

BAB III

LANDASAN TEORI

3.1 COVID-19

Pada desember 2019, di kota wuhan, Tiongkok, terjadi kejadian luar biasa kasus radang paru-paru (Pneumonia) yang disebabkan oleh virus dari golongan virus corona, namun virus tersebut belum pernah ditemukan sebelumnya. Dan pada 11 februari 2020 secara resmi WHO mengumumkan penamaan baru virus tersebut dengan nama *Severa Acute Syndrome Coronavirus-2* dan penyakit yang ditimbulkan adalah *Coronavirus Disease 2019* atau yang lebih dikenal dengan COVID-19. COVID-19 adalah penyakit yang menyerang saluran pernapasan manusia yang ditularkan lewat droplet. Masa inkubasi virus corona 1-14 hari. Selain berbahaya virus ini juga dikenal dengan penularannya yang sangat cepat [31].

3.2 Machine Learning

Machine learning adalah teknik untuk melakukan inferensi terhadap data dengan pendekatan matematis. Inti dari machine learning adalah untuk membuat model (matematis) yang merefleksikan pola-pola data. Machine learning juga banyak memanfaatkan statistika dan aljabar linear. Penerapan machine learning pada suatu domain membutuhkan pengetahuan/keahlian pada domain tersebut. Teknik-teknik pada machine learning sangatlah berguna untuk berbagai macam permasalahan. Tetapi perlu diingat bahwa teknik-teknik yang ada belum mampu memodelkan proses berpikir manusia dengan benar. Proses berpikir manusia

sangatlah kompleks, model-model pada machine learning harus diberikan banyak sampel untuk dipelajari agar menghasilkan hasil dan akurasi yang baik [32].

3.3 Clustering

Clustering adalah pengelompokan data dengan sifat yang mirip. Data untuk *clustering* tidak memiliki label (kelas). Secara umum, algoritma *clustering* dapat dikategorikan menjadi dua macam berdasarkan hasil yang diinginkan, yaitu *partitional* dan *hierarchical*. *Partitional* yaitu menentukan partisi sebanyak K atau jumlah dari cluster yang diinginkan, sedangkan *hierarchical* yaitu mengelompokkan data berdasarkan struktur taksonomi. Contoh algoritma *partitional* adalah *K-Means*, sedangkan algoritma *hierarchical* adalah *agglomerative clustering*. Selain itu ada juga algoritma berbasis kepadatan (*density-based*) seperti *DBSCAN* [33].

3.4 K-Means

K-means merupakan salah satu metode cluster analysis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokkan kedalam *cluster* yang lain. Metode K-Means *Clustering* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain [34].

K-Means juga merupakan bagian dari *unsupervised learning* (tanpa diawasi). Dalam melakukan pembelajaran menggunakan k-means langkah-langkahnya sebagai berikut :

1. Tentukan jumlah kelompok yang kita inginkan.
2. Melakukan inisiasi centroid untuk setiap kelompok. centroid adalah data yang merepresentasikan suatu kelompok.
3. Menghitung jarak suatu data terhadap centroid, kemudian memasukan data tersebut ke kelompok yang centroidnya memiliki sifat terdekat dengan dirinya.
4. Memilih kembali centroid untuk masing-masing kelompok, yaitu dari anggota kelompok tersebut (untuk memilih centroid yang baru)
5. Iterasi dilakukan sampai tidak ada perubahan anggota untuk semua kelompok.

Berikut merupakan rumus-rumus yang digunakan dalam algoritma k-means sebagai berikut:

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} ,$$

Rumus di atas untuk menentukan nilai centroid ketika sudah melakukan iterasi. Dimana V_{ij} adalah centroid/ rata-rata cluster ke-i untuk variabel ke-j, kemudian N_i adalah jumlah data yang menjadi anggota cluster ke-I, i dan k adalah indeks dari cluster, j adalah indeks dari variabel. Dan X_{kj} adalah nilai data ke-k yang ada di dalam cluster tersebut untuk variabel ke-j.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2},$$

Rumus di atas digunakan untuk menghitung jarak antara titik centroid dengan titik tiap objek. Untuk menghitung jarak tersebut rumus yang digunakan adalah Euclidean Distance. Dimana D_e adalah Euclidean Distance, kemudian I adalah banyaknya objek, (x,y) merupakan koordinat objek, dan (s,t) merupakan koordinat centroid.

