BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era digital seperti saat ini, segala hal menjadi serba mudah, cepat, dan instan dengan adanya internet. Kebutuhan akan informasi dan komunikasi dapat dengan mudah diakses melalui mesin pencarian Google dan melalui media sosial seperti WhatsApp, Instagram, Facebook, dan lain-lain. Kemudahan ini berdampak pada kecepatan pertumbuhan informasi yang semakin tinggi sehingga membuat data menjadi semakin kompleks dan bervariasi dengan volume yang terus meningkat. Fenomena tersebut dikenal dengan istilah *big data*. *Big data* terdiri dari berbagai jenis data yang terstruktur maupun tidak terstruktur seperti teks, audio, video, dan lain-lain. Pada tahun 2020 volume data diperkirakan mencapai 44 *zettabyte* atau 44 triliun *gigabyte* (Turner, 2014).

Dalam mencari informasi – informasi penting dari data yang sangat besar tentu tidak mudah. Kita memerlukan suatu teknik yang dapat mengektraksi dan mengidentifikasi informasi penting dari berbagai *database* yang besar dengan cepat (Suyanto, 2019). Teknik tersebut dikenal dengan *data mining* atau penambangan data. *Data mining* secara umum diartikan sebagai pencarian pola tersembunyi yang mungkin ada pada database yang besar. *Data mining* memiliki tujuan yaitu menemukan hubungan dan karakteristik yang menarik yang mungkin ada secara implisit dalam basis data. Menurut Fayyad dan Smyth (1996), terdapat enam prosedur yang dapat dilakukan dalam penambangan data, yaitu klasterisasi, klasifikasi, regresi, deteksi anomali, perangkuman, dan pembelajaran aturan asosiasi.

Dalam penelitian ini, prosedur penambangan data yang akan dilakukan yaitu klasterisasi atau analisis klaster. Analisis klaster adalah salah satu teknik multivariat yang mempunyai tujuan utama yaitu mengelompokkan objek – objek berdasarkan karakteristik yang dimilikinya. Jadi dengan analisis klaster, kumpulan data yang besar akan dipartisi menjadi beberapa kelompok kecil sesuai dengan kesamaan karakteristik yang dimilikinya sehingga data menjadi lebih mudah dipahami.

Metode - metode dalam analisis klaster telah berkembang sejak tahun 1950-an. Menurut L. Kaufman dan P.J Rousseeuw (1990), secara garis besar metode analisis klaster dibagi menjadi dua yaitu metode hierarki dan metode partisi. Metode hierarki adalah suatu metode analisis klaster yang membentuk tingkatan tertentu seperti pada struktur pohon (dendogram) karena proses pengklasteran dilakukan secara bertahap. Sedangkan metode partisi adalah metode analisis klaster yang bekerja dengan cara membagi atau mempartisi data ke dalam sejumlah kelompok dengan menentukan k objek perwakilan yang terletak di pusat klaster yang didefinisikan. Perbedaan antara kedua metode tersebut adalah pada metode partisi banyaknya klaster atau kelompok ditentukan terlebih dahulu sedangkan pada hierarki jumlah klaster tidak ditentukan terlebih dahulu.

Menurut Suyanto (2019), metode berbasis hierarki terbagi menjadi metode *agglomerative* (pemusatan) dan *divisive* (penyebaran). Metode *agglomerative* dimulai dengan menganggap setiap objek tunggal sebagai sebuah klaster, kemudian secara iteratif menggabungkannya untuk membentuk klaster-klaster yang lebih besar. Sedangkan metode *divisive* dimulai dengan sebuah klaster besar yang berisi semua objek dalam himpunan data, kemudian secara iteratif dipecah ke dalam klaster - klaster yang lebih kecil. Kelemahan dari metode hierarki yaitu tidak dapat membatalkan apa yang telah dilakukan sebelumnya. Maksudnya yaitu ketika dua objek digabungkan maka kedua objek tersebut akan selalu berada dalam klaster yang sama dan begitupun sebaliknya. Selain itu, banyaknya klaster yang terbentuk bersifat subjektivitas peneliti dengan melihat dendogram dan tidak efektif dalam pengelompokkan dataset besar.

