PENGELOMPOKKAN DATA RAWAN BENCANA ALAM DI BERBAGAI KOTA DAN KABUPATEN DI INDONESIA BERBASIS ALGORITMA K-MEANS *CLUSTERING*

(Skripsi)

Oleh

Sherly Martina Mulyadi

2015061021

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

PENGELOMPOKKAN DATA RAWAN BENCANA ALAM DI BERBAGAI KOTA DAN KABUPATEN DI INDONESIA BERBASIS ALGORITMA K-MEANS *CLUSTERING*

Oleh

Sherly Martina Mulyadi

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Program Studi Teknik Elektro
Jurusan Teknik Elektro
Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ABSTRAK

PENGELOMPOKKAN DATA RAWAN BENCANA ALAM DI BERBAGAI KOTA DAN KABUPATEN DI INDONESIA BERBASIS ALGORITMA K-MEANS *CLUSTERING*

Oleh

SHERLY MARTINA MULYADI

Indonesia terletak di wilayah cincin api pasifik dan beriklim tropis menyebabkan rawan terhadap bencana alam seperti banjir, gempa bumi, angin puting beliung, letusan gunung api, tanah longsor, dan kekeringan. Bencana alam memberikan dampak kerugian bagi manusia dan lingkungan. Kerugian dapat dikurangi dengan melakukan upaya mitigasi. Salah satu strategi mitigasi, yaitu dengan informasi terkait daerah rawan bencana. Berdasarkan situasi tersebut, penelitian ini memiliki tujuan untuk *clustering* data rawan bencana alam di Indonesia berdasarkan jumlah terjadinya bencana alam sehingga dapat membantu pemangku kepentingan dalam mengidentifikasi daerah rawan bencana. *Clustering* dilakukan menggunakan algoritma *data mining K-Means* dan metode pengembangan *Cross Industry Standard Process for Data Mining* (CRISP-DM). Jenis bencana alam yang digunakan dalam penelitian ini adalah banjir, tanah longsor, kebakaran hutan dan lahan, gelombang pasang atau abrasi, kekeringan, serta angin puting beliung. Metode *elbow* dimanfaatkan untuk menentukan jumlah *cluster* terbaik. Hasil dari penelitian memperlihatkan bahwa *elbow method* dapat digunakan untuk menentukan jumlah *cluster* yang optimal. Hal ini dibuktikan dengan kekohesifan antara objek dengan *centroid* terbaik terdapat di *cluster* yang dihasilkan oleh model menggunakan atribut kekeringan dengan *silhouette coefficient* sebesar 0,9 dan *davies bouldin index* sebesar 0,39. Penelitian ini menghasilkan lima *cluster,* yaitu *cluster* 0 terdiri atas 391 daerah, *cluster* 1 sebanyak 2 daerah, *cluster* 2 sebanyak 64 daerah, *cluster* 3 sebanyak 25, dan *cluster* 4 sebanyak 11 daerah. Penelitian ini juga membangun visualisasi hasil *clustering* berupa *mapping* daerah rawan bencana alam di Indonesia menggunakan GeoPandas yang mencakup 494 daerah.

Kata kunci : bencana alam, *clustering*, CRISP-DM, elbow method, K-Means

ABSTRAK

CLUSTERING OF NATURAL DISASTER PRONE DATA IN VARIOUS CITIES AND DISTRICTS IN INDONESIA BASED ON K-MEANS CLUSTERING ALGORITHM

Oleh

SHERLY MARTINA MULYADI

Indonesia is located in the Pacific Ring of Fire and has a tropical climate, making it prone to natural disasters such as floods, earthquakes, tornadoes, volcanic eruptions, landslides, and droughts. Natural disasters cause losses to humans and the environment. Losses can be reduced by making mitigation efforts. One of the mitigation strategies is information related to disaster-prone areas. Based on this situation, this study aims to cluster data on natural disaster prone areas in Indonesia based on the number of occurrences of natural disasters so that it can help stakeholders in identifying disaster-prone areas. Clustering is done using the K-Means data mining algorithm and the Cross Industry Standard Process for Data Mining (CRISP-DM) development method. The types of natural disasters used in this research are flood, landslide, forest and land fire, tidal wave or abrasion, drought, and tornado. The elbow method was utilized to determine the best number of clusters. The results of the study show that the elbow method can be used to determine the optimal number of clusters. This is evidenced by the cohesiveness between objects with the best centroid in the cluster produced by the model using the drought attribute with a silhouette coefficient of 0.9 and a davies bouldin index of 0.39. This study produced five clusters, namely cluster 0 consisting of 391 regions, cluster 1 as many as 2 regions, cluster 2 as many as 64 regions, cluster 3 as many as 25, and cluster 4 as many as 11 regions. This research also built a visualization of clustering results in the form of mapping natural disaster prone areas in Indonesia using GeoPandas which covers 494 regions.

Keywords: clustering, CRISP-DM, elbow method, K-Means, natural disaster

Judul Skripsi	: **PENGELOMPOKKAN DATA RAWAN BENCANA ALAM DI BERBAGAI KOTA DAN KABUPATEN DI INDONESIA BERBASIS ALGORITMA K-MEANS *CLUSTERING***
Nama Mahasiswa	: Sherly Martina Mulyadi
Nomor Pokok Mahasiswa	: 2015061021
Jurusan	: Teknik Elektro
Fakultas	: Teknik

MENYETUJUI

1. Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Ir. Muhamad Komarudin, S.T., M.T.	**Ir. Titin Yulianti, S.T., M.Eng.**
NIP. 196812071997031006	NIP. 198807092019032015

2. Mengetahui

Ketua Jurusan Teknik Elektro	Ketua Program Studi Teknik Informatika
Herlinawati, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 197103141999032001	NIP. 197312262000122001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Ir. Muhamad Komarudin, S.T., M.T.**

   Sekretaris : **Ir. Titin Yulianti, S.T., M.Eng.**

   Penguji : **Yessi Mulyani, S.T., M.T.**

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.
NIP. 197509282001121002

Tanggal Lulus Ujian Skripsi : **19 Januari 2024**

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya dengan judul "Pengelompokkan Data Rawan Bencana Alam di Berbagai Kota dan Kabupaten di Indonesia Berbasis Algoritma K-Means *Clustering*" dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 26 Januari 2024

Pembuat pernyataan,

Sherly Maruna Mulyadi

NPM 2015061021

RIWAYAT HIDUP

Penulis dilahirkan di Bandarlampung, pada tanggal 27 Maret 2002. Penulis merupakan anak satu-satunya dari pasangan Bapak Aos Mulyadi dan Ibu Anita.

Penulis menyelesaikan pendidikannya di SD Negeri 2 Rajabasa pada tahun 2014, SMP Negeri 2 Bandar Lampung pada tahun 2017, dan SMA Negeri 9 Bandar Lampung pada tahun 2020. Pada tahun 2020, penulis terdaftar sebagai mahasiswa Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik Universitas Lampung melalui jalur SNMPTN. Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan, antara lain:

1. Menjadi anggota biasa Himpunan Mahasiswa Teknik Elektro Universitas Lampung, Departemen Pendidikan dan Pengembangan Diri, Divisi Pendidikan periode 2021/2022 dan Departemen Pengembangan Keteknikan, Divisi Penelitian dan Pengembangan periode 2022/2023.
2. Menjadi asisten Laboratorium Teknik Digital Universitas Lampung pada tahun 2022 sampai tahun 2023.
3. Mengikuti program Studi Independen Kampus Merdeka dari Kementerian Pendidikan dan Budaya dengan mengambil kelas Data Analytics di Zenius Education pada tahun 2022.
4. Mengikuti program Magang Bersertifikat Kampus Merdeka Batch VI di Bank Indonesia sebagai Data Analyst pada 6 Maret 2023 sampai 27 Juni 2023.
5. Melaksanakan Kuliah Kerja Nyata di Desa Belu, Kecamatan Kota Agung Barat, Kabupaten Tanggamus, Provinsi Lampung pada bulan Januari sampai dengan Februari 2023.

6. Mengikuti penelitian mengenai Prediksi Kemampuan Membayar *Customer* menggunakan *Machine Learning* bersama dengan dosen PSTI pada tahun 2023.

MOTTO

“Pada akhirnya takdir Allah selalu baik walaupun terkadang perlu air mata untuk menerimanya.”
(**Umar Bin Khattab**)

“Maka sesungguhnya bersama kesulitan ada kemudahan.”
(**QS. Al-Insyirah: 5**)

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya.”
(**QS. Al-Baqarah: 286**)

“Sesungguhnya Allah tidak akan merubah suatu nasib kaum sebelum mereka mengubah keadaan diri mereka sendiri.”
(**QS. Ar-Ra’d: 11**)

“Keberhasilan bukanlah milik orang yang pintar. Keberhasilan adalah kepunyaan mereka yang senantiasa berusaha.”
(**B.J. Habibie**)

PERSEMBAHAN

Bismillahirrahmanirrahim, segala puji syukur atas kehadirat Allah SWT, yang telah melimpahkan rahmat, hidayah, dan karunia-Nya sehingga saya dapat menyelesaikan skripsi ini.