3.5 DBSCAN

DBSCAN adalah algoritma *density-based spatial clustering of applications with noise* yang merupakan salah satu metode *clustering* berbasis kepadatan (*density-based*) dimana pengamatan dilakukan dengan cara mengelompokkan data yang relative berdekatan. DBSCAN sering digunakan untuk data-data yang mengandung noise, hal tersebut disebabkan karena DBSCAN tidak memasukan data yang dianggap noise kedalam suatu cluster atau kelompok tertentu. Algoritma ini tidak memerlukan penentuan cluster di awal tetapi hanya memerlukan 2 parameter input yaitu epsilon (ϵ) dan minimum points ($minpts$) sebelum melakukan *clustering*. Epsilon diketahui sebagai jarak maksimal antar 2 data pada suatu kelompok yang diizinkan, sedangkan minimum points ($minpts$) adalah banyaknya data minimal pada jarak epsilon hingga terbentuk suatu cluster/kelompok [35].

3.6 Biostatistika dan Epidemiologi

Biostatistika merupakan metode statistik yang diterapkan pada ilmu-ilmu terkait kesehatan, seperti kedokteran dan kesehatan masyarakat, Metode Statistik kesehatan dikenal dengan istilah “biostatistik”. Biostatistik berguna untuk membantu memahami tentang karakteristik populasi, dan hubungan/ pengaruh variabel pada suatu populasi. Sedangkan epidemiologi adalah disiplin ilmu yang merupakan bagian dari pengetahuan ilmu kesehatan masyarakat yang mempelajari tentang sifat, penyebab, pengendalian, dan faktor-faktor yang mempengaruhi frekuensi, distribusi, dan kematian yang disebabkan oleh suatu penyakit pada suatu populasi [36].

Salah satu Ukuran pada biostatistika dan epidemiologi dalam melihat tingkat keparahan, resiko, dan hubungan suatu penyakit pada populasi adalah mortality, dalam ukuran mortality terdapat banyak pengukuran diantaranya yang akan digunakan dalam penelitian ini seperti [37]:

$$\text{Case fatality rate /CFR} = \frac{\text{Jumlah kematian suatu penyakit}}{\text{jumlah yang terkonfirmasi penyakit tersebut}} \times 100$$

$$\text{Mortality Rate /MR} = \frac{\text{Jumlah kematian yang disebabkan suatu penyakit}}{\text{jumlah populasi}} \times 100$$

$$\text{Recovery Rate/RR} = \frac{\text{Jumlah penderita yang sembuh}}{\text{jumlah kasus}} \times 100$$

Ukuran-ukuran di atas akan digunakan sebagai variabel pada penelitian ini ditambah dengan 1 variabel lain yaitu density (angka kepadatan penduduk/km²).

3.7 Davies-Bouldin Index

Davies-Bouldin Index merupakan ukuran atau metode evaluasi dalam melihat kualitas dari setiap cluster, dimana seberapa baik setiap objek dikelompokkan kedalam suatu cluster. Metode evaluasi ini juga merupakan gabungan dari 2 metode yaitu *cohesion* dan *separation*. Dimana *cohesion* didefinisikan sebagai kedekatan setiap data terhadap centroid dari cluster atau kelompok data itu berada, sedangkan *separation* adalah jarak antar *centroid* dari clusternya. Metode tersebut digunakan sebagai evaluasi performa dan pengukuran akurasi dalam membandingkan metode *K-Means* dan *DBSCAN* [38]. Langkah-langkahnya sebagai berikut :

1. Mencari *Sum of square within cluster* (SSW) dengan tujuan mengetahui koheisi dari sebuah cluster ke-i yang dirumuskan sebagai berikut:

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i)$$

2. Mencari *Sum of square between cluster* (SSB) dengan tujuan mengetahui separasi antar cluster yang dirumuskan sebagai berikut:

$$SSB_{i,j} = d(c_i, c_j)$$

3. Mencari nilai nilai *Rasio* untuk mengetahui perbandingan cluster ke-i dan cluster ke-j yang dirumuskan sebagai berikut :

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

4. Dari nilai *Rasio* yang didapatkan akan digunakan untuk mencari nilai dari

Davies-Bouldin Index (DBI). Yang dirumuskan sebagai berikut :

$$DBI = \frac{1}{K} \sum_{i=1}^K \max_{i \neq j} (R_{i,j})$$

Semakin kecil nilai dari *Davies-Bouldin Index (DBI)*, maka semakin baik pengelompokan metode tersebut.