Metode berbasis partisi dapat dibagi menjadi metode *k-means* dan *k-medoids. K-means* dikembangkan oleh TJ.A Hartigan dan M.A. Wong pada tahun 1979. Metode ini banyak digunakan pada data atau objek berskala kecil maupun menengah. Pada *k-means*, pusat-pusat klaster diperbaharui berdasarkan rata-rata jarak dari objek yang ada di setiap klaster. Oleh karena itu algoritma ini sangat lemah terhadap keberadaan data pencilan. Hal ini karena data pencilan dapat membuat nilai rata-rata (*mean*) menjadi bias yang berakibat pada hasil klasterisasi menjadi kurang representatif dalam menggambarkan karakteristik data. Untuk memudahkan, pencilan seringkali dibuang, padahal acapkali mengandung suatu informasi penting (Suyanto, 2019).

Untuk mengatasi kelemahan pada metode *k-means*, maka dapat digunakan metode lain yang kuat terhadap pencilan yaitu *k-medoids*. Pada metode *k-medoids* penentuan atau pembaharuan pusat-pusat klaster bukan berdasarkan nilai rata-rata melainkan dengan memilih sebuah objek representatif dari suatu klaster yang disebut *medoid.* Kelebihan *k-medoids* selain kuat terhadap keberadaan data pencilan yaitu klaster yang terbentuk tidak bergantung pada urutan objek yang diperiksa dan hasil klasterisasi tidak akan berubah sehubungan dengan dilakukannya transformasi ortogonal pada titik data (Ng & Han, 2002). Algoritma pertama dalam metode *k-medoids* adalah *Partitioning Around Medoids* (PAM) yang dikembangkan oleh Kaufman dan Rousseeuw pada tahun 1990. Pada algoritma ini, semua objek non representatif diperhitungkan dalam proses penggantian objek representatif sehingga iterasi yang dilakukan memerlukan waktu yang lama. Oleh karenanya algoritma ini hanya efektif untuk data dengan jumlah kecil maupun menengah yaitu dibawah 100 (Kaufman & Rousseeuw, 1990). Untuk mengatasi kelemahan pada algoritma PAM, Kaufman dan Rousseeuw kemudian membangun algoritma baru yang disebut *Clustering LARge Application* (CLARA) guna mengelompokkan dataset besar. CLARA bekerja dengan membentuk sampel dari himpunan data besar, kemudian menerapkan algoritma PAM pada sampel tersebut sehingga diperoleh *k* medoid sampel. Kelemahan pada algoritma ini yaitu *k* medoid yang terpilih dari sampel seringkali

bukan *k* medoid terbaik dalam dataset besarnya, sehingga kualitas klaster yang dihasilkan seringkali tidak sebaik PAM.

Guna memperbaiki algoritma PAM dan CLARA, Raymond T. Ng dan Jiawei Ham memperkenalkan algoritma baru yang disebut *Clustering Large Application based on RANdomized Search* (CLARANS). CLARANS bekerja dengan mengkombinasikan teknik *sampling* dan algoritma PAM. Pada pengelompokkan dataset besar, CLARANS lebih efisien dibanding PAM sebab tidak mengecek seluruh objek non medoid. Sedangkan jika dibandingkan dengan CLARA, algoritma CLARANS menghasilkan kualitas klasterisasi yang lebih baik sebab pencarian tidak terbatas pada area lokal saja (Ng & Han, 2002).

