Daftar Istilah

Istilah		Deskripsi	Halaman pertama kali digunakan
Covid-19	:	*Virus Corona* atau *severe acute respiratory syndrome* corona virus 2 (SARS-CoV-2) adalah virus yang menyerang sistem pernapasan. Penyakit karena infeksi virus ini disebut COVID-19.	1
FCM	:	*Fuzzy C-Means* memiliki kemampuan untuk mengelompokkan data yang besar, data dalam suatu kelompok ditentukan oleh derajat keanggotaannya, Penentuan pusat kelompok dilakukan secara berulang sehingga diperoleh data yang akurat.	1
PSO	:	*Particle Swarm Optimization* adalah algoritma optimasi yang mana berguna untuk mendapatkan nilai akurasi yang baik pada cluster yang akan dibentuk	1
DBI	:	*Davies-Bouldin Index* adalah teknik yang digunakan untuk mengukur validitas sebuah *cluster*.	1
SSW	:	*Sum of square within cluster* adalah persamaan yang digunakan untuk menentukan matriks kohesi dari cluster ke-i. Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap pusat cluster dari sebuah cluster	25
SSB	:	*Sum of Square Between-cluster* (SSB) digunakan untuk mencari hasil separasi antar cluster.	26
Ratio	:	Tujuannya untuk mengetahui nilai perbandingan antara *cluster* i dan *cluster* j.	26

BAB I PENDAHULUAN

I.1 Latar Belakang

Wabah Covid 19 berdampak parah bagi kesehatan, mata pencaharian, lingkungan, psikologi, pendidikan dan transportasi masyarakat yang ada di seluruh dunia. Beberapa keputusan penting harus dibuat untuk memastikan keamanan publik selama krisis ini. Untuk menghindari penyebaran virus, beberapa pembatasan yang telah diberlakukan adalah adanya pembatasan pada mobilitas publik.

Berbagai bentuk Mobilitas dikurangi dengan diberlakukannya kegiatan belajar-mengajar dari rumah (*online*), kantor memberlakukan prosedur bekerja dari rumah (*Work From Home*), dan melakukan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) wilayah dengan jumlah kasus positif yang terlampau banyak. PPKM diumumkan selama bulan Juli 2021. Dan maksud penelitian ini adalah mengelompokkan wilayah di Indonesia berdasarkan tingkat mobilitas selama periode Januari hingga Desember 2021.

Sekitar 47% orang populasi dunia terdampak Virus Covid-19. Penyebab utama penyebaran Covid-19 ini dikarenakan permasalahan mobilitas masyarakat yang tidak baik. Mobilitas terjadi ketika orang-orang berkumpul di berbagai tempat dan ini telah menjadi alasan mengapa Covid-19 menyebar dari satu ke yang lain. Maka dari itu perlu adanya analisis pada daerah yang terkena dampak Covid-19 dan diklasterisasi menjadi daerah berisiko tinggi, sedang dan rendah berdasarkan tingkat mobilitasnya. Pemilihan metode pengelompokan atau metode clustering dalam penelitian ini dimaksudkan untuk mengklasifikasikan karakteristik pada wilayah yang berada dalam satu kelompok dengan wilayah kelompok lain memiliki sifat yang berbeda maupun sama. Clustering pada dasarnya dimaksudkan untuk mengoptimumkan pusat cluster (Centroid) (Jain et al., 1999). Salah satu metode yang dapat digunakan untuk mengelompokkan wilayah berdasarkan kehomogenan maupun perbedaan karakteristiknya yaitu *fuzzy c-means* (FCM) yang menyebutkan kelebihan dari metode FCM adalah mudah diimplementasikan, memiliki kemampuan untuk mengelompokkan data yang besar, data dalam suatu kelompok ditentukan oleh derajat keanggotaannya,

Penentuan pusat kelompok dilakukan secara berulang sehingga diperoleh data yang akurat. “Algoritma FCM adalah salah satu teknik pengelompokan fuzzy yang paling populer karena efisien, lugas, dan mudah diimplementasikan” (Izakian & Abraham, 2011).

