BAB II

LANDASAN TEORI

2.1 Tinjauan penelitian sebelumnya

Tabel 2. 1 Tabel Tinjauan Penelitian Sebelumnya

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
1.	*The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19* data[7].	Pengelompokan provinsi berdasarkan kasus COVID-19 di Indonesia merupakan upaya untuk menentukan kedekatan atau kesamaan suatu provinsi[7].	Penelitian ini membahas tentang pengaplikasian K-means untuk meng*cluster* tingkat Covid-19[7].	Pada bagian diagram tidak dijelaskan mengenai *cluster* secara detail [7].	Penelitian ini melakukan *clustering* yang bertujuan untuk mengetahui tingkat resiko covid-19 berdasarkan data pasien Covid-19 di Indonesia[7].	Hasil penelitian ini terdapat 3 *cluster* pada Covid-19 di Indonesia[7].

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
2.	*A Two-Stage Time Series Clustering Framework for Explaining the Varying Patterns of COVID-19 Deaths across the U.S. (Preprint)*[9].	Penelitian ini membahas bagaimana *cluster* didistribusikan secara geografis, dan faktor apa yang memengaruhi kemungkinan keanggotaan *cluster* dari hasil menganalsis kematian COVID-19 terkonfirmasi tingkat kabupaten dari Minggu, 1	Penelitian ini menggunakan kerangka kerja analitik data 2 tahap yang dapat menjelaskan berbagai tingkat agregasi temporal untuk hasil pandemi dan predictor. Pengelompokan deret waktu untuk mengidentifikasi kluster. Regresi logistik multinomial digunakan untuk	Penilitian ini kurang menjelaskan detail dari data yang diambil[9].	Penelitian ini di lakukan untuk menentukan berapa banyak *cluster* berbeda dari deret waktu yang ada untuk kematian COVID-19 di 3108 kabupaten yang berdekatan di Amerika Serikat[9].	Hasil penelitian ini memberikan bukti bahwa pola kematian COVID-19 tingkat kabupaten berbeda dan dapat dijelaskan[9].

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
		Maret 2020, hingga Sabtu, 27 Februari 2021[9].	menjelaskan hubungan antara prediktor tingkat komunitas dan penugasan klaster[9].			
3.	Implementasi Metode K-Means Untuk Pengelompokan Kasus Covid-19 Tingkat Provinsi Di Indonesia[10].	Metode yang digunakan dalam penelitian ini ialah k-means *clustering* dan *silhouette index* (SI)[10].	Metode k-means *clustering* dipakai untuk klasterisasi data COVID-19 dan metode SI dipakai untuk uji keakurasian klasterisasi data COVID-19[10].	Kurang dijelaskan metode SI yang digunakan[10].	Penelitian ini bertujuan untuk mengidentifikasi dan mengetahui Tujuan dari penelitian ini untuk mengetahui tingkat kasus penyebaran COVID-19 di Indonesia[10].	Berdasarkan hasil dari pengujian algoritma k-means *clustering* dinyatakan bahwa uji validasi menunjukkan nilai SI dengan tingkat akurasi sebesar 0,857 atau 85,7%[10].

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
6.	*A dynamic K-means clustering for Data Mining* oleh [11].	Penelitian ini membandingkan metode yang di usulkan oleh penulis dengan algoritma K-Means yang ada[11].	Metode yang diusulkan secara dinamis membentuk *cluster* untuk kumpulan data yang diberikan[11].	Penelitian ini tidak menjelaskan metode yang diusulkan tersebut digunakan untuk tipe kasus seperti yang bagaimana[11].	Penelitian ini mengusulkan algoritma K-Means baru untuk menghilangkan kesulitan dari K-Means yang ada[11].	Hasilnya menunjukkan bahwa metode yang diusulkan mengungguli metode yang ada untuk kumpulan data iris yang terkenal[11].