SAYA PERSEMBAHKAN SKRIPSI INI KEPADA:

"Kedua orang tua saya yang senantiasa memberikan dukungan, kasih sayang, dan doa-doa yang diberikkan untukku. Saya mengucapkan terimakasih karena telah membesarkan saya dengan penuh kasih sayang, mendidik, dan mengajarkan saya berbagai hal. Terimakasih ayah dan ibu atas pengorbanan telah yang dilakukan selama ini untuk memberikan yang terbaik untuk saya. Terimakasih telah memberikan saya kesempatan menimba ilmu di perguruan tinggi. Semoga ilmu yang saya dapatkan dan cita-cita yang saya gapai nanti dapat menjadi amal Jariyah bagi ayah dan ibu."

"Diri saya sendiri yang telah melakukan yang terbaik hingga skripsi ini selesai. Terimakasih karena sudah selalu bekerja keras, tidak menyerah, dan tetap yakin bisa melawati seluruh rintangan. Semoga seluruh impian dan cita-cita dapat segera tercapai."

"Among dan Ajong Ua yang senantiasa mendoakan dan memberikan dukungan kepadaku untuk kemudahan dan kelancaran dalam setiap hal yang saya lakukan. Semoga Among dan Ajong Ua selalu diberi kesehatan dan selalu dalam perlindungan Allah SWT."

SANWACANA

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayat-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi ini dengan judul "Pengelompokkan Data Rawan Bencan Alam di Berbagai Kota dan Kabupaten di Indonesia Berbasis Algoritma K-Means *Clustering*". Dalam pelaksanaan dan pembuatan skripsi ini penulis menerima dukungan baik secara moral maupun materil yang sangat berharga dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada semua pihak yang telah membantu, khususnya kepada:

1. Bapak Dr. Eng. Helmy Fitriawan, S.T., M.Sc., selaku Dekan Fakultas Teknik Universitas Lampung;
2. Ibu Herlinawati, S.T., M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung;
3. Ibu Yessi Mulyani, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung, penguji, dan pembimbing akademik yang telah memberikan saran serta masukan terhadap penelitian ini;
4. Bapak Ir. Muhamad Komarudin, S.T., M.T. selaku Pembimbing Utama yang selalu meluangkan waktunya untuk memberikan bimbingan, arahan dalam proses pengerjaan penelitian, dan dukungan kepada penulis dalam penyelesaian penelitian ini;
5. Ibu Ir. Titin Yulianti, S.T., M.Eng., selaku Pembimbing Pendamping yang telah memberikan dukungan, bimbingan, masukan dan arahan secara detail terhadap penyelesaian skripsi ini;
6. Para dosen, civitas akademika di Jurusan Teknik Informatika Unila, dan Mbak Rika selaku Admin Program Studi Teknik Informatika yang telah banyak membantu penulis dalam segala urusan administrasi selama perkuliahan;

7. Kedua orang tua tercinta yang tidak hentinya memberikan doa, semangat, dukungan, dan materi sehingga penulis dapat menyelesaikan skripsi dengan baik;
8. Teman-teman Uts Gabuts Michel, Amanda, Feny, Rio, Niki, Hamzah, dan Fajar yang telah membantu penulis dan menjadi tempat bertukar pikiran dari semester satu hingga sekarang. Terutama Michel dan Amanda yang selalu berbagi informasi terkait pengerjaan skripsi, melakukan bimbingan, dan mengurus berkas wisuda bersama dengan penulis.
9. Bejo dan Yuki, adikku yang selalu menghibur penulis dengan tingkah dan mukanya yang lucu sehingga menghilangkan rasa lelah saat mengerjakan skripsi.
10. Seluruh pihak yang terlibat langsung maupun tidak langsung dalam penyelesaian skripsi.

Akhir kata, semoga laporan ini dapat menjadi referensi bagi pengembangan keilmuwan di bidang Teknik Informatika dan bermanfaat bagi yang membacanya

Bandar Lampung, 26 Januari 2024
Penulis,

Sherly Martina Mulyadi

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Bencana alam merupakan bencana yang diakibatkan oleh peristiwa alam. Negara Indonesia secara geografis terletak di Cincin Api Pasifik (*Ring of Fire*), yaitu wilayah pertemuan antara Lempeng Eurasia, Lempeng Indo-Australia, Lempeng Filipina, dan Lempeng Pasifik. Aktivitas tektonik yang dihasilkan dari keempat lempeng tersebut menyebabkan terjadinya bencana alam gempa bumi di Indonesia. Gempa bumi dari kegiatan tektonik dapat memicu terjadinya bencana tsunami. Lalu, terdapat sabuk vulkanik (*volcanic arc*) yang terbentang dari pulau Sumatera, Jawa, Nusa Tenggara, hingga Sulawesi dimana sisi dari sabuk berupa pegunungan vulkanik. Jumlah gunung api aktif yang terdapat di wilayah Indonesia adalah 127 gunung. Aktivitas dari gunung api aktif menyebabkan terjadinya bencana letusan gunung api di negara Indonesia. Kondisi geografis tersebut mengakibatkan Indonesia rawan terhadap bencana alam geologi, seperti gempa bumi, tsunami, dan letusan gunung api [1].

Negara Indonesia berada di garis khatulistiwa sehingga merupakan negara yang beriklim tropis. Indonesia hanya memiliki dua musim saja, yaitu musim hujan dan musim kemarau. Saat musim hujan, curah hujan di Indonesia cukup tinggi. Hal tersebut dapat menyebabkan terjadinya bencana banjir karena tempat penampungan air telah mencapai batas maksimal dan kurangnya daerah resapan air. Saat musim kemarau, Indonesia mengalami cuaca yang cerah dan suhu yang panas. Keadaan ini dapat menyebabkan bencana kekeringan serta kebakaran hutan dan lahan karena curah hujan yang rendah membuat jumlah air yang tersedia terbatas. Kondisi iklim tersebut mengakibatkan Indonesia rawan terhadap bencana

alam hidrometeorologi, seperti banjir, tanah longsor, kebakaran hutan dan lahan, gelombang ekstrem dan abrasi, serta kekeringan [1].

Menurut data bencana alam yang terjadi di Indonesia dari Badan Nasional Penanggulangan Bencana (BNPB) sepanjang tahun 2022 tercatat bencana banjir sebanyak 1.531 kasus, cuaca ekstrem sebanyak 1.068 kasus, tanah longsor sebanyak 634 kasus, kebakaran hutan dan lahan sebanyak 252 kasus, gempa bumi sebanyak 28 kasus, gelombang pasang dan abrasi sebanyak 26 kasus, kekeringan sebanyak 4 kasus, serta erupsi gunung api sebanyak 1 kasus. Dampak dari bencana alam tersebut bagi masyarakat meliputi 858 korban meninggal, 8.733 korban luka-luka, 37 korban hilang, dan 6.144.534 korban menderita serta harus mengungsi. Selain itu dampak kerusakan yang ditimbulkan oleh bencana alam tahun 2022, yaitu 95.403 rumah mengalami kerusakan, 1.241 fasilitas pendidikan rusak, 647 fasilitas ibadah rusak, 95 fasilitas kesehatan rusak, dan 342 jembatan rusak [2].

Dampak dari bencana alam pada tahun 2022 menunjukkan bahwa bencana alam memberikan kerugian bagi masyarakat dan lingkungan di daerah yang terjadi bencana alam karena menimbulkan korban jiwa, menggangu aktivitas masyarakat, merusak fasilitas umum, memberikan rasa trauma bagi korban bencana, dan menghilangkan harta benda. Untuk mengurangi potensi kerugian yang ditimbulkan oleh bencana alam, maka perlu dilakukan upaya mitigasi bencana.

Mitigasi bencana merupakan upaya yang dilakukan untuk meminimalisir risiko dari terjadinya bencana alam. Dalam menyusun strategi mitigasi bencana alam bagi kabupaten atau kota di Indonesia, dibutuhkan informasi terkait daerah yang rawan bencana alam. Dengan mengetahui daerah yang rawan terhadap bencana alam, maka pemerintah di setiap kabupaten atau kota dapat menentukan upaya mitigasi yang tepat dan efektif sesuai dengan jenis bencana alam yang rawan terjadi di daerah tersebut sehingga dapat mengurangi dampak yang ditimbulkan semaksimal mungkin.

Informasi terkait daerah yang rawan bencana alam di Indonesia diperoleh dari hasil *clustering* data rawan bencana alam yang terjadi menurut kota atau kabupaten. Teknologi yang digunakan untuk mengolah dan menganalisis data rawan bencana alam adalah *data mining*. *Data mining* secara umum merupakan proses pengolahan data untuk mengambil informasi penting dalam data. Salah satu teknik *data mining* adalah *clustering* yang berfungsi untuk mengelompokkan data menjadi beberapa kelompok. Jenis bencana alam yang terdapat dalam data adalah tanah longsor, banjir, kebakaran hutan dan lahan, gelombang pasang atau abrasi, angin puting beliung, serta kekeringan. Algoritma *K-Means* merupakan salah satu teknik *clustering.* Algoritma ini digunakan untuk membagi kota atau kabupaten di Indonesia menjadi beberapa *cluster* berdasarkan jumlah bencana alam yang pernah terjadi.