Namun kelemahan dengan FCM adalah bahwa ia mengambil centroid awal secara acak. Ini mengarah ke hasil optimal lokal dan menyebabkan keterlambatan konvergensi (Bezdek et al., 1984). Untuk mengoptimalkan akurasi pada FCM, maka diperlukan optimasi untuk menutup kekurangan yang ada dengan tujuan untuk mendapatkan nilai akurasi yang baik pada cluster yang akan dibentuk oleh *Particle Swarm Optimization* (PSO).

PSO bertujuan Untuk menentukan matriks keanggotaan untuk pengelompokkan (Brouwer & Groenwold, 2010). Nilai akurasi dan waktu komputasi digunakan untuk mengukur seberapa baik metode PSO yang diterapkan (Bisilisin et al., 2017).

Penelitian ini serupa dengan penelitian yang mencermati hubungan korelasi antara mobilitas penduduk dengan jumlah terkonfirmasi kasus positif Covid-19 di Jakarta (Ghiffari, 2020). Perbedaannya terletak pada sumber data yang digunakan. Penelitian di Jakarta menggunakan sumber data hasil pelacakan telepon seluler, sementara penelitian ini menggunakan sumber data teragregasi dari Google (Covid-19 *Community Mobility Reports*) dari negara Indonesia. Selain itu, penelitian di Jakarta dilakukan dalam jangka waktu Maret-April 2020, sementara penelitian ini mengamati dinamika mobilitas masyarakat pada periode waktu yang jauh lebih panjang, yaitu dari Januari 2021 sampai Desember 2021. Wilayah dan waktu pengamatan yang berbeda akan memberikan wawasan (*insight*) yang berbeda, dengan manfaat yang berbeda pula.

Google Covid-19 *Community Mobility Reports* mengambil data masyarakat berdasarkan dari pengguna yang mengaktifkan setelan Histori Lokasi pada ponsel mereka. Maka dari itu untuk meningkatkan validitas data yang diambil sebelum mengolah data, di dapatkan data pengguna ponsel android di seluruh Indonesia. Data didapatkan bersumber dari BPS (Badan Pusat Statistik), dan Survei Sosial

Ekonomi Nasional (Susenas) pada tahun 2020. Pengguna ponsel android di Indonesia didapatkan sekitar sekitar 62,84 %. Lalu diperinci lagi untuk setiap provinsinya, untuk Aceh sekitar 59,60 %, Sumatera Utara 59,56%, Sumatera Barat 61,83%, Riau 66,35%, Jambi 64,06%, Sumatera Selatan 60,68%, Bengkulu 60,19%, Lampung 59,03%, Kep. Bangka Belitung 66,61%, Kep. Riau 74,33 %, DKI Jakarta 77,57%, Jawa Barat 64,83%, Jawa Tengah 60,87%, DI Yogyakarta 67,62%, Jawa Timur 61,82%, Banten 64,40%, Bali 69,93%, NTB 56,36%, NTT 44,12%, Kalimantan Barat 58,11%, Kalimantan Tengah 68,56%, Kalimantan Selatan 67,69%, Kalimantan Timur 76,71%, Kalimantan Utara 71,95%, Sulawesi Utara 67,13%, Sulawesi Tengah 57,89%, Sulawesi Selatan 65,14%, Sulawesi Tenggara 63,86%, Gorontalo 61,42%, Sulawesi Barat 54,41%, Maluku 56,16%, Maluku Utara 54,88%, Papua Barat 66,61%, Papua 40,44%.

I.2 Perumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan, maka rumusan masalah yang akan dibahas dalam Tugas Akhir ini adalah sebagai berikut:

1. Data apa yang akan digunakan untuk pengembangan sistem pengelompokan data ini?
2. Metode apa yang akan digunakan pada pengerjaan Tugas Akhir ini?
3. Bagaimana memperoleh hasil kategorisasi daerah terdampak Covid 19 dengan kombinasi *Fuzzy C-Means* dan algoritma optimasi *Particle Swarm Optimization*?
4. Bagaimana menganalisis hasil mobilitas daerah terdampak Covid 19?