Di awal tahun 2020, masyarakat dunia diresahkan dengan adanya pandemi *Coronavirus Disease* (Covid-19). Penyakit ini pertama kali terkonfirmasi di Wuhan, China pada tanggal 31 Desember 2019. Kemudian menyebar dengan cepat ke berbagai negara termasuk Indonesia. Kasus Covid-19 pertama kali terkonfirmasi di Indonesia pada tanggal 2 Maret 2020 dan dengan cepat menyebar ke berbagai daerah (Kemenkes RI, 2020). Melansir dari laman Covid19.go.id, diketahui per tanggal 1 Februari 2021 total kasus Covid-19 di Indonesia mencapai 1.089.308 kasus. Dimana DKI Jakarta, Jawa Barat, dan Jawa Tengah menjadi tiga provinsi dengan jumlah kasus terbanyak. Guna memutus rantai penyebaran Covid-19, berbagai upaya telah dilakukan oleh pemerintah seperti melakukan karantina wilayah dengan menerapkan Pembatasan Sosial Berskala Besar (PSBB) dan menghimbau masyarakat untuk selalu mematuhi protokol kesehatan.

Pandemi Covid-19 yang berkepanjangan menyebabkan timbulnya berbagai permasalahan sosial ekonomi bagi masyarakat. Salah satu masalah yang timbul yaitu meningkatnya angka kemiskinan. Menurut Badan Pusat Statistik (BPS), jumlah penduduk miskin di Indonesia tahun 2020 mengalami peningkatan sebanyak 2,76 juta jiwa jika dibandingkan dengan tahun sebelumnya. Hal ini merupakan dampak lanjutan dari permasalahan sosial ekonomi lainnya seperti terjadinya resesi ekonomi, tingginya angka pengangguran akibat Pemutusan Hubungan Kerja (PHK), dan menurunnya daya beli masyarakat. Permasalahan ini

tentu berdampak negatif bagi masyarakat karena dapat menyebabkan tingginya angka kriminalitas, tertutupnya akses pendidikan, dan tingginya angka kematian. Oleh karena itu, diperlukan penanganan yang serius dari pemerintah guna mengatasi permasalahan ini.

Berdasarkan pemaparan yang telah disampaikan, penulis tertarik untuk mengkaji lebih dalam mengenai pengelompokkan dengan algoritma CLARANS pada dataset besar. Fokus studi kasus dari penelitian ini yaitu mengelompokkan kota/kabupaten di Indonesia berdasarkan tingkat kemiskinannya. Oleh karena itu, penelitian ini diberi judul "Analisis Klaster pada Dataset Besar dengan Algoritma CLARANS (Studi Kasus : Tingkat Kemiskinan di 221 Kota/Kabupaten di Indonesia Tahun 2020)".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana hasil penerapan analisis klaster dengan menggunakan algoritma CLARANS pada data tingkat kemiskinan di 221 kota/kabupaten di Indonesia tahun 2020?
2. Bagaimana karakteristik yang terbentuk dari setiap klaster dengan menggunakan algoritma CLARANS?

1.3 Tujuan Penelitian

Sesuai dengan rumusan masalah di atas, maka tujuan penelitian ini adalah:

1. Mengimplementasikan dan memperoleh hasil dari analisis klaster dengan menggunakan algoritma CLARANS pada data tingkat kemiskinan di 221 kota/kabupaten di Indonesia tahun 2020.
2. Mengetahui karakteristik dari setiap klaster yang terbentuk dengan menggunakan algoritma CLARANS.

1.4 Pembatasan Masalah

Agar tujuan dari penelitian ini tercapai, maka diperlukan adanya pembatasan masalah. Data yang digunakan dalam penelitian ini merupakan data sekunder yang bertipe numerik. Pengelompokkan didasarkan pada ukuran jarak dan ukuran jarak yang digunakan adalah jarak Euclid.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Secara Teoritis

   Secara teoritis hasil penelitian ini diharapkan dapat bermanfaat dalam memberikan informasi dan pengetahuan menyangkut implementasi metode *k-medoids* dengan algoritma CLARANS pada studi kasus tingkat kemiskinan di 221 kota/kabupaten di Indonesia tahun 2020. Selain itu diharapkan dapat menjadi bahan referensi bagi penelitian - penelitian selanjutnya yang memilik topik yang selaras.

2. Secara Praktis

   Secara praktis hasil penelitian ini diharapkan dapat menambah wawasan tentang seberapa pentingnya penanganan masalah kemiskinan di Indonesia akibat dampak pandemi Covid-19.