1.3 Rumusan Masalah

Berdasarkan latar belakang tersebut, maka rumusan masalah dari penelitian ini sebagai berikut:

1. Bagaimana mengklasterisasi data rawan bencana alam di Indonesia menjadi beberapa *cluster* kerawanan bencana berdasarkan jumlah bencana alam yang pernah terjadi?
2. Bagaimana melakukan pengelompokkan data rawan bencana alam di Indonesia menggunakan algoritma *data mining K-Means*?
3. Apakah *elbow method* dapat menentukan jumlah *cluster* yang optimal dalam pengelompokkan data rawan bencana alam di Indonesia?
4. Bagaimana membangun visualisasi data *clustering* yang diperoleh menggunakan GeoPandas?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini sebagai berikut:

1. Melakukan klusterisasi data rawan bencana alam di Indonesia beberapa *cluster* kerawanan bencana berdasarkan jumlah bencana alam yang pernah terjadi.

2. Mengelompokkan data rawan bencana alam di Indonesia menggunakan algoritma *data mining* K-Means.
3. Menerapkan *elbow method* dalam menentukan jumlah *cluster* optimal untuk pengelompokkan data rawan bencana alam di Indonesia.
4. Membangun visualisasi data *clustering* yang diperoleh menggunakan *GeoPandas*.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini sebagai berikut :

1. Memberikan referensi metode yang dapat digunakan oleh akademisi untuk melakukan *clustering* daerah rawan bencana alam.
2. Membantu pemangku kepentingan dalam mengidentifikasi daerah-daerah yang rawan bencana alam sehingga dapat menentukan upaya mitigasi yang tepat dan efektif.
3. Memberikan informasi kepada masyarakat terkait daerah rawan bencana di Indonesia sehingga dapat meningkatkan kesiapsiagaan masyarakat dalam menghadapi bencana alam.

1.6 Batasan Masalah

Batasan masalah dari penelitiannya ini adalah data yang digunakan pada penelitian ini merupakan data bencana alam yang terjadi di Indonesia dari bulan Januari 2020 hingga bulan Desember tahun 2022 dan tidak melakukan pengujian terhadap performa visualisasi yang dibangun.

1.7 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang mengapa perlu dilakukannya penelitian, rumusan masalah yang memuat masalah yang

akan diteliti di penelitian, tujuan dilakukannya penelitian, dan batasan masalah yang dibahas dalam penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini berisi mengenai teori-teori yang digunakan sebagai referensi dalam penelitian seperti bencana alam, Badan Nasional Penanggulangan Bencana, indeks rawan bencana, data mining, *Unsupervised Learning*, teknik *clustering*, algoritma *K-Means*, *Outlier*, *Elbow Method*, *Silhouette Coefficient, Davies Bouldin Index,* CRISP-DM, Python, Google Colaboratory, GeoPandas, dan penelitian terkait.

BAB III METODOLOGI PENELITIAN

Bab ini membahas terkait waktu dan tempat penelitian dilakukan, alat dan bahan yang digunakan dalam mengerjakan penelitian, tahapan dari pengerjaan penelitian, dan tahapan dari metodologi yang digunakan pada penelitian, yaitu CRISP-DM.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas tahapan *data understanding*, tahapan *data preparation*, tahapan *modeling*, tahapan *evaluation*, tahapan *deployment* serta memberikan analisis terhadap hasil clustering daerah rawan bencana alam.

BAB V KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan yang diperoleh dari pembahasan hasil melakukan penelitian dan saran-saran untuk pengembangan penelitian lebih lanjut.

II. TINJAUAN PUSTAKA

2.1 Bencana Alam

Bencana alam merupakan peristiwa yang mengganggu dan merugikan kehidupan serta pencaharian masyarakat yang disebabkan oleh faktor alam. Jenis bencana yang termasuk bencana alam antara lain gempa bumi, banjir, tsunami, kekeringan, letusan gunung api, tanah longsor, kebakaran hutan dan lahan, dan angin topan. Bencana alam memberikan dampak negatif bagi manusia dan lingkungan karena menimbulkan korban jiwa, merusak lingkungan, merugikan harta benda, serta memberikan dampak psikologis bagi korban bencana [3]. Berdasarkan Data Informasi Bencana Indonesia (DIBI), jenis bencana alam yang disebabkan oleh iklim merupakan bencana alam yang paling banyak terjadi di Indonesia dari tahun 2010 sampai dengan tahun 2019. Bencana banjir terjadi sebanyak 7.691 kasus, angin puting beliung sebanyak 7.172 kasus, tanah longsor sebanyak 5.240 kasus, kebakaran hutan dan lahan sebanyak 1.833 kasus, serta kekeringan sbenyak 877 kasus [4].

2.1.1 Jenis Bencana Alam

1. Banjir adalah bencana alam dimana air menggenangi suatu wilayah selama jangka waktu tertentu. Bencana banjir disebabkan oleh curah hujan yang tinggi dan turun secara terus menerus sehingga terjadi luapan air sungai, laut, danau, atau saluran air karena jumlah air melebihi batas maksimum daya tampung. Banjir dengan arus air yang cepat dapat menghanyutkan manusia, hewan, dan harta benda sehingga dapat menimbulkan korban jiwa. Air banjir dapat merusak pondasi bangunan, jembatan, dan lainnya yang dilalui. Banjir merusak tanaman karena merendam tanaman dengan material yang dibawa oleh air [5].

2. Tanah longsor merupakan bencana yang biasanya disebabkan oleh gabungan dari turunnya curah hujan yang tinggi di lereng dengan tanah yang bersifat kurang padat dan terjal, terjadi pengikisan, kurangnya tutupan vegetasi, serta terdapat getaran. Bencana longsor seringkali terjadi dengan cepat sehingga tidak dapat melakukan evakuasi mandiri karena terbatasnya waktu. Dampak yang ditimbulkan oleh tanah longsor adalah rusaknya jalan, kabel, dan pipa karena gerakan dari tanah longsor atau tertimbun material tanah longsor. Lalu, tanah longsor menciptakan rekahan pada tanah yang membuat fondasi dari bangunan terpisah dan runtuhan batuan yang meluncur dapat merusak bangunan atau pemukiman yang terdapat di bawahnya [5].
3. Angin puting beliung merupakan bencana yang diakibatkan oleh peristiwa hidrometeorologis. Waktu terjadinya puting beliung sulit diperkirakan dan akan meningkat intesitasnya pada masa peralihan musim. Bencana ini menyebabkan rumah rusak dan pohon tumbang [5].
4. Kekeringan adalah bencana alam yang terjadi saat mengalami kekurangan pasokan air dari curah hujan selama jangka waktu tertentu seringkali satu musim atau lebih. Kekurangan pasokan air menyebabkan jumlah air yang tersedia jauh di bawah jumlah yang dibutuhkan. Kekeringan membuat tanaman mati karena kekurangan air untuk proses fotosintesis. Apabila hal tersebut berlangsung dalam waktu yang lama maka akan mengurangi jumlah bahan pangan karena tanaman pangan dan ternak mati menyebabkan masyarakat kesulitan untuk mendapatkan bahan pangan [5].
5. Kebakaran hutan dan lahan merupakan kondisi saat hutan dan lahan dilanda api sehingga menyebabkan perubahan baik secara langsung maupun tidak langsung terhadap sifat fisik dari hutan atau lahan dan hayati yang terdapat di dalamnya. Hal tersebut mengurangi fungsi hutan atau lahan dalam menopang kehidupan yang berkesinambungan. Kerugian yang ditimbulkan dari jenis bencana alam ini adalah mengganggu ekosistem karena menghilangnya flora dan fauna, merusak fasilitas umum, menimbulkan korban jiwa, dan lebih lanjut menimbulkan gangguan pernafasan [5].
6. Gelombang pasang atau abrasi adalah bencana yang disebabkan oleh badai yang mengakibatkan terjadinya perubahan besar pada permukaan air laut di sepanjang pantai. Dampak dari bencana gelombang pasang atau abrasi adalah

merusak tempat tinggal di daerah bencana menyebabkan warga kehilangan tempat tinggal dan menggenangi lahan pertanian dan pertambakan sehingga penduduk kehilangan mata pencaharian dan penghasilan menjadi berkurang. Selain itu, gelombang yang besar dapat menyebabkan kerusakan jalan, bangunan, jembatan, dan merusak infrastruktur lainnya yang berada di tepi pantai [6].