I.3 Tujuan Tugas Akhir

Dari beberapa permasalahan yang ada maka tujuan yang ingin dicapai dari tugas akhir ini adalah sebagai berikut

1. Melakukan pengelompokkan daerah terdampak Covid-19 berdasarkan tingkat mobilitas masyarakat Indonesia.
2. Mengimplementasikan algoritma *hybrid Fuzzy C-Means* serta *Particle Swarm Optimization* untuk memperoleh fungsi keanggotaan input dan output.

3. Membuktikan bahwa terjadi peningkatan performa algoritma *Fuzzy-C Means* setelah mengimplementasikan *Particle Swarm Optimization*

I.4 Batasan Tugas Akhir

Batasan tugas akhir ini adalah sebagai berikut:

1. Data yang akan digunakan dalam tugas akhir ini adalah data mobilitas masyarakat Indonesia dari Januari sampai Desember 2021 yang bersumber dari Google Covid-19 *Community Mobility Reports.*
2. Metode *data mining* yang diimplementasikan adalah *clustering* dengan Algoritma *Fuzzy C-Means.*
3. Algoritma optimasi yang diimplementasikan adalah *Particle Swarm Optimization.*
4. Pada penelitian ini data berisikan tempat di mana orang berkumpul untuk tujuan penting dikategorikan menjadi enam kelompok seperti: Ritel dan Rekreasi, Toko Kelontong dan Apotek, Taman, Stasiun Transit, Tempat Kerja, dan kawasan perumahan

I.5 Manfaat Tugas Akhir

Manfaat tugas akhir ini:

1. Bagi pemerintah, penelitian ini bermanfaat dalam memprediksi penurunan penyebaran Covid-19 karena tindakan analisis yang diambil.

2. Bagi peneliti, tugas akhir ini bermanfaat dalam implementasi metode/model/konsep dalam upaya miningkatkan efisiensi dan efektivitas dalam suatu organisasi.

I.6 Sistematika Penulisan

Tugas akhir ini diuraikan dengan sistematika penulisan sebagai berikut:

Bab I Pendahuluan

Bab ini memuat uraian tentang latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, keterbatasan masalah, dan sistematisasi penulisan.

Bab II Tinjauan Pustaka

Bab ini mencakup literatur yang terkait dengan masalah yang diteliti dan juga membahas temuan dari penelitian sebelumnya yang relevan dengan penelitian saat ini.

Bab III Metodologi Penyelesaian Masalah

Bab ini menjelaskan langkah-langkah penelitian secara rinci termasuk apa yang dilakukan pada saat penelitian dan memiliki model makalah yang konseptual dan sistematis sehingga dapat lebih jelas menggambarkan apa yang dilakukan saat ini.

Bab IV Pengumpulan dan Pengolahan Data

Bab ini menjelaskan mengenai gambaran umum objek penelitian meliputi fungsi algoritma *Fuzzy C-Means* dan algoritma optimasi *Particle Swarm Optimization* yang terlibat dalam penelitian ini.

Bab V Analisa Hasil dan Evaluasi

Pada bab Analisis dan Pembahasan akan membahas hasil dari sebuah rancangan Algoritma *hybrid Fuzzy C-Means* dan *Particle Swarm Optimization* serta Hasil Pengujiannya.

Bab VI Kesimpulan dan Saran

Pada bab ini dijelaskan kesimpulan dari hasil penelitian yang dilakukan serta saran terhadap penelitian yang telah dilakukan.

BAB II TINJAUAN PUSTAKA

II.1 Penelitian Terkait

Penelitian dengan menggunakan metode *Fuzzy C-Means* sudah banyak dilakukan sudah ada penelitian terdahulu dengan pengelompokan data menggunakan metode *Clustering*. Sehingga penelitian ini mengambil literatur-literatur untuk dijadikan bahan tinjauan pustaka dalam pemaparan dan pendukung dalam mengimplementasikan metode *hybrid Fuzzy C-Means* dengan *Particle Swarm Optimization* serta untuk membantu dalam menganalisa pada daerah yang terkena dampak Covid-19 dan mengkategorikannya menjadi daerah berisiko tinggi, sedang dan rendah berdasarkan tingkat mobilitasnya dengan menggunakan *Google Community Mobility Report*.

