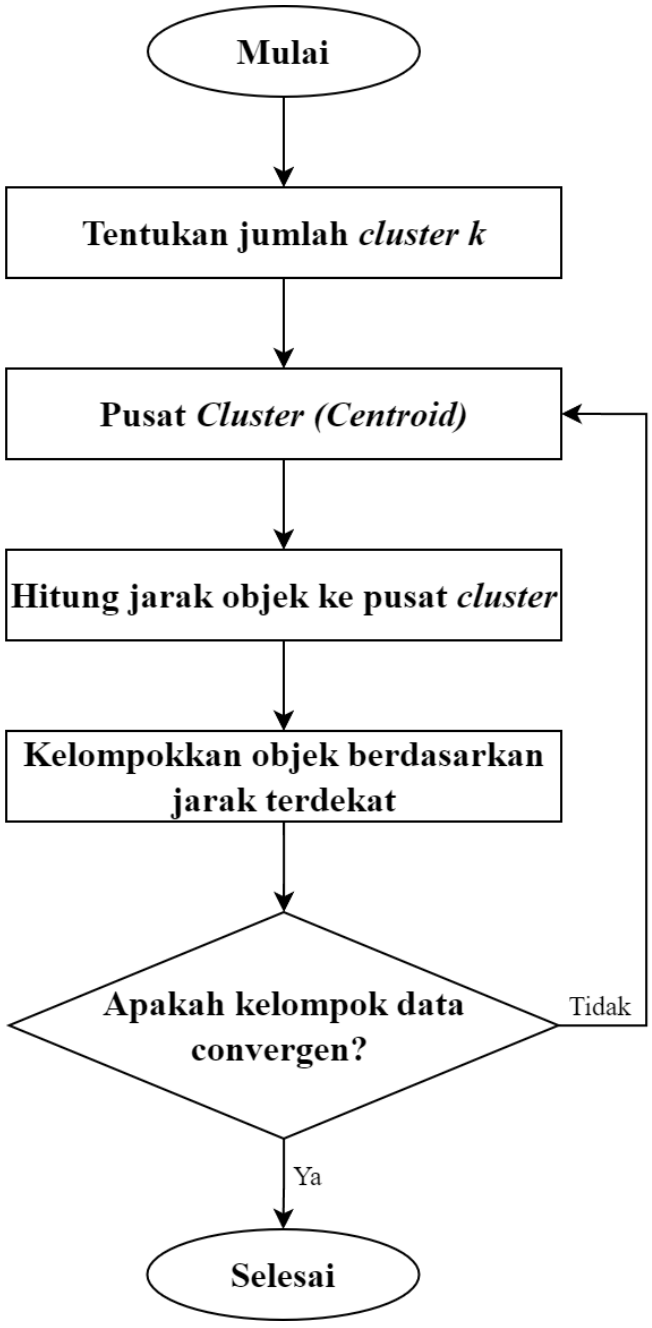
Berdasarkan perhitungan normalisasi data menggunakan persamaan 2.1, jika X_{lama} sebesar 15, rata-rata sebesar 3, dan standar deviasi sebesar 4, maka didapatkan X_{baru} sebesar 3.

2.2.5 Metode *Elbow*

Metode *Elbow* merupakan salah satu teknik yang digunakan untuk menemukan nilai k optimal pada *algoritma clustering*. Metode ini dilakukan dengan cara membandingkan nilai *SSE* yang diperoleh dari berbagai percobaan *clustering* dengan nilai k yang berbeda, kemudian hasilnya ditampilkan dalam grafik. Nilai k optimal akan ditemukan pada titik "*elbow*" atau patahan pada grafik tersebut, di mana penambahan jumlah klaster tidak lagi signifikan memperkecil nilai *SSE* [31].

2.2.6 Algoritma *K-Means Clustering*

Clustering merupakan proses pengelompokkan benda atau produk yang sama ke dalam kelompok yang berbeda, sehingga setiap kelompok memiliki arti yang bermanfaat. Algoritma *clustering* terdiri dua bagian yaitu hierarkis dan non-hierarkis. Algoritma hierarkis digunakan untuk menemukan *cluster* secara berurutan, sedangkan algoritma partisional digunakan untuk menentukan semua kelompok pada waktu tertentu. Salah satu jenis algoritma partisional adalah *Algoritma K-Means* yang memulai proses *clustering* dengan menentukan jumlah awal kelompok dan nilai *centroid* awal. Algoritma *K-Means* digunakan untuk mengelompokkan item dalam *dataset* ke dalam beberapa kelompok berdasarkan jarak terdekat di dalam kluster [32]. *Algoritma K-Means* adalah sebuah algoritma yang mengambil input parameter dan kemudian membagi himpunan objek ke dalam *cluster-cluster* dengan tingkat kemiripan anggota dalam satu *cluster* yang tinggi, sementara tingkat kemiripan antar anggota pada cluster yang berbeda sangat rendah [33].