2.2 Badan Nasional Penanggulangan Bencana

Badan Nasional Penanggulangan Bencana (BNPB) didirikan pada tanggal 26 Januari tahun 2008 seiring dengan dikeluarkannya Peraturan Presiden Nomor 8 Tahun 2008 tentang Badan Nasional Penanggulangan Bencana [7]. BNPB memiliki tugas dan fungsi sesuai dengan peraturan presiden tersebut, sebagai berikut :

a. Tugas BNPB [8]
   1. Memberikan arahan dan panduan yang adil dan merata pada usaha penanggulangan bencana yang meliputi pencegahan bencana, tanggap darurat, rekontruksi, dan rehabilitasi;
   2. Menetapkan persyaratan dan standar dalam menyelenggarakan penanggulangan bencana sesuai dengan peraturan perundang-undangan;
   3. Menyediakan informasi kepada masyarakat tentang aktivitas penanggulangan bencana;
   4. Melaporkan kepada Presiden setiap waktu saat dalam keadaan darurat bencana dan sebulan sekali saat keadaan normal mengenai penyelenggaraan penanggulangan bencana;
   5. Memanfaatkan dan memberikan pertanggungjawaban sumbangan atau bantuan baik yang berasal dari dalam negeri maupun luar negeri;
   6. Memberikan pertanggungjawaban terhadap penggunaan dana yang diperoleh dari Anggaran Pendapatan dan Belanja Negara;
   7. Melakukan tugas-tugas lain sesuai dengan peraturan perundang-undangan; dan
   8. Membuat dasar pembentukan Badan Penanggulangan Bencana Daerah.

b. Fungsi BNPB [8]

1. Merumuskan dan menetapkan kebijakan penanggulangan bencana dan penanganan pengungsi dengan merespon secara cepat, tepat, efisien, serta efektif.
2. Mengoordinasian pelaksanaan operasi penanggulangan bencana dengan menyeluruh, terpandu, dan terencana.

Geoportal Data Bencana Indonesia merupakan layanan resmi milik BNPB. Layanan ini menyediakan data bencana alam yang terjadi di Indonesia yang dapat dilihat dan diunduh oleh siapapun, jurnal kebencanaan, infografis mengenai jumlah bencana alam yang terjadi dalam periode waktu tertentu, dan *mapping* untuk memantau bencana. Data dan layanan Geoportal data Bencana Indonesia dikelola oleh BNPB [9]. Data yang digunakan pada penelitian ini diperoleh dari layanan Geoportal Data Bencana Indonesia, yaitu data bencana alam yang terjadi di Indonesia dari bulan Januari tahun 2020 sampai dengan bulan Desember tahun 2022. Data bencana alam yang digunakan berjumlah 14.785 meliputi bencana banjir, angin puting beliung, tanah longsor, kebakaran hutan dan lahan, kekeringan serta gelombang pasang atau abrasi.

2.3 Indeks Rawan Bencana

Indeks rawan bencana digunakan untuk menginformasikan tingkat rawan bencana dari setiap kota atau kabupaten di Indonesia. Indeks rawan bencana dikeluarkan oleh Badan Nasional Penanggulangan Bencana (BNPB) pada tahun 2011. Komponen yang mempengaruhi Indeks Rawan Bencana Indonesia adalah frekuensi terjadinya bencana, banyaknya korban luka-luka, korban meninggal, kepadatan penduduk, kerusakan rumah, dan kerusakan fasilitas umum. Besaran bobot dari setiap komponen, yaitu frekuensi bencana sebesar 30%, jumlah korban meninggal, luka-luka, dan kepadatan penduduk sebesar 50%, serta jumlah kerusakan rumah dan kerusakan fasilitas umum sebesar 20%. Besar nilai dari setiap komponen dijumlahkan untuk mendapatkan total nilai dari sebuah kota atau kabupaten. Total nilai digunakan untuk menentukan tingkat kerawanan dari kota

atau kabupaten. Hasil dari penilaian indeks rawan bencana menghasilkan tiga kategori rawan bencana, yaitu rendah, sedang, dan tinggi [10]. Salah satu komponen indeks rawan bencana adalah jumlah terjadinya bencana alam di suatu kota atau kabupaten dimana komponen tersebut berbanding lurus terhadap indeks rawan bencana. Oleh karena itu, apabila jumlah terjadinya suatu bencana alam tinggi, maka indeks kerawanan dari bencana alam tersebut juga tinggi.

Tabel 1. Penilaian Skor Parameter [10]

PARAMETER	% Bobot	Nilai	Kelas	Bobot	SKOR (Kelas x Bobot)
1. JUMLAH KEJADIAN BENCANA					
BANJIR	30%	> 0 dan < 4	1	5	5
		4 - 15	2		10
		>15	3		15
GEMPA BUMI		< 2	1	5	5
		2 - 3	2		10
		> 4	3		15
GEMPA BUMI DAN TSUNAMI		< 2	1	3	3
		2 - 3	2		6
		> 4	3		9
KEBAKARAN PERMUKIMAN		< 5	1	3	3
		5 - 24	2		6
		> 24	3		9
KEKERINGAN		1	1	3	3
ANGIN TOPAN		< 4	1	3	3
		4 - 15	2		6
		> 15	3		9
BANJIR DAN TANAH LONGSOR	30%	< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
TANAH LONGSOR		< 4	1	5	5
		4 - 15	2		10
		> 15	3		15
LETUSAN GUNUNG API		< 2	1	5	5
		2 - 3	2		10
		> 3	3		15
GELOMBANG PASANG / ABRASI		< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
KEBAKARAN HUTAN DAN LAHAN		< 3	1	3	3
		3 - 8	2		6
		> 8	3		9
KECELAKAAN INDUSTRI		< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
KECELAKAAN TRANSPORTASI		< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
KONFLIK / KERUSUHAN SOSIAL	30%	< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
KEJADIAN LUAR BIASA		< 2	1	3	3
		2 - 3	2		6
		> 3	3		9
2. JUMLAH KORBAN MENINGGAL	50%	< 40	1	5	5
		40 - 1599	2		10
		> 1599	3		15
3. JUMLAH KORBAN LUKA-LUKA		< 40	1	3	3
		40 - 1599	2		6
		> 1599	3		9
4. JUMLAH KEPADATAN PENDUDUK		< 25	1	5	5
		25 - 624	2		10
		> 624	3		15
5. JUMLAH KERUSAKAN RUMAH	20%	< 50	1	4	4
		50 - 2499	2		8
		> 2499	3		12
6. JUMLAH KERUSAKAN FASILITAS UMUM DAN INFRASTRUKTUR		< 20	1	4	4
		20 - 399	2		8
		> 399	3		12

Cara menghitung skor dari setiap parameter adalah mengalikan jumlah dari parameter dengan bobot. Lalu, menentukan kelas dan bobot dari parameter berdasarkan nilai dari hasil perkalian. Kemudian mengalikan kelas dan bobot untuk mendapatkan skor dari parameter. Selanjutnya menjumlahkan seluruh skor dari setiap parameter untuk mendapatkan skor total agar mengetahui kategori rawam bencana dari suatu daerah.

Tabel 2. Klasifikasi Tingkat Tawan Bencana [10]

Kelas	Kategori Rawan Bencana
1	Kerawanan Rendah
2	Kerawanan Sedang
3	Kerawanan Tinggi

2.4 *Data Mining*

Data mining adalah proses menggali atau mencari informasi dan pola-pola tidak diketahui dari data yang besar (*big data*) dan bermanfaat bagi seluruh pihak yang memiliki kepentingan. Informasi dan pola-pola yang ditemukan dapat dimanfaatkan sebagai alat pendukung dalam pengambilan keputusan dan digunakan untuk menyelesaikan permasalahan. Dengan menggunakan *data mining*, perusahaan dapat menemukan informasi dalam *big data* melalui pengolahan dengan menggunakan metode pada *data mining* [11].

2.3.1 Tahapan *Data Mining*

Adapun tahapan yang dimiliki oleh *data mining* adalah sebagai berikut [11]:

1. Identifikasi tujuan adalah tahapan yang dilakukan untuk mengetahui domain atau area yang akan diambil serta tujuan yang akan dicapai.
2. Spesifikasi permasalahan adalah tahapan merencanakan dan menyusun domain aplikasi yang akan dilaksanakan untuk menyelesaikan permasalahan, mengumpulkan informasi terkait pengetahuan-pengetahuan yang berkaitan dari para ahli, dan memutuskan tujuan akhir.

3. *Data selection* adalah tahapan memilih data dari database dengan melihat hubungan data dan diselarasakan dengan pengetahuan para ahli yang terkait.
4. *Preprocessing data* adalah tahapan yang mencakup operasi membersihkan data (*data cleaning)*, melakukan integrasi data, mentransformasikan data (*data transformation)*, dan mengurangi jumlah data termasuk memilih serta mengekstraksi fitur dalam dataset.
5. *Data mining* adalah tahapan yang bertujuan untuk mengekstraksi pola-pola data yang valid dan menentukan jenis *data mining* beserta algoritma *data mining* yang akan digunakan untuk mencari pola-pola data, yaitu algoritma *K-Means.*
6. *Evaluation and interpretation* adalah tahapan membuat perkiraan, mengidentifikasi, dan mengartikan pola-pola penting yang diperoleh dari tahapan *data mining*.
7. *Discovered knowledge* adalah tahapan untuk menyatukan pengetahuan yang diperoleh dengan sistem lain untuk proses lebih dalam seperti menyatukan pengetahuan dengan sistem visualisasi untuk menyampaikan dan memberitahukan pengetahuan kepada pengguna.

