# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Fenomena *Covid-19* telah menggemparkan dunia, Pada akhir tahun 2019 dunia di kejutkan dengan penemuan virus baru yang pada awalnya menyerang pedagang pasar hasil ikan laut di Wuhan, China. Dalam waktu singkat, virus tersebut mulai menyebar ke pedagang lainnya, masyarakat di sekitar pasar dan hungga saat ini virus tersebut terus menerus menyebar luas ke berbagai Negara di dunia dan telah menginfeksi 213 negara di dunia, tidak terkecuali Indonesia, (Anies, 2020) menyebutkan, Indonesia adalah salah satu negara dimana masyarakatnya terdampak dari virus tersebut. Data dari satgas penanganan Covid-19. Indonesia menyatakan bahwa per tanggal 31 juli 2020 berada pada posisi 24 dengan jumlah kasus terbesar di dunia. Dengan adanya pandemi ini, berbagai penanganan pemerintah dilakukan seperti pemberlakuan *System Lockdown*, PPKM (pemberlakuan pembatasan masyarakat) dan berbagai macam cara di lakukan agar memutus rantai penyebaran virus Covid-19 tersebut.

Pada penelitian ini kali ini dilakukan klasterisasi penyebaran virus Covid-19 di DKI Jakarta, kota tersebut dipilih berdasarkan angka kasus tertinggi di Indonesia. Alasan dilakukannya klasterisasi ini berkaitan dengan mengelompokkan kasus persebaran covid di daerah-daerah DKI Jakarta dimana nantinya akan dilakukan untuk menentukan penanganan Covid-19 . Dengan menerapkan metode data mining. Pengelompokan dilakukan berdasarkan parameter jumlah dirawat, sembuh, meninggal dan isolasi mandiri. Metode K-Means dan metode *Cosine Similarity*, ini menghasilkan *prototipe* pengelompokan data persebaran pasien terinfeksi Covid-19. Pengklasteran dilakukan berdasarkan penyebaran kasus terbanyak di provinsi DKI Jakarta.

*Clustering* merupakan salah satu metode dalam data mining yaitu teknik pengelompokkan data, pengamatan atau memperhatiakn dan membentuk kelas obyek yang memiliki kemiripan. Salah satu metode *clustering* yang sangat terkenal adalah algoritma *k-means*, karena *k-means* memiliki algoritma yang

sederhana dan efisien. Sehingga *k-means* mudah untuk dipelajari (Gustientiedina, Adiya, & Desnelita, 2019). *k-means* merupakan metode yang cukup tangguh untuk digunakan di berbagai jenis data (Solichin & Khairunnisa, 2020). Oleh karena itu pada penelitian ini *k-means* dipilih untuk proses klasterisasi pada data persebaran COVID-19 di wilayah Jakarta. Metode K-Means digunakan, karena Mudah dilakukan saat penggunaan dan saat dijalankan. Waktu yang di gunakan dalam melakukan pembelajaran relative lebih cepat. Sangat fleksibel, adaptasi yang mudah untuk dilakukan.

Penggunanya sangat umum, menggunakan prinsip sederhana dapat di jelaskan dalam non-statistik. Namun di samping kemudahan, terdapat kekurangan, yaitu sebelum algoritma dijalankan, titik K di insialisasikan secara random sehingga pengelompokan data yang di dapatkan menjadi tidak optimal. Apabila terjebak dalam kasus yang di sebut dengan *curse of dimensionality*. Hal ini pun akan terjadi apablila salah satu data untuk melakukan pelatihan memiliki dimensi yang banyak, sebagai contoh; jika ada data sebuah pelatihan yang terdiri dari 2 buah atribut saja maka dimensinya ada 2 dimensi pula, namun akan berbeda jika ada 20 atribut maka akan ada 20 dimensi yang di miliki. Salah satu dari cara kerja algoritma cluster ini ialah untuk mencari jarak terdekat antara k titik dangan titik lainnya.

Apabila ingin mencari jarak antar titik dari 2 dimensi hal itu masih mudah di lakukan, namun dengan 20 buah dimensi hal tersebut akan menjadi sulit untuk di lakukannya pencarian jarak. Apabila terdapat beberapa buah titik sampel data yang ada, maka hal yang mudah untuk melakukan sebuah penghitungan dan juga mencari jarak titik terdekat dengan k titik yang telah di lakukan inisialisasi secara acak. Namun jika ada banyak titik data, misalkan satu juta data, maka perhitungan dan pencarian titik terdekat akan sangat membutuhkan waktu yang lama. Proses tersebut dapat dipercepat pengerjaanya namun, dibutuhkan sebuah struktur data yang rumit seperti kD-tree atau hashing untuk melakukan proses itu. Adanya penggunaan k buah random, tidak adanya jaminan untuk menemukan kumpulan cluster yang optimal. Maka dari itu menutupi kelemahan dari metode K-Means di sempurnakan dengan tambahan metode *cosine similiarty*.

*Cosine Similarity* merupakan metode yang digunakan untuk mengukur kemiripan. Ada beberapa jenis *similarity measure* seperti *Dice Coeficient, Jaccard Coeficient, Cosine Similarity, Euclidean Distance* dan lain lain (S. Christina 2014). *Eucledian Distance* dianggap sebagai *distance matrix* yang mengadopsi prinsip *Phytagoras*. Hal ini dikarenakan pola perhitungannya yang menggunakan aturan pangkat dan akar kuadrat. Eucledian akan memberikan hasil jarak yang relatif kecil karena menggunakan aturan akar kuadrat.