2.2 Dasar Teori

2.2.1 Data Mining

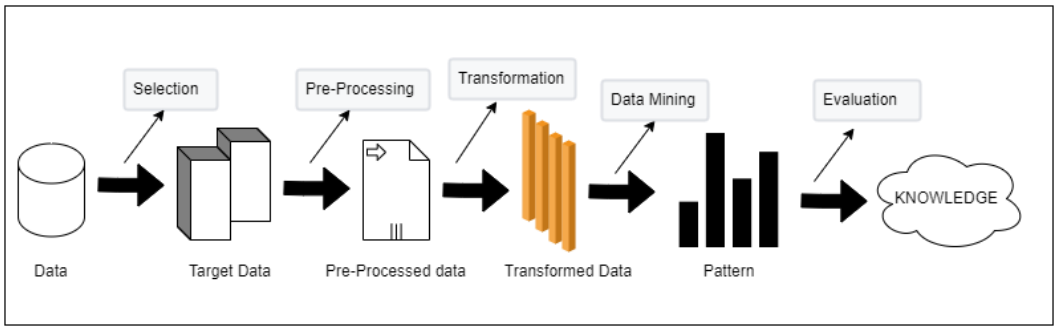
Data Mining merupakan suatu proses penggalian data dengan cara mengekstraksi dan mengenali data pada sebuah *database*. *Data mining* merupakan proses logis untuk menemukan informasi yang berguna. Setelah ditemukan informasi dan pola dapat di gunakan untuk alat pendukung dalam pengambilan keputusan dalam mengembangkan bisnis. Alat *data mining* dapat memberikan jawaban untuk berbagai pertanyaan yang terkait dengan bisnis dan sulit diselesaikan. *Data mining* juga dapat digunakan untuk meramalkan tren masa depan yang memungkinan pebisnis membuat keputusan yang eefektif, proaktif, dan dinamis. Data - data yang diolah dengan menggunakan teknik *data mining* juga mampu menghasilkan pengetahuan yang sesuai dengan harapan. Secara umum *data mining* dapat diklasifikasikan menjadi dua kategori, yaitu *supervised data mining technologies* dan *unsupervised data mining technologies. Supervised data mining technologies* bertujuan untuk mempelajari hubungan yang lebih rumit antara banyak variabel, biasanya digunakan untuk melakukan prediksi atau klasifikasi sedangkan *unsupervised data mining technologies* merupakan teknologi data mining yang cocok dalam menemukan struktur intrinsik, hubungan dan keterkaitan sebuah data. Metode ini dapat menemukan operasi atau pola tersembunyi dalam membangun sistem data operasi[12]. *Data mining* digunakan untuk mencari pengetahuan atau data dalam jumlah besar dalam sebuah basis data. *Data mining* memiliki fungsi umum yang diterapkan[13]:

1. *Assosiation*: Proses menemukan aturan asosiasi antara suatu kombinasi item dalam suatu waktu
2. *Sequence*: proses penemuan aturan asosiasi antara sebuah kombinasi item dalam suatu waktu dan diterapkan lebih dari satu periode
3. *Clustering*: Proses pengelompokkan sejumlah data ke dalam suatu kelompok data sehingga dalam setiap kelompok terisi data yang mirip atau similar

4. *Classification*: Membedakan konsep atau kelas data untuk memperkirakan kelas dari suaut objek yang belum diketahui
5. *Regression*: Proses pemetaan data
6. *Forecaston*: Pengestimasian nilai prediksi berdasarkan pola
7. *Solution*: Proses penemuan akar masalah dan penyelesaian masalah dari suatu persoalan yang ada.

2.2.2 KDD (*Knowledge Discovery in Data Process*)

KDD (*Knowledge Discovery in Data Process)* adalah suatu rangkaian yang berhubungan dengan visualisasi dari pola-pola pada sejumlah data. Proses KDD (*Knowledge Discovery in Data Process)* dapat melibatkan iterasi yang signifikan dan memungkinkan berisi loop antara dua langkah. Proses KDD adalah ditunjukkan pada Gambar 2.1 sebagai berikut[14]:



Gambar 2. 1 Gambar Knowledge Discovery in Data Process

1. *Data Selection*: Membuat atau memilih dataset dasar dari sebuah variabel atau contoh data pada objek yang akan dilakukan
2. *Pre-processing*: Pemrosesan atau pembersihan data seperti menghapus noise dan memilih strategi untuk mengatasi jika ada data *fields* yang hilang.
3. *Transformation*: Data diubah menjadi bentuk yang sesuai untuk dilakukan penggalian.
4. *Data Mining*: Pemilihan algoritma untuk pencarian suatu pattern, pengklasifikasian aturan, regresi, *clustering*, *sequence*, *modelling*, pengekstrasian pola dari data yang sudah ada
5. *Interpretation/Evaluation* : Proses intrepretasi pola

2.2.3 *Cluster*ing

Clustering adalah salah suatu metode dalam *data mining* yang digunakan untuk menganalisis atau mengelompokkan data yang besar ke dalam suatu *cluster*. *Clustering* bertujuan untuk mengelompokan data, pola ataupun sebuah dokumen. Konsep pengelompokan pada *data mining* adalah pengelompokan data yang memiliki karakteristik sama ke dalam suatu *cluster* yang sama dan data dengan karakteristik berbeda ke *cluster* lain. *Cluster* adalah sekelom-pok atau sekumpulan objek-objek data yang similar satu sama lain dalam *cluster* yang sama dan disimilar terhadap objek-objek yang berbeda *cluster*. Objek akan dikelomokkan ke dalam satu atau lebih *cluster* sehingga objek-objek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan yang lainnya. Menggunakan *clustering* ini, dapat dilakukan pengklasifikasian daerah yang padat, menemukan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut data. Dalam data mining, usaha difokuskan pada metode-metode penemuan untuk *cluster* pada basis data berukuran besar secara efektif dan efisien. Beberapa kebutuhan *cluster*ing dalam data mining meliputi skalabilitas, kemampuan untuk menangani tipe atribut yang berbeda, mampu menanganidimensionalitas yang tinggi, mena-ngani data yang mempunyai *noise,* dan dapat diterjemakan dengan mudah[15]. Sebuah objek di *cluster* berdasarkan *principle* dari *maximizing & minimizing interclass similarity*. *Clustering* dilakukan sehingga suatu objek diantara *cluster* memilki nilai kemiripan yang lebih tinggi jika dibandingkan dengan yang lain. Setiap *cluster* dapat di katakan sebagai *class* dari sebuah objek, dimana aturan dapat diturunkan[16]. *Clustering* akan menghasilkan tingkat kesamaan yang tinggi terhadap dua buah objek dalam suatu class dan tingkat kesamaan yang rendah antar class. *Clustering* memiliki empat tipe data : Variabel berskala interval; Variabel biner; Variabel nominal, ordinal dan rasio[17].

2.2.4 Analisis *Cluster*ing

Analisis *cluster*ing bertujuan untuk mengelompokkan objek-objek pada sebuah data sesuai dengan kemiripan karakteristik setiap objek tersebut. Analisis *Cluster*ing masuk ke dalam kategori analisis statistic multivariate metode

independen yang menjadkan tujuan analisis *cluster*ing tidak digunakan untuk menghubungkan atau menjadi pembeda dengan sample ataupun variabel lain. Analisis *cluster* memliki 2 metode yaitu Hirarki dan Non-Hirarki dimana metode hirarki merupakan jumlah kelompok yang akan di peroleh namun belum diketahui sedangkan non hirari dapat di artikan bahwa ada nilai k terlebih dahulu[18]. Penjelasan 2 metode pada analisis *cluster*ing sebagai berikut[19] :

1. Metode Hirarki: Metode ini diawali dengan melakukan pengelompokkan dua atau lebih objek yang memiliki kemiripan terdekat. Proses selanjutnya dilanjutkan ke objek lain yang memiliki kemiripan atau kedekatan kedua. Tahap itu dilakukan sampai kelompok atau *cluster* terbentuk menjadi suatu hirarki atau tingkatan yang jelas antar objek.
2. Metode Non-Hirarki: Pada metode ini penghitungan dilakukan dengan cara menghitung jarak menggunakan *euclidian* guna menentukan nilai kemiripan antar setiap objek. Hasil *cluster* pertama akan menjadi objek observasi pertama. Hasil objek kedua adalah observasi lengkap berikutnya yang dipisahkan dari hasil pertama menggunakan jarak minimum khusus.