2.3.2 Pengelompokkan *Data Mining*

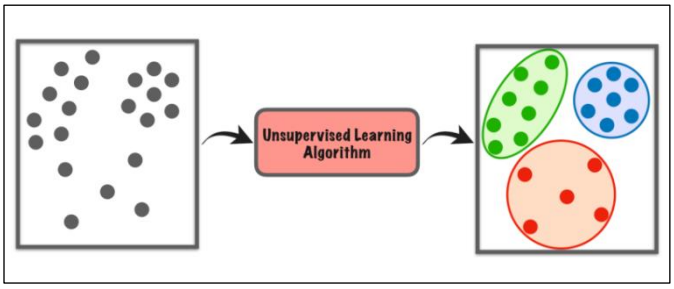
Data mining secara garis besar dibagi menjadi dua, yaitu [11]:

1. Strategi prediktif adalah model yang terdiri atas beberapa prediktor yang merupakan variabel yang condong untuk berpengaruh pada hasil prediksi. Model ini dimanfaatkan untuk meramalkan hasil di masa depan (*outcome*) dengan melibatkan fungsi *supervised learning* guna meramalkan nilai target. Metode yang tergolong dalam strategi prediktif, yaitu c*lassification, time series analysis, regression, outlier analysis,* dan *evolution analysis*.
2. Strategi deskriptif merupakan strategi untuk membuat hubungan sebab-akibat, tabulasi silang, frekuensi, dan lainnya. Strategi ini digunakan untuk memperoleh keadaan yang teratur dalam data dan untuk menunjukkan pola yang didapatkan. Fokus dari strategi deskriptif adalah meringkas dan mengubah data menjadi informasi yang berarti untuk melakukan pelaporan dan

pengamatan. Teknik yang termasuk dalam strategi ini adalah *clustering, summarization, association rules, correlations*, dan *characterization*.

2.5 *Unsupervised Learning*

Unsupervised Learning merupakan teknik pembelajaran *machine learning* yang tidak memerlukan pengawasan terhadap modelnya dan memperbolehkan model untuk bekerja sendiri agar dapat memperoleh informasi yang diperlukan [12]. Proses pada teknik ini diawali dengan melatih model untuk mendalami dan mencari pola yang didapatkan dari data input. Lalu, melakukan analisis terhadap pola-pola yang ditemukan dan mengelompokkan data-data yang memiliki kemiripan ke dalam area tertentu. *Unsupervised learning* tidak membutuhkan label pada data karena hanya mengelompokkan data, tidak melakukan prediksi data. Teknik ini disebut *learning by reasonin*g karena melatih model untuk menggunakan nalar dan memutuskan fitur yang serupa atau berlainan dari setiap data. Secara umum teknik *unsupervised learning* dibagi menjadi empat jenis kasus, yaitu *clustering, association, anomaly detection, dan dimensionality reduction* [13].



Gambar 1. Unsupervised Learning [14]

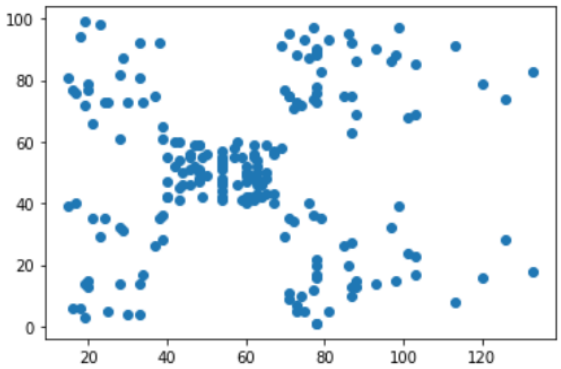
2.6 *Clustering*

Clustering merupakan salah satu jenis dari teknik *unsupervised learning* yang mengelompokkan objek berdasarkan kesamaan karakteristik (*similarity*) yang dimiliki antara satu data dengan data yang lain. *Clustering* bertujuan untuk mengelompokkan data dimana data yang terdapat dalam sebuah *cluster* harus

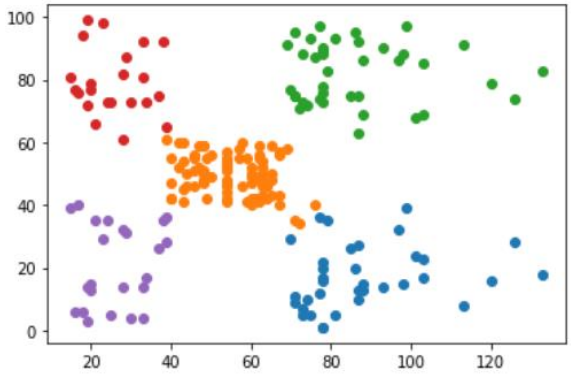
hampir sama serta data yang terdapat dalam satu *cluster* harus berbeda dibandingkan data dalam *cluster* lainnya [15]. *Clustering* dalam mengelompokkan sekumpulan data menjadi beberapa kelompok tanpa pengawasan sehingga tidak melakukan latihan atau *training* dan tidak membutuhkan target *output* [16]. Metode *clustering*, antara lain *Fuzzy C-Means, K-Medoids, Self-Organizing Map (SOM), K-Means*, dan lainya.

Metode *hierarchical clustering* adalah suatu metode *clustering* yang diawali dengan melakukan pengelompokkan beberapa data berdasarkan kemiripan paling dekat yang dimiliki oleh setiap data. Lalu, dilanjutkan dengan mengelompokkan data dengan kemiripan terdekat kedua dan seterusnya hingga *cluster* membentuk seperti pohon yang memiliki tingkatan atau hierarki yang jelas antar data dimulai dari yang paling mirip hingga yang paling tidak mirip.

Metode *non-hierarchical clustering* merupakan metode yang dimulai dengan menentukan jumlah *cluster* yang diinginkan terlebih dahulu. Setelah itu, metode melakukan proses pengelompokkan data berdasarkan kemiripan yang dimiliki tetapi tidak menggunakan proses tingkatan atau hierarki. Salah satu contoh dari metode ini adalah algoritma *K-Means*.



Gambar 2. Data Sebelum *Clustering*



Gambar 3. Data Setelah *Clustering*

2.7 Algoritma *K-Means*

Algoritma *K-Means* merupakan salah satu metode *clustering* untuk mengelompokkan data yang tidak memiliki label dan dilakukan tanpa proses hierarki. Algoritma *K-Means* menggunakan sistem partisi dalam proses pengelompokkan data menjadi beberapa *cluster* berdasarkan jarak [17]. Pengelompokkan data ke dalam satu *cluster* berbasis kesamaan atribut yang dimiliki. Cara untuk mengetahui kesamaan adalah menghitung jarak setiap data dengan titik pusat *cluster* atau *centroid*. Data dengan karakteristik yang sama dikelompokkan dalam satu *cluster* sedangkan data yang mempunyai karakteristik berlainan dikelompokkan ke *cluster* lain [18]. Tahapan-tahapan dalam mengelompokkan data menggunakan algoritma *K-Means*, sebagai berikut [19]:

1. Menentukan jumlah *cluster*.
2. Melakukan inisialisasi titik pusat *cluster* atau *centroid* awal secara random.
3. Mengukur jarak setiap objek terhadap setiap *centroid* menggunakan metode *Euclidean Distance* dengan rumus persamaan (2.1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_j)^2} \tag{2.1}$$

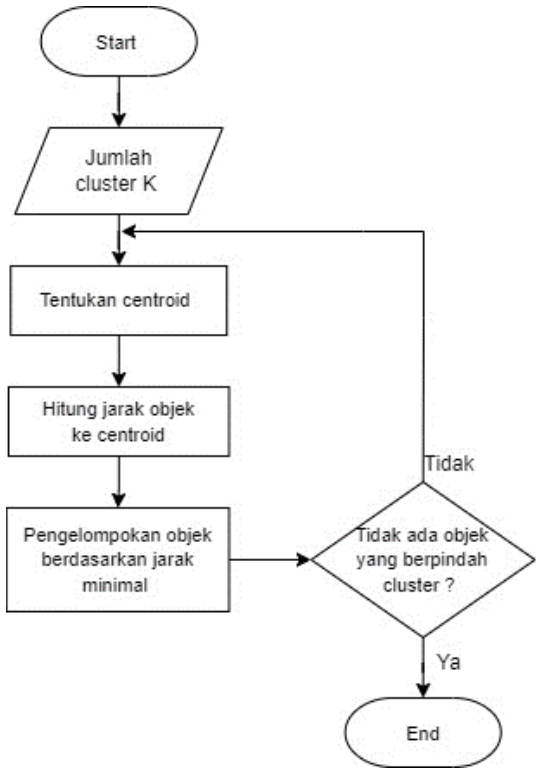
Keterangan :

$d(x,y)$ = jarak antar data ke-i dengan pusat cluster j

x_i = variabel data ke-i di atribut data ke-n

y_j = variabel titik pusat ke-j pada atribut ke-n

4. Menghitung kembali pusat *cluster* baru dengan keanggotaan *cluster* yang baru untuk setiap *cluster.* Pusat *cluster* diperoleh dengan menghitung rata-rata dari seluruh objek yang terdapat dalam *cluster*.
5. Menghitung kembali jarak setiap objek dengan *cluster* yang baru atau melakukan perulangan pada langkah 3 dan 4 sampai dengan tidak ada anggota *cluster* yang berubah.