Penelitian Wang & Yamamoto, (2020) yang berjudul "*Using a partial differential equation with Google Mobility data to predict COVID-19 in Arizona*" membuat model perkiraan yang dirancang untuk memprediksi kasus Covid-19 di Arizona dengan bantuan Google CMR. Pada model yang berbasis *Partial Differential Equation* (PDA) ini digunakan untuk pengelompokan wilayah tingkat kabupaten berdasarkan data historis yang dirumuskan dengan metode orde keempat Runge Kuttam dalam menyelesaikan PDA. Dalam penelitian ini, aktivitas manusia yang melintasi batas-batas dari berbagai kabupaten dipertimbangkan dan menghasilkan akurasi prediksi diatas 94%.

Penelitian yang dilakukan Zhou et al., (2020) yang berjudul "*Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data*", peneliti mensimulasikan efek berkurangnya mobilitas atas penyebaran Covid-19 di kota Shenzhen China dengan mengumpulkan data ponsel dari penyedia layanan bernama China Unicom. Orang-orang dari kelompok usia 15 hingga 65 tahun dipertimbangkan dan pergerakan mereka dilacak. Orang-orang dari berbagai tahap Covid-19 seperti tahap suspek, terpapar, terinfeksi, dan pulih dimodelkan menggunakan persamaan diferensial dan simulasi berbagai rentang pembatasan mobilitas diterapkan sehingga keputusan lebih lanjut mengenai mobilitas dapat diambil.

Penelitian yang dilakukan Cartenì et al., (2020) yang berjudul "*How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study*" melakukan analisis pengaruh rutinitas mobilitas terhadap penyebaran pandemi di Italia. Data dikumpulkan dari laporan harian yang diberikan oleh kementerian kesehatan dan transportasi Italia. Riwayat perjalanan 21 hari terakhir dari orang yang terinfeksi dilacak dan model regresi linier berganda dikembangkan dengan mempertimbangkan variabel wilayah geografis, lingkungan, mobilitas dan kesehatan. Hasilnya menegaskan bahwa peningkatan mobilitas berkontribusi pada penyebaran Covid-19.

Penelitian yang dilakukan Jia et al., (2020) yang berjudul "*Population flow drives spatio-temporal distribution of COVID-19 in China*" peneliti menganalisis efektivitas karantina selama pandemi Covid-19. Peneliti juga mengidentifikasi berapa banyak populasi yang pindah dari Wuhan dan mengembangkan model risiko untuk menentukan rasio penyebaran. Algoritma Levenberg–Marquardt digunakan untuk analisis dan hasilnya dapat digunakan untuk membuat keputusan mengenai mobilitas. Sedangkan pada penelitian (Pepe et al., 2020) yang berjudul "*COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown*" peneliti menilai dampak *lockdown* di Italia dengan membangun jaringan kedekatan untuk memperkirakan radius girasi sehingga mengekstraksi pergerakan mingguan dan harian orang. Hasilnya dibandingkan dengan sumber data lain seperti Google CMR.

Penelitian yang dilakukan Azarafza et al., (2020) yang berjudul "*Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran*" melakukan analisis pola penyebaran infeksi Covid-19 di Iran dilakukan dengan menggunakan *clustering* dan sistem informasi geografis. Data diskalakan dan dendrogram dibentuk untuk menemukan penyebaran infeksi dan Teheran dan Qom ditemukan sebagai sumber penyakit yang paling rentan di Iran. Sedangkan penelitian yang dilakukan Sasidharan et al., (2020) yang berjudul "*A vulnerability-based approach to human-mobility reduction for countering COVID-19 transmission in London while considering local air quality*" melakukan penelitian tentang korelasi antara polusi udara dan tingkat kematian Covid-19 di London.