Analisis *Cluster* memiliki beberapa tahapan atau proses sebagai berikut[20]:

1. Menentukan ukuran kemiripan dari kedua objek

   Pada tahap ini dilakukan pengukuran seberapa jauh jarak kemiripan atau kesamaan disetiap objek-objek yang dipilih
2. Melakukan Proses Standarisasi data

   Proses ini dilakukan hanya jika memang di perlukan. Standarisasi data dilakukan dengan syarat nilai dari masing-masing setiap variabel memiliki perbedaan skala. Pada tahap ini memiliki metode yang biasa digunakan untuk melakukan standarisasi data yaitu perhitungan dengan metode Min-Max seperti sebagai berikut:

$$X_1 = \frac{\mathrm{X} - X_{min}}{X_{max} - X_{min}} \quad (2.\ 1)$$

Keterangan:

X1: Data hasil dari dilakukannya normalisasi

X: Data asli sebelum di normalisasi

Xmax: Nilai maksimal dari data

Xmin: Nilai Min dari data

2.2.5 K-Means

Algoritma *K*-Means merupakan sebuah algoritma berdasarkan pembagian dimana algoritma ini merupakan sebuah algoritma *clustering* berulang-ulang yang diawali dengan menetapkan nilai *cluster* secara random dan nilai tersebut menjadi pusat dari *cluster*[21]. *K-Means* merupakan saalah satu metode data *cluster*ing non hierarki 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 *cluster* yang sama. Pengelompokkan algoritma *k-means* bertujuan untuk meminimalkan indeks kinerja *cluster,* kesalahan kuadrat dan kriteria kesalahan. *K-Means* adalah metode *clustering* berbasis jarak yang membagi data ke dalam sejumlah *cluster* dan algoritma ini hanya bekerja pada atribut *numeric*. Algoritma *k-means* termasuk *partitioning clustering* yang memisahkan data ke *k* daerah bagian yang terpisah. Algoritma *k-means* sangat terkenal karena kemudahan dan kemampuannya untuk meng-*cluster* data yang besar sangat cepat. Dalam algoritma *k-means*, setiap data harus termasuk ke *cluster* tertentu dan bisa dimungkinkan bagi setiap data yang termasuk *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke *cluster* lainnya[11]. Hasil optimalisasi pada algoritma ini, mencoba dengan mencari nilai K untuk memenuhi kriteria tertenu. Diawali dengan memilih beberapa titik untuk mewakili titik focus *cluster* awal. Selanjutnya mengumpulkan titik sampel yang tersisa terhadap titik fokus awal sesuai dengan kriteria jarak minimum maka, akan mendapatkan nilai klasisfikasi awal. Proses tersebut dilakukan secara berulang

sampai Nilai dari *center* tidak berubah yang berarti telah ditemukan jarak yang paling dekat dari setiap data dengan *means*[22]. Tahapan dan rumus metode *K-means* sebagai berikut[23] :

$$d_{(x,y)} = \sqrt{\sum_{i=0}^{n} (xi - yi)^2} \qquad (2.\ 2)$$

Keterangan:

i(x, y) = Jarak data ke x ke pusat *cluster* y

xi = data x pada observasi ke-i

yi = titik pusat ke y

observasi ke-i

n = banyaknya observasi

Metode *K-Means* memiliki beberapa tahapan dan proses seperti sebagai berikut : *Tahapan* pada Metode *K-Means:*

1. Menentukan jumlah *Cluster* (K)
2. Menentukan nilai *centroid* atau nilai tengah secara acak
3. Menghitung nilai centroid atau nilai tengah pada masing
4. Memberikan label pada setiap data sesuai dengan *centroid* terdekat
5. Memilih nilai *centroid* yang baru, sesuai dengan *cluster* baru yang telah terbaru
6. Data di label ulang sesuai dengan jarak paling terdekat dengan titik *centroid* yang paling baru