Gambar 4. Diagram Alur Algoritma *K-Means*

Algoritma *K-Means* dimulai dari menentukan jumlah dari *cluster*. Lalu, memilih beberapa objek secara random dalam populasi untuk dijadikan sebagai *centroid*. Kemudian, menghitung jarak dari setiap objek terhadap *centroid* yang telah dipilih. Selanjutnya mengelompokkan objek-objek ke dalam *cluster* berdasarkan jarak terdekat objek ke *centroid*. Setelah itu, menghitung kembali *centroid* baru dan jarak objek ke *centroid* baru apabila terdapat objek yang berpindah *cluster*. Apabila tidak terdapat objek yang berpindah *cluster* maka perulangan selesai.

2.6.1 Kelebihan dan Kekurangan Algoritma *K-Means*

Kelebihan yang dimiliki oleh algoritma *K-Means*, antara lain mudah untuk diimplementasikan, waktu yang dibutuhkan untuk mempelajarinya cepat, dapat melakukan pengelompokkan pada data berjumlah besar dengan cepat dan efisien, serta algoritma *K-Means* bersifat fleksibel dan sederhana [[17], [20]]. Sedangkan kelemahan dari algoritma *K-Means*, antara lain merupakan sebuah keharusan untuk menentukan jumlah *cluster* terlebih dahulu, hanya dapat menggunakan data yang atributnya bertipe numerik dan rata-ratanya dapat ditentukan, serta sensitif terhadap data yang memiliki *noisy* dan *outlier* [21]. Adapun parameter dari algoritma K-Means untuk melakukan pengelompokkan data dapat dilihat pada Tabel 3.

Tabel 3. Parameter Algoritma K-Means [22]

Parameter	Keterangan
n_clusters	jumlah cluster yang akan dibentuk dalam proses *clustering*.
Init	parameter yang digunakan untuk melakukan inisialisasi *centroid* awal.
n_init	jumlah menjalankan algoritma *K-Means* dengan *centroid* yang berbeda.
max_iter	jumlah maksimum melakukan iterasi algoritma *K-Means* dalam sekali *running*.
Tol	toleransi relatif dari perbedaan centroid untuk dua iterasi yang berbeda.
Verbose	mode verbositas
random_state	menentukan jumlah pembangkitan bilangan acak untuk inisialisasi *centroid.*
copy_x	untuk menentukan apakah data awal telah dimodifikasi atau tidak.
Algorithm	algoritma K-Means yang digunakan.

2.8 *Outlier*

Data pencilan atau *outlier* merupakan data yang memiliki nilai yang berbeda jauh dengan sebagian besar nilai yang terdapat dalam kelompoknya atau disebut nilai ekstrem [23]. *Outlier* akan membuat sebuah *cluster* menjadi tidak padat karena akan dianggap berada di daerah yang berkepadatan rendah. Deteksi *outlier* bertujuan untuk memisahkan data yang bernilai ekstrem dengan data yang bernilai normal. Selain itu, deteksi *outlier* digunakan untuk menemukan anomali yang terdapat dalam kelompok data sehingga dikenal juga sebagai *unsupervised anomaly detection* [24].

Deteksi *outlier* dapat dilakukan menggunakan metode *Inter Quartilr Range* (IQR) dan metode *Percentile.* Metode IQR mendeteksi *outlier* yang nilainya kurang dari hasil pengurangan persentil ke-25 dari data dengan (1,5x *inter quartile range*) dan nilainya lebih dari hasil penjumlahan persentil ke-75 dengan (1,5x *inter quartile range*). IQR didefinisikan sebagai hasil dari pengurangan persentil ke-75 (Q3) dengan persentil ke-25 (Q1) [25]. Metode *percentile* berbeda dengan metode IQR dalam mendeteksi *outlier*. Pada metode *percentile* dapat memperpanjang rentang batasan untuk mendeteksi *outlier* seperti data yang memiliki distribusi besar yang mana jika menggunakan metode IQR akan banyak data yang dideteksi sebagai *outlier* sehingga metode tersebut mungkin kurang optimal maka dapat menggunakan metode *percentile.* Contohnya, besar Q1 menjadi persentil ke-5 dan besar Q3 menjadi persentil ke-95 [26].

2.9 *Elbow Method*

Salah satu kelemahan yang dimiliki oleh algoritma *k-means* adalah harus mengetahui jumlah *cluster* yang ingin dibentuk sebelum membangun model. Agar tidak terjadi *overfitting* data karena jumlah *cluster* yang terlalu banyak atau *underfitting* karena jumlah *cluster* yang terlalu sedikit [27]. Terdapat beberapa metode yang dapat digunakan untuk menentukan jumlah *cluster* yang optimal, antara lain, *silhouette coefficient, davies bouldin index, calinski harabasz index,* dan *elbow method*. Elbow method dibandingkan dengan metode lain dapat bekerja

pada dataset yang memiliki baris data besar. Selain itu, metode ini dapat digunakan pada algoritma pengelompokkan dan menggunakan metrik jarak dalam menghitung kesamaan dari setiap data.

Metode *Elbow* merupakan metode yang dimanfaatkan untuk memberikan informasi mengenai jumlah *cluster* yang terbaik dengan cara menampilkan persentase hasil perbandingan antara jumlah *cluster* dalam bentuk siku pada suatu titik. Jumlah *cluster* yang terbaik didapatkan dengan menghitung *Sum of Square Error* (SSE) dari setiap *cluster*. SSE berfungsi menghitung selisih antara data yang didapatkan dengan model perkiraan yang telah dilakukan. *Cluster* terbaik berada pada titik siku terakhir sebelum *plot* menurun membentuk garis lurus [28]. Langkah-langkah dalam menentukan nilai *cluster* yang terbaik sebagai berikut [29] :

1. Persamaan rumus SSE sebagai berikut.

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_k} \| X_i - C_k \|_2^2 \quad (2.2)$$

   Keterangan :
   K = jumlah cluster yang terbentuk
   C_i = cluster ke i
   X = data yang yang terdapat dalam setiap cluster
2. Secara acak menentukan titik pusat dari *cluster. Centroid* awal ditentukan secara acak dari objek yang tersedia dengan jumlah sama dengan *cluster* K. Lalu, menghitung *centroid cluster* ke-i selanjutnya menggunakan rumus persamaan 2.3 berikut.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \text{ dengan i = 1, 2, 3, ,,, n} \quad (2.3)$$

3. Menghitung jarak antara setiap objek dengan setiap *centroid* menggunakan *Euclidian Distance* dengan rumus persamaan 2.4.

$$d(x, y) = \| x - y \| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \text{ ; i = } 1, 2, 3, , , , \text{n} \quad (2.4)$$

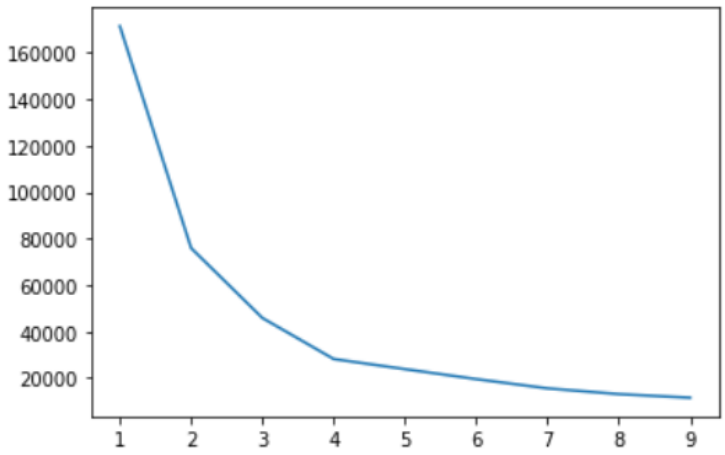
Keterangan :

x_i = variabel pada objek x ke i

y_i= variabel keluaran y

n = jumlah objek

4. Mengalokasikan tiap objek ke dalam *centroid* terdekat.
5. Mengalokasi objek ke dalam setiap *cluster* pada iterasi dengan K-Means dengan jarak yang paling dekat dengan titik pusat *cluster*.
6. Melakukan iterasi lalu proses menentukan posisi *centroid* yang baru dengan rumus pada langkah ke 2.
7. Apabila letak *centroid* yang baru tidak sama dengan yang lama maka melakukan kembali langkah ke 3.