Gambar 2. 1 Tahapan *K-Means Clustering*

Tahapan-tahapan yang dilakukan dalam pengelompokkan data menggunakan algoritma *K-Means Clustering* dijelaskan sebagai berikut[34]:

1) Melakukan penetuan nilai jumlah *cluster(k).*

2) Melakukan inisialisasi *centroid* awal *(initial centroid)* secara acak. Untuk menentukan pusat *cluster* baru dalam *algoritma K-Means*, dilakukan dengan cara menghitung rata-rata dari nilai total jarak dalam suatu *cluster*. Adapun rumus yang digunakan seperti persamaan (2.2) berikut[35]:

$$\boldsymbol{C_i} = \frac{\sum d_i}{n} \quad (2.2)$$

Keterangan:

$\boldsymbol{C_i}$ = *centroid* baru ke *i*

$\boldsymbol{d_i}$ = jumlah nilai jarak yang masuk dalam tiap *cluster*

$\boldsymbol{n}$ = jumlah data pada tiap *cluster*

3) Melakukan perhitungan jarak pada masing-masing data pada setiap *centroid* menggunakan rumus jarak *Euclidean Distance* berikut:

$$\boldsymbol{d(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ldots + (X_{ki} - X_{kj})^2} \quad (2.3)$$

Keterangan:

$\boldsymbol{d(i,j)}$ = jarak data ke *i* ke pusat *cluster j*

$\boldsymbol{X_{ki}}$ = data ke *i* pada atribut data ke *k*

$\boldsymbol{X_{kj}}$ = titik pusat ke *j* pada atribut data ke *k*

Berikut ini merupakan contoh perhitungan jarak pada data pertama ke pusat *cluster* 1 menggunakan 3 atribut/paramet dengan persamaan *Euclidean Distance:*

$$\boldsymbol{d(1,1)} = \sqrt{(92-22)^2 + (31-16)^2 + (75-13)^2}$$

$$\boldsymbol{d(1,1)} = \sqrt{(70)^2 + (15)^2 + (62)^2}$$

$$\boldsymbol{d(1,1)} = \sqrt{4900 + 225 + 3844}$$

$$\boldsymbol{d(1,1)} = \sqrt{8969}$$

$$\boldsymbol{d(1,1)} = 94{,}7048$$

Hasil yang diperoleh dari perhitungan jarak data pertama ke pusat *cluster* 1 sebesar **94,7048**.

4) Melakukan pengelompokkan dengan menghitung jarak antara setiap objek dengan *centroid* dan memilih *centroid* terdekat sebagai representasi *cluster*

5) Tahap terakhir dalam proses ini adalah melakukan uji konvergensi antara kelompok data yang baru dibentuk dengan kelompok data pada iterasi sebelumnya. Jika kedua kelompok data tersebut sudah sama, proses *clustering* akan dihentikan. Namun, jika kedua kelompok data masih berbeda, maka akan dilakukan iterasi dan menentukan pusat *cluster* baru.

2.2.7 Pariwisata

Pariwisata adalah salah satu sektor yang diunggulkan dalam meningkatkan pendapatan nasional, menyerap tenaga kerja, serta berperan sebagai penyumbang devisa negara. Dengan adanya pariwisata disetiap daerah diharapkan pemerinah daerah dapat memanfaatkan peluang tersebut untuk memaksimalkan potensi sumber daya alam [4]. Banyak orang menganggap pariwisata sebagai alternatif penting untuk pembangunan, terutama di negara-negara yang memiliki sumber daya alam yang terbatas. Pariwisata merujuk pada aktivitas perjalanan yang dilakukan oleh orang perorangan atau dalam kelompok dengan maksud untuk tujuan rekreasi, pengembangan pribadi, ataupun mempelajari keunikan dari obyek-obyek wisata [36]. Terdapat beberapa jenis pariwisata berdasarkan tujuan perjalanan, yaitu [37] :

a. Pariwisata untuk menikmati perjalanan

   Jenis pariwisata yang dilakukan oleh seseorang dengan tujuan berlibur, mencari udara segar, dan menikmati keindahan alam

b. Pariwisata untuk rekreasi

   Jenis pariwisata ini dilakukan ketika seseorang ingin memulihkan diri dari keletihan dan kelelahan yang dialami selama berada di tempat rekreasi.