2.2.6. Davies Bouldin(DBI)

Davies-Bouldin Index (DBI) adalah salah satu pengukur evaluasi algoritma *clustering*. DBI digunakan untuk mengevaluasi pengelompokan sejumlah *cluster* pada *K-Means clustering* dengan sejumlah *cluster* tertentu yang diberikan[24]. *Davies Bouldin Index* (DBI) dihitung sebagai rata-rata kesamaan pada setiap *cluster* dengan *cluster* yang paling mirip dengan *cluster* tersebut. Pada *Davies-Bouldin Index* (DBI) jika semakin rendah rata-rata kesamaan, semakin baik *cluster* yang

dipisahkan maka akan semakin baik hasil *clustering* yang dilakukan. *Davies Bouldin Index* memiliki rumus sebagai berikut [25]:

$$DBI = \frac{1}{k}.\sum_{i=1}^{k} R_i$$

$$\text{Dengan } R_i = \max R_{ij}$$

$$\text{dan } R_{ij} \frac{var(C_i)+var(c_j)}{\|c_i - c_j\|} \qquad (2.\ 3)$$

dimana : C_i : *cluster i* dan c_i adalah nilai *centroid* dari *cluster* i.

2.2.6 Covid-19

Covid-19 atau *Corona Virus* adalah suatu penyakit yang menyerang sistem pernafasan. Penyakit ini merupakan keluarga besar dengan penyakin *MERS* (*Middle East Respiratory Syndrome)* dan *SARS* (*Severe Acute Respiratory Syndrome)*. *Coronavirus* adalah virus RNA dengan ukuran partikel 120-160 mm. Virus ini utamanya menginfeksi hewan, termasuk di antaranya adalah kelelawar dan unta. Sebelum terjadinya wabah COVID-19, ada 6 jenis *coronavirus* yang dapat menginfeksi manusia, yaitu *alphacoronavirus* 229E, *alphacoronavirus* NL63, *betacoronavirus* OC43, *betacoronavirus* HKU1, *Severe Acute Respiratory Illness Coronavirus* (*SARS-CoV*), dan *Middle East Respiratory Syndrome Coronavirus* (*MERS-CoV*).14 *Coronavirus* yang menjadi etiologi COVID-19 termasuk dalam *genus betacoronavirus*. Hasil analisis filogenetik menunjukkan bahwa virus ini masuk dalam subgenus yang sama dengan coronavirus yang menyebabkan wabah *Severe Acute Respiratory Illness* (SARS) pada 2002-2004 silam, yaitu *Sarbecovirus*. 15 Atas dasar ini, *International Committee on Taxonomy of Viruses* mengajukan nama *SARS-CoV*-2.16 Virus ini dilaporkan pertama kali di kota Wuhan, China dengan menyerang organ atau sistem pernafasan. Penyebaran virus ini diketahui sangat cepat sekali ditandai dengan penambahan jumlah kasus yang sangat banyak hanya dalam kurun waktu kurang dari satu bulan. Virus ini sudah menyebar ke berbagai belahan dunia seperti Jepang, Korea Selatan, Thailand[26].

WHO (*World Health Organization)* menyatakan Covid-19 sebagai *pandemic global* pada Maret 2020 dikarenakan jumlah yang terinfeksi sudah mencapai 200.000 pasien dengan jumlah kematian 8000 di 160 negara. Hingga tahun 2022, Indonesia masih menempati posisi ke 4 se Asia dengan jumlah masyarakat terpapar covid sebanyak 42 juta. Covid-19 dapat menyebabkan kematian yang dimana jumlah kematian yang terjadi sangat banyak dari berbagai provinsi di Indonesia[27].

Provinsi di Indonesia yang terinfeksi Covid-19, salah satunya adalah Kalimantan utara, dimana Kalimantan Utara adalah provinsi dengan luas wilayah ±75.467,70 km2 dan penduduknya berjumlah ±716.407 jiwa. Kalimantan juga memiliki Kabupaten/Kota antara lain Tarakan, Malinau, Kepulauan Tana Tidung, Nunukan dan Tanjung Selor[28].