Beberapa penelitian tentang *Cosine Similarity* telah dilakukan diantaranya Loura Yasni, Imam Much Ibnu Subroto Sam Farisa Chaerul Haviana dengan judul Implementasi *cosine similarity matching* dalam Penentuan dosen pembimbing tugas akhir, Namun untuk penelitian yang hampir sama dengan penelitian kali ini masih belum ada. Di gunakannya *Metode cosine similarity* di karenakan metode ini digunakan untuk menghitung *similarity* (tingkat kesamaan) antara dua buah objek. *Metode cosine similarity* ini mengitung *similarity* antara dua buah objek (misalkan D1 dan D2) yang di nyatakan dalam dua buah vector dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran. *Cosine similarity* bisa digunakan sebagai metode normalisasi dokumen yang panjang selama perbandingan masih berjalan. Dalam pengambilan informasi, cosine similarity dari 2 buah dokumen akan berkisar dari 0 hingga 1, karena *term frequency* (menggunakan pembobotan tf-idf) tidak boleh negatif. Sudut antara dua *vektor term frequency* tidak boleh lebih besar dari 90.

Untuk tahapan akhir yaitu metode validasi untuk menguji sebuah cluster yang paling baik. Pada penelitian ini menggunakan metode *Davies-Bouldin Index* diperkenalkan oleh David L. *Davies and Donald W. Bouldin* pada tahun 1979. *Davies-Bouldin Index* (DBI). *Davies-Bouldin Index* (DBI) merupakan metode validasi untuk menguji sebuah cluster yang paling baik. Proses perhitungan *Davies-Bouldin Index* (DBI) (Al-Anazi dkk, 2016). Davies Bouldin Index (DBI) adalah metrik untuk mengevaluasi hasil algoritma clustering Untuk penelitiaan mengenai DBI telah dilakukan yaitu oleh Irhamni, Firli Damayanti, Fitri K, Bain Khusnul A, Mifftachul dengan judul optimalisasi pengelompokan kecamatan berdasarkan indikator pendidikan menggunakan metode *clusterin*g dan *davies*

*bouldin index*. Dapat dilhat bahwa penilitian kali ini membahas tentang covid-19 sedangkan penelitian sebelumnya mengenai pengelompokan kecamatan berdasarkan indikator pendidikan.

Sebelumnya telah dilakukan penelitian berkaitan dengan judul peneliti, diantaranya Achmad Solichin, A, Khairunnisa, K. (2020). Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means. Dan penelitian berikutnya yaitu oleh, Cahyo, M. U, Anggraini, L. B, , Maria Andini c , Hesti Retnosari d, M. Anas Nasrulloh e (2021). Penerapan metode k-means clustering data COVID-19 di Provinsi Jakarta. Perbedaan penelitian tersebut dengan peneliti kali ini terdapat pada penggunaan metode *cosine similiarty*.Selain penggunaan *cosine similiarty*, Perolehan data yang di buat tentunya berbeda, di mana penelitian terdahulu mengambil kasus catatan Covid-19 pada bulan maret-juni 2020. Kemudian penelitian selanjutnya memperoleh data per bulan Januari-april 2020. Sedangkan pada penelitian kali ini data di peroleh pada bulan oktober-desember 2020. Peneliti menggunakan metode *cosine similiarty* dan DBI guna untuk menyempurnakan hasil penelitian sebelumnya.

Pada tugas akhir ini berkaitan dengan mengelompokkan daerah penanganan Covid-19 di DKI Jakarta guna untuk menentukan penanganan Covid-19 terbaik. metode yang digunakan yaitu mengklasterisasi persebaran Virus Corona di DKI Jakarta dengan menggunakan Metode *K-Means Clustering* dan di sempurnakan dengan *cosine similiarty*, dan diuji dengan metode DBI. Hasil penelitian ini diharapkan dapat membantu pemerintah DKI Jakarta dalam mengambil keputusan strategis dalam mengurangi kasus persebaran virus Covid-19 di DKI Jakarta.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan diatas, maka dapat dirumuskan pokok permasalahan, yaitu :

a. Berapa tingkat perhitungan DBI dengan menggunakan metode K-Means tanpa *cosine* ?

b. Berapa tingkat perhitungan DBI dengan menggunakan metode K-Means *cosine* ?

## 1.3 TujuanPenelitian

a. Untuk mengetahui tingkst perhitungan K-Means tanpa cosine menggunakan DBI
b. Untuk mengetahui tingkat perhitungan K-Means Cosine menggunakan DBI

## 1.4 Manfaat Penelitian

1) Pemerintah dapat memberi usulan daerah-daerah dengan klaster yang mirip sehingga dapat membantu proses pengambilan keputusan penanganan covid di daerah tersebut.
2) Manfaat bagi peneliti lainnya, dapat dijadikan sumber informasi tambahan atau referensi penelitian berikutnya.
3) Manfaat bagi penulis yaitu, dapat menyelesaikan penelitian yang dikerjakan dan memenuhi tugas penelitian yang dilaksanakan.

## 1.5 Batasan Masalah

Agar pembahasan pada penelitian kali ini tidak terlalu luas dan lebih pokus maka dibatasi hanya pada :

1) Data ini di dapat dari alamat,https://riwayat-file-covid-19-dki-jakarta-jakartagis.hub.arcgis.com/
2) Catatan kasus Covid-19 diperoleh dari data berurutan bulan oktober, november, dan desember tahun 2020.
3) Tools yang di gunakan adalah *Rapid Minner*.
4) Bahwa *Cluster* yang akan dihitung antara 2-10