Gambar 5. *Elbow Method*

2.10 *Silhouette Coefficient*

Silhouette coefficient adalah metrik yang digunakan untuk mengevaluasi performa *clustering*. Variabel yang digunakan pada metrik ini adalah variabel nilai jarak rata-rata antara seluruh objek yang terdapat dalam satu *cluster* dan variabel nilai jarak rata-rata antara objek dengan seluruh objek yang terdapat dalam *cluster* terdekat yang bukan *cluster*-nya sendiri [30]. *Silhouette coefficient* mengevaluasi seberapa baik kualitas pengelompokkan dan pemisahan objek-objek.

Rentang nilai dari *silhouette coefficient* adalah -1 hingga 1. Jika nilai *silhouette coefficient* mendekati 1 maka objek-objek yang terdapat dalam satu *cluster* yang

sama memiliki jarak antar objek yang semakin dekat dan semakin jauh dengan *cluster* lain. Jika nilai mendekati 0 maka menandakan bahwa terdapat *cluster* yang tumpang tindih. Jika nilai kurang dari 0 atau mendekati -1 mengindikasi bahwa objek dimasukkan ke dalam *cluster* yang salah karena terdapat *cluster* lain yang jaraknya lebih dekat dengan objek tersebut atau jarak antar objek di dalam *cluster* semakin jauh sedangkan jarak antara objek dengan *cluster* lain semakin dekat. Rumus untuk *silhouette coefficient* adalah sebagai berikut [31]:

$$Silhouette\ Coefficient = \frac{(b-a)}{\max(a,b)} \tag{2.5}$$

Keterangan :
b = jarak rata-rata antara objek dengan seluruh objek di *cluster* lain yang terdekat
a = jarak rata-rata antara seluruh objek dalam satu *cluster*

Adapun parameter dari *silhouette coefficient* untuk melakukan pengelompokkan data dapat dilihat pada Tabel 4.

Tabel 4. Parameter *Silhouette Coefficient* [32]

Parameter	**Keterangan**
x	Array yang menyimpan data-data atribut yang digunakan untuk *clustering.*
labels	Cluster yang didapatkan dari hasil *clustering* untuk setiap data.
Metric	Untuk menghitung jarak antara objek yang terdapat dalam array atribut.
Sample_size	Ukuran dari sampel yang digunakan saat menghitung *silhouette coefficient*.
Random_state	menentukan jumlah pembangkitan bilangan acak untuk inisialisasi *centroid.*

Adapun interpretasi dari nilai *Silhouette Coefficient* dapat dilihat pada tabel 5 berikut.

Tabel 5. Interpretasi *Silhouette Coefficient* [33]

Silhouette Coefficient	Interpretasi
0,71 – 1,00	*Cluster* memiliki struktur yang kuat
0,51 – 0,70	*Cluster* memiliki struktur standar (*medium*)
0,26 – 0,50	*Cluster* memiliki struktur yang lemah
≤ 0,25	*Cluster* tidak memiliki struktur

2.11 *Davies Bouldin Index*

Davies Bouldin Index (DBI) merupakan metrik evaluasi menggunakan nilai separasi dan kohesi dari objek untuk mengetahui jumlah k cluster yang optimal. Nilai separasi adalah jarak antara nilai pusat (*center*) dari *cluster*. Nilai kohesi adalah jarak rata-rata antara setiap objek dalam satu *cluster* dengan *centroid* dari *cluster*-nya. *Sum of Square Within Cluster* (SSW) digunakan untuk mencari nilai kohesi dengan rumus 2.6 [34]:

$$SSW_I = \frac{1}{mi}\sum_{j=1}^{mi} d(x_j, c_i) \tag{2.6}$$

Keterangan :

m_i = jumlah objek dalam *cluster* ke-i

c_i = *centroid* dari cluster ke-i

$d(x_j, c_i)$ = jarak dari setiap objek terhadap *centroid* i yang dihitung menggunakan metode *Euclidean Distance*.

Adapun rumus untuk menghitung jarak setiap objek dengan *centroid* menggunakan metode *Euclidean Distance,* sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_j)^2} \tag{2.7}$$

Keterangan :

$d(x, y)$ = jarak antar data ke-i dengan pusat cluster j

x_i = variabel data ke-i di atribut data ke-n

y_j = variabel titik pusat ke-j pada atribut ke-n

Nilai separasi dihitung menggunakan *Sum of Between Cluster* (SSB) dengan rumus 2.8.

$$SSB_{I,J} = D(C_I, C_J) \tag{2.8}$$

Keterangan :

$D(C_I, C_J)$ = jarak *centroid* ke-i terhadap *centroid* ke-j di cluster yang lain

Setelah itu, mencari nilai rasio dengan cara membandingkan nilai dari penjumlahan matrik kohesi *cluster* ke i dan matrik kohesi *cluster* ke j dengan nilai separasi dari *cluster* ke i dan *cluster* ke j menggunakan rumus 2.9.

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{2.9}$$

Keterangan :

SSW_i = *Sum of Square Within Cluster* di *centroid* i

$SSB_{i,j}$ = *Sum of Between Cluster centroid ke-i* terhadap *centroid* ke-j dimana j terletak di *cluster* yang berbeda.

Nilai rasio kemudian ditambahkan hingga jumlah *cluster* terakhir dimana nilai ini digunakan sebagai nilai dari *Davies Bouldin Index*.

$$DBI = \frac{1}{K}\sum_{K=1}^{K} max_{i \neq j}(R_{i,j}) \tag{2.10}$$

Dimana K adalah banyaknya *cluster*.

Rentang nilai dari *davies bouldin indeks* adalah 0 sampai dengan 1. Semakin nilai *davies bouldin indeks* mendekati 0 maka menunjukkan semakin baik *cluster-cluster* dipisahkan dan semakin padu *cluster* yang dihasilkan. Sebaliknya, apabila nilai semakin mendekati nilai 1 maka semakin buruk sebuah *cluster* dipisahkan dan *cluster* semakin tidak padu. Adapun parameter dari algoritma *Davies Bouldin Index* untuk melakukan pengelompokkan data dapat dilihat pada Tabel 6.

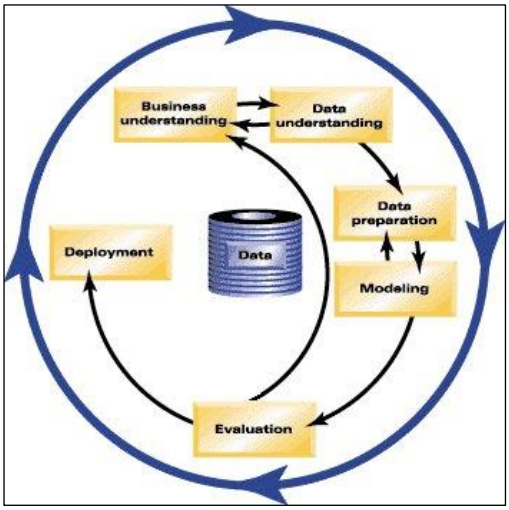
Tabel 6. Parameter *Davies Bouldin Index* [35]

Parameter	Keterangan
X	Array yang menyimpan data-data atribut yang digunakan untuk *clustering.*
Labels	Cluster yang didapatkan dari hasil *clustering* untuk setiap data.

2.12 *CRISP-DM*

Cross-Industry Standard Process for Data Mining atau CRISP-DM merupakan cara yang digunakan dalam dunia industri untuk mengarahkan dalam melakukan proyek yang berkaitan dengan *data mining* [36].

Proyek *data mining* yang menggunakan model CRISP-DM mempunyai 6 fase dalam siklus hidup, yaitu *Business Understanding* (Memahami Bisnis), *Data Understanding* (Memahami Data), *Data Preparation* (Mempersiapkan Data), *Modeling* (Memodelkan), *Evaluation* (Mengevaluasi), dan *Deployment*. Seluruh fase yang berurutan tersebut bersifat adaptif. Fase selanjutnya dalam urutan siklus bergantung pada keluaran dari fase sebelumnya. Model CRISP-DM bersifat fleksibel dan dapat disesuaikan dengan kebutuhan dalam menyelesaikan proyek [37].