Karena berbagai pembatasan dalam mobilitas, polusi udara menurun ke tingkat yang lebih besar. Analisis ini membantu pemerintah untuk mengambil tindakan yang tepat dalam membatasi mobilitas.

Penelitian yang dilakukan oleh Mekhmoukh & Mokrani, (2015) yang berjudul "*Improved Fuzzy C-Means based Particle Swarm Optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation. Computer Methods and Programs in Biomedicine*" menyatakan bahwa fungsi keanggotaan FCM sensitif terhadap *outlier* dan tidak mengintegrasikan informasi spasial ke dalam segmentasi citra, belum lagi ketergantungan pada inisialisasi *cluster-centric*. Untuk meningkatkan sensitivitas *outlier* pada algoritma FCM yang didukung oleh PSO, pusat *cluster* dalam FCM dipilih secara acak (*random*) dan pusat *cluster* dipilih secara optimal menggunakan PSO. Hasil yang diperoleh dalam penelitian ini menunjukkan efektivitas algoritma yang diusulkan.

Penelitian yang dilakukan Silva Filho et al., (2015) yang berjudul "*Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization*" menyatakan bahwa *fuzzy clustering* telah menjadi area penelitian yang penting dengan banyak aplikasi untuk masalah dunia nyata. Dari teknik pengelompokan *fuzzy*, FCM adalah salah satu yang paling dikenal karena kesederhanaan dan efisiensinya, tetapi memiliki beberapa kelemahan. Secara khusus, ia cenderung jatuh ke minimum lokal. Banyak teknik pengelompokan berbasis optimasi telah diusulkan dalam literatur untuk mengatasi kelemahan ini. Beberapa metode ini hanya mengandalkan optimasi metaheuristik, seperti PSO adalah teknik *hybrid* yang menggabungkan metaheuristik dengan teknik *clustering* partisi tradisional seperti FCM.

Penelitian yang dilakukan Dandekar & Barbastathis, (2020) yang berjudul "*Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning*" peneliti merancang sebuah model epidemiologi untuk mengetahui jumlah orang yang terkena dampak pada setiap titik waktu. Jumlah yang terinfeksi sebenarnya dapat ditemukan dengan menggunakan model *neural network.* Sedangkan penelitian yang dilakukan oleh Saha et al., (2020) yang

berjudul "*Lockdown for COVID-19 and its impact on pupil mobility in India: an analysis of the COVID-19 Community Mobility Reports*" menganalisis Dampak mobilitas masyarakat di India. Peneliti menganalisis data menggunakan lembar excel dan rata-rata kenaikan atau penurunan mobilitas di negara bagian dan teritori persatuan India diwarnai dengan pemetaan data spatio-temporal.

Penelitian yang dilakukan Hadjidemetriou et al., (2020) yang berjudul "*The impact of government measures and human mobility trend on COVID-19 related deaths in the UK*" mengidentifikasi dampak mobilitas dalam hal menggunakan kendaraan pribadi, berjalan kaki, dan kendaraan umum di Inggris. Serta meneliti pengurangan tingkat mobilitas selama *lockdown* memiliki efek serius pada penyebaran Covid-19. Studi ini membantu pemerintah untuk mengambil langkah-langkah efektif dan memutuskan tentang strategi untuk mengatasi situasi tersebut. Dalam penelitian yang dilakukan Kraemer et al., (2020) yang berjudul "*The effect of human mobility and control measures on the COVID-19 epidemic in China*" mempelajari tren mobilitas sebelum dan selama *lockdown* di China yang disesuaikan untuk mengendalikan penyebaran dan tingkat keparahan Covid-19. Rata-rata keterlambatan 6,5 hari Covid-19 diamati karena mobilitas yang berkurang.

II.2 Mobilitas Manusia Dalam Mitigasi Pandemi Covid-19

Pandemi Covid-19 saat ini menimbulkan risiko yang sangat besar di semua tingkatan, mulai dari nasional hingga kelompok masyarakat dalam skala terbatas. Risiko besar ini muncul karena tingginya tingkat ketidakpastian dalam berbagai aspek sehingga menyulitkan para pengambil keputusan untuk membuat prediksi (Aven & Bouder, 2020). Seringkali "nasib" bangsa terancam oleh keputusan strategis yang dibuat berdasarkan situasi seperti ini (Enserink & Kupferschmidt, 2020).