c. Pariwisata untuk kebudayaan

   Pariwisata jenis ini dilakukan untuk mempelajari adat, kelembagaan, dan budaya masyarakat yang berbeda melalui kunjungan ke tempat bersejarah, pusat seni, agama, atau festival seni.

d. Pariwisata untuk olahraga

   Pariwisata olahraga juga dapat dilakukan oleh orang-orang yang ingin berlatih dan mencoba sendiri aktivitas seperti pendakian gunung, olahraga

berkuda, berburu, memancing, dan lain sebagainya. Pariwisata olahraga juga dapat dilakukan oleh orang-orang yang ingin berlatih dan mencoba sendiri aktivitas seperti pendakian gunung, olahraga berkuda, berburu, memancing, dan lain sebagainya.

e. Pariwisata untuk usaha dagang

Jenis pariwisata ini terkait dengan kepentingan bisnis atau pekerjaan, di mana seseorang melakukan perjalanan untuk tujuan kerja seperti rapat bisnis, konferensi, presentasi, ataupun kunjungan resmi.

f. Pariwisata untuk berkonvensi

Jenis pariwisata ini menarik banyak negara untuk mengembangkan pariwisata konvensi karena dapat menarik banyak peserta yang menginap dalam jangka waktu tertentu saat ada konvensi atau pertemuan diadakan di negara tersebut.

2.2.8 Waikato Environment for Knowledge Analysis (WEKA*)*

WEKA, singkatan dari *Waikato Environment for Knowledge Analysis*, adalah sebuah perangkat lunak yang terkenal dalam bidang pembelajaran mesin dan ditulis dengan menggunakan bahasa pemrograman *Java. WEKA* dikembangkan oleh *University of Waikato* di Selandia Baru. WEKA menyediakan berbagai algoritma dan visualisasi yang digunakan untuk menganalisis data dan membuat model prediksi. *WEKA* mencakup berbagai teknik pembelajaran, seperti pohon keputusan, *Support Vector Machines (SVM)*, logistik dan linier, *multi-layer perceptrons*, dan metode *nearest neighbor*[38]. *WEKA* pertama kali dikembangkan pada tahun 1994 dan telah menjadi perangkat lunak sumber terbuka paling populer untuk *data mining. WEKA* memiliki banyak kelebihan, antara lain memiliki berbagai algoritma *data mining* dan *machine learning* yang beragam, mudah digunakan, dan selalu terbarukan dengan algoritma-algoritma terbaru [39]. Penggunaan aplikasi *WEKA* pada penelitian ini sebagai alat untuk pengolahan data agar lebih cepat dan akurat.

2.2.9 Sum of Square Error (SSE)

Sum of Square Error atau disebut dengan *SSE* adalah metode untuk melakukan validasi *cluster* dengan menghitung jumlah kuadrat jarak setiap anggota *cluster* ke pusatnya. Teknik *SSE* digunakan untuk mengevaluasi jumlah *cluster* k yang dihasilkan dari pengujian dengan *K-Means*[40]. Dalam melakukan validasi data, *SSE* dapat dihitung untuk setiap nilai *cluster*, yang akan memberikan nilai validasi. Semakin banyak jumlah *cluster* k, semakin kecil nilai *SSE*, sehingga akan semakin baik dalam validasi[41]. Untuk menghitung *SSE* menggunakan persamaan [42]:

$$\boldsymbol{SSE} = \sum_{k=1}^{k} \sum_{x_1} ||X_i - C_k||^2 \qquad (2.4)$$

Keterangan:

k : Jumlah *cluster*

X_i : Data ke-i

C_k : *Centroid cluster*

Sebelum menghitung nilai *SSE*, menghitung jarak antara setiap data dengan *centroid* dalam *cluster* dan jumlahkan kuadrat jaraknya untuk setiap *cluster*. Berikut ini merupakan contoh perhitungan *SSE* menggunakan persamaan 2.4.

Cluster 1: *SSE* = $(0)^2 + (5)^2 = 25$

Cluster 2: *SSE* = $(3)^2 + (4)^2 = 25$

Cluster 3: *SSE* = $(1)^2 + (1)^2 = 2$

Setelah menghitung nilai *SSE* pada setiap *cluster*, kemudian menjumlahkan SSE dari semua *cluster* untuk mendapatkan *SSE* total:

SSE Total = *SSE* (Cluster 1) + *SSE* (Cluster 2) + *SSE* (Cluster 3)

= 25 + 25 + 2

= 52

Dari hasil perhitungan nilai *SSE* pada tiap *cluster* diatas, maka didapatkan total *SSE* sebesar 52.