Dalam situasi krisis secara umum, pengambilan keputusan mengenai mitigasi risiko dapat dilakukan di berbagai tingkatan, mulai dari tingkat pemerintah pusat hingga tingkat unit pemerintahan yang paling rendah. Pada semua tingkatan ini, pengambilan keputusan membutuhkan dukungan data yang mengungkapkan fakta tentang situasi saat ini. Salah satu aspek yang sering mendapat perhatian dalam

penanganan krisis penyakit menular adalah mobilitas manusia. Penularan penyakit biasanya karena dua penyebab (Wesolowski et al., 2016).

1. Mobilitas membawa patogen pembawa penyakit ke populasi manusia yang rentan terhadap penyakit tersebut, atau
2. Meningkatkan kontak antara manusia pembawa penyakit dan manusia rentan lainnya.

Beberapa penelitian mengkonfirmasi hubungan antara mobilitas manusia terhadap penyebaran penyakit, seperti influenza (Ferguson et al., 2005), malaria (Wesolowski et al., 2012), DHF (Cummings et al., 2004), dan cacar (Grenfell et al., 2001).

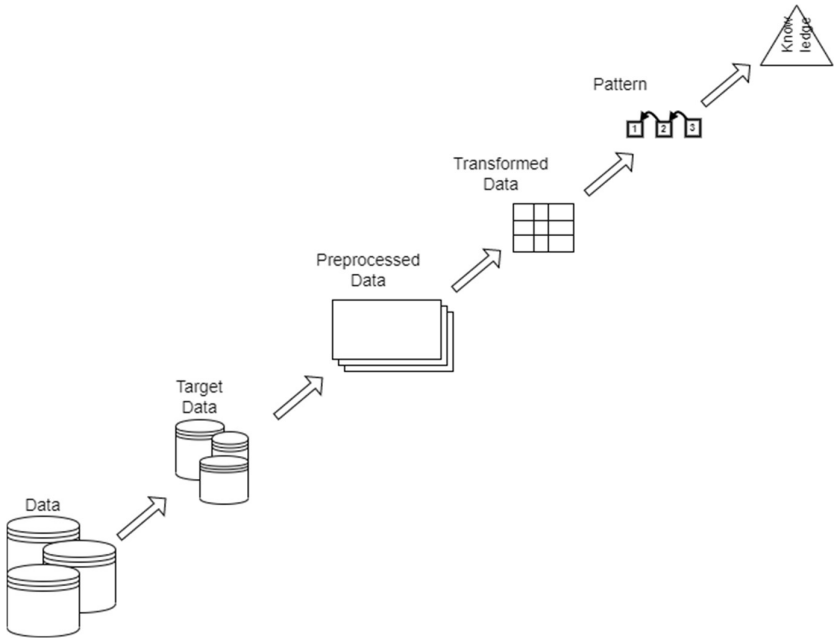
Karena mobilitas manusia terbukti mempengaruhi laju penyebaran, aspek mobilitas ini sering dianggap sebagai faktor penting dalam upaya pengendalian penyebaran penyakit menular. Model mobilitas jaringan global yang dikombinasikan dengan model epidemiologi lokal telah dikembangkan untuk memprediksi dinamika dan pengendalian epidemi Covid-19 di Eropa (Linka et al., 2020). Di China, pola penularan virus SARS-CoV-2 yang terjadi di kota-kota China telah dipelajari dan dipahami menggunakan data mobilitas penduduk di Wuhan (Kraemer et al., 2020).Terkait persebaran Covid-19 di Indonesia, penelitian yang mengkaji hubungan mobilitas orang dari dalam dan luar kota Jakarta terhadap penyebaran Covid-19 di Jakarta menunjukkan korelasi yang tinggi, oleh karena itu, rekomendasi tersebut mencerminkan pentingnya penerapan kebijakan yang efektif yaitu dengan membatasi mobilitas penduduk (Ghiffari, 2020). Studi yang meneliti hubungan mobilitas penduduk dengan laju penyebaran Covid-19 menunjukkan pentingnya pemahaman ini dalam pengambilan keputusan, terutama dalam upaya meminimalkan dampak dan mencegah penyebaran Covid-19.

Ada banyak cara untuk mendapatkan data mobilitas masyarakat, namun ciri khas sumber data ini adalah sifatnya yang personal dan mudah diagregasikan untuk membentuk data *crowd*. Data mobilitas dapat diperoleh dari data lokasi ponsel yang bersifat anonim (Jumadi, 2018). Data mobilitas juga dapat diperoleh dari matriks *Origin-Destination* (OD) yang dibuat dari data panggilan ponsel (*call*

detail records – CDR). Metode ini dapat digunakan untuk mendeteksi perubahan pola mobilitas penduduk dan pengaruhnya terhadap kegiatan ekonomi lokal (Giannotti et al., 2020). Cuitan (*tweets*) di Twitter juga bisa menjadi sumber data mobilitas. *Geotag* yang melekat pada setiap data cuitan digunakan untuk menentukan perpindahan lokasi pengguna, baik yang terjadi dalam hari yang sama atau pada hari yang berbeda (Huang et al., 2020). Sumber data mobilitas yang bersifat agregat juga dapat digunakan untuk menggambarkan kondisi pandemi saat ini, seperti yang dilakukan di Meksiko dengan menggunakan data Covid-19 *Community Mobility Reports* dari Google (Mas, 2021) sedangkan di China yang lebih menggunakan data *mobility index* Baidu (Xi et al., 2020).

II.3 *Knowledge Discovery In Databases* (KDD)

KDD (*Knowledge discovery in Databases*) adalah suatu proses pemancingan atau penambangan data. KDD adalah penerapan metode ilmiah untuk data mining dengan mengidentifikasi pola (*pattern*) dalam data. KDD memberikan solusi untuk memproses, menganalisis, mengatur, dan mengelompokkan data dalam jumlah besar melalui sejumlah proses yaitu seleksi data, *pre-processing data*, transformasi, data mining, dan evaluasi.



Gambar II. 1 Proses KDD

Dari gambar II.1 diatas bisa dilihat beberapa proses KDD yang dimulai dari seleksi data sampai pada proses evaluasi. Berikut adalah langkah-langkah proses KDD sebagai berikut:

II.3.1 Seleksi Data

Seleksi data adalah proses memilih data dari kumpulan data operasional. engan proses awal ini dimungkinkan untuk membuat dataset yang ditargetkan. Data dalam database seringkali tidak semuanya terpakai, sehingga hanya data yang sesuai untuk analisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang menguji faktor kecenderungan orang dalam membeli barang seperti kasus *market basket analysis*, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

II.3.2 *Pre-processing* (*Cleaning Data*)

Pre-processing adalah prosedur operasi dasar untuk membersihkan data dari *noise*, *output*, duplikasi, tipografi, dan lain-lain. Proses ini akan membuat data menjadi lebih kecil tanpa mengubah isi data. Pembersihan dapat dilakukan dengan mengoreksi kesalahan, mengisi data dengan nilai yang hilang, menghilangkan outlier, dan memeriksa data yang tidak konsisten.

II.3.3 Transformasi

Transformasi adalah proses modifikasi data atau penambahan parameter sesuai kebutuhan tujuan yang diinginkan. Proses ini tergantung pada jenis atau pola yang akan ditemukan dalam data. Ada beberapa cara untuk melakukan transformasi data, yaitu:

1. *Smoothing*, yang bertugas untuk menghilangkan *noise* dari data.
2. *Attribute construction*, melakukan penambahan atribut baru untuk membantu proses data mining.
3. *Aggregation*, bertugas untuk operasi agresi pada data.
4. *Normalization*, membuat data atribut dalam skala tertentu sehingga menjadi data yang lebih kecil.
5. *Discretization*, nilai mentah atribut numerik dikonversi menjadi data dengan interval label.