Gambar 6. Siklus Model CRISP-DM [36]

Penjelasan dari setiap fase awal pada siklus model CRISP-DM adalah sebagai berikut [37]:

1. Fase *Business Understanding*

   Pada fase ini bertujuan untuk menentukan tujuan proyek, memastikan kebutuhan bisnis atau penelitian dari proyek secara detail, menyiapkan strategi awal untuk mencapai tujuan, dan membuat permasalahan *data mining* dari tujuan dan batasan proyek.
2. Fase *Data Understanding*

   Pada fase ini dilakukan pengumpulan data yang akan digunakan dalam melakukan proyek, melakukan analisis data agar mendapatkan informasi awal dan memahami mengenai data yang digunakan, serta melakukan evaluasi kualitas data. Setelah melakukan fase ini, dapat kembali ke fase sebelumnya untuk memastikan apakah data dapat yang telah dikumpulkan dapat digunakan untuk mencapai tujuan proyek.
3. Fase *Data Preparation*

   Pada fase ini bertujuan untuk menghasilkan data yang siap digunakan dalam fase *modeling*. Fase ini sangat penting karena memproses data awal menjadi data yang akan digunakan di seluruh fase selanjutnya. Pada fase ini dilakukan pemilihan variabel yang sesuai dengan kebutuhan untuk dilakukannya analisis dan melakukan perubahan pada variabel-variabel tertentu jika diperlukan.
4. Fase *Modeling*

   Pada fase ini dilakukan pemilihan dan pengaplikasian teknik *modeling* yang sesuai. Fase ini dapat kembali ke fase sebelumnya untuk memproses data agar sesuai dengan spesifikasi untuk menerapkan teknik *data mining* yang dipilih.
5. Fase *Evaluation*

   Pada fase ini bertujuan untuk mengevaluasi seluruh model yang digunakan dalam fase *modeling* untuk mengetahui model yang memiliki kualitas dan efektivitas terbaik. Fase ini juga untuk mengetahui apakah terdapat model yang mencapai tujuan yang ditetapkan, memastikan apakah terdapat masalah bisnis atau penelitian yang tidak diselesaikan dengan baik, dan membuat keputusan terkait penggunaan hasil *data mining*.

6. Fase *Deployment*

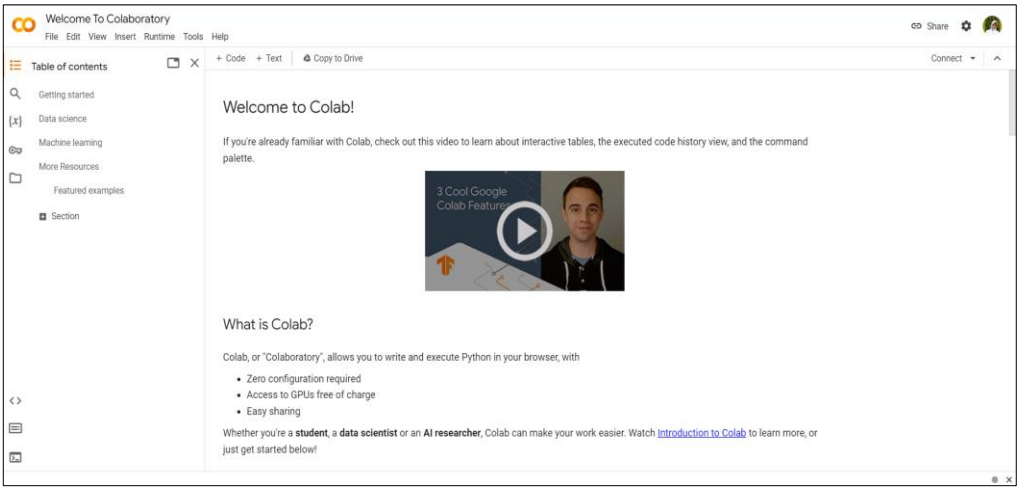
   Pada fase ini bertujuan untuk menyajikan data dalam bentuk yang dapat membuat hasil mudah dipahami dan digunakan oleh pengguna seperti membuat laporan.

2.13 Python

Python adalaha bahasa pemrogram tingkat tinggi karena kode programnya hampir mirip dengan bahasa manusia sehingga mudah untuk dipahami oleh manusia. Metode pemrosesan yang digunakan oleh python adalah *interpreted* membuat kode program dijalankan perbaris dan tidak memerlukan proses *compile*. Python dibuat pada awal 1990-an oleh Guido van Rossum di Scithchting Mathematisch Centrum (CWI) di Belanda. Seluruh versi dari python diluncurkan dalam bentuk *open source.* Keunggulan dari python adalah python telah memiliki banyak modul, *library,* dan *framework* membuat python cepat untuk digunakan karena hanya perlu mengimpor atau memodifikasi yang telah ada serta python bersifat fleksibel karena dapat digunakan untuk proyek kecil maupun besar, dapat digunakan secara *offline* maupun *online* dan dapat digunakan untuk proyek dengan GUI maupun tidak [38].

2.14 Google Colaboratory

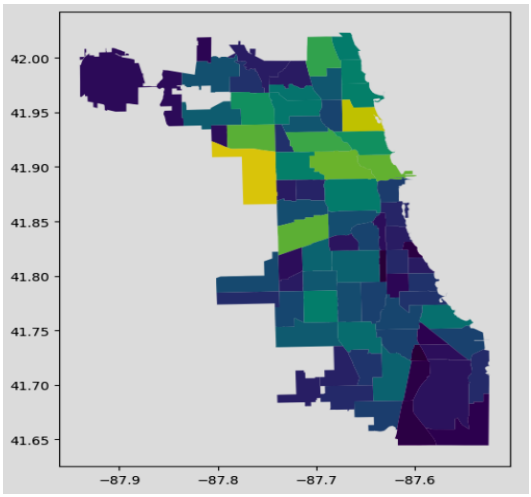
Google Colaboratory merupakan salah satu produk dari *Google Research*. Google Colaboratory tidak membutuhkan *setup* dan dapat digunakan secara gratis. Dengan menggunakan Google Colaboratory, memungkinkan siapapun untuk menulis dan menjalankan kode python melalui browser khususnya untuk *data science* dan *machine learning*. Google Colaboratory berbasis Jupyter Notebook sehingga dapat digunakan tanpa perlu mengunduh, menjalankan, atau menginstal apapun. Notebook dari Google Colaboratory disimpan di Google Drive atau di Github dengan format Jupyert Notebook (.ipynb). Notebook dapat dibagikan dengan mudah sama seperti membagikan Google Docs atau Google Sheets. Notebook dapat dijalankan dalam waktu hingga 12 jam dan *runtime* akan berhenti apabila tidak terdapat aktivitas pada notebook dalam jangka waktu tertentu [39].



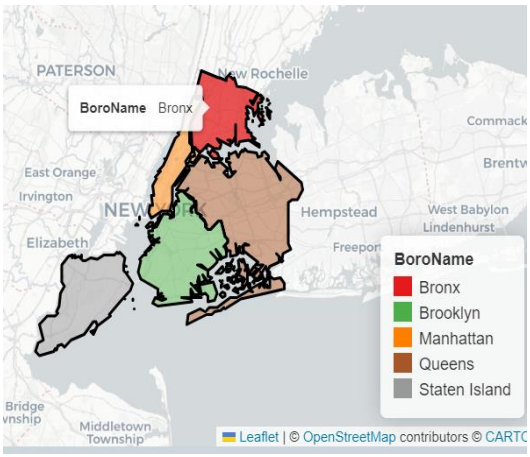
Gambar 7. Tampilan Google Colaboratory [40]

2.15 GeoPandas

GeoPandas adalah proyek *open source* yang dapat digunakan untuk mempermudah pengolahan data geospasial menggunakan python menjadi objek pandas dan menyediakan antarmuka tingkat tinggi untuk data spasial. GeoPandas membuat python dapat melakukan operasi spasial pada tipa data *geometry*. Operasi geometris dilakukan menggunakan library *shapely*. Untuk mengakses file dalam GeoPandas menggunakan library *fiona* dan library *matplotlib* untuk melakukan *plot* data. Terdapat dua jenis visualisasi yang dimiliki oleh GeoPandas, yaitu GeoDataFrame.plot dan GeoDataFrame.explore. Jenis GeoDataFrame.plot menghasilkan visualisasi data geospasial berupa *plot* menggunakan *matplotlib*. Sedangkan, GeoDataFrame.explore menampilkan peta iteraktif yang berdasarkan pada GeoPandas dan folium. GeoPandas mengimplementasikan GeoDataFrame dan GeoSeries yang merupakan *subclass* dari pandas.DataFrame dan pandas.Series. GeoPandas dapat digunakan secara gratis oleh siapapun [41].



Gambar 8. GeoPandas Plot [41]



Gambar 9. GeoPandas Explore [41]

2.16 Penelitian Terkait

Adapun lima penelitian terdahulu yang dijadikan sebagai referensi dalam melakukan penelitian, yaitu:

Penelitian yang dilakukan oleh Suwardi Annas, dkk. [42] bertujuan untuk memprediksi kelompok wilayah di Indonesia menggunakan algoritma *K-Means* dan memetakan daerah yang beresiko bencana alam berdasarkan jumlah bencana alam yang terjadi. Atribut yang digunakan adalah tanah longsor, banjir, banjir bandang, gempa bumi, gelombang pasang laut, puting beliung, erupsi gunung api, dan kebakaran hutan. Hasil yang diperoleh dari penelitian adalah *cluster* 1 (daerah yang rawan bencana banjir, puting beliung, dan tanah longsor) beranggota 61

daerah, *cluster* 2 (daerah rawan bencana banjir, dan tanah longsor) beranggota 4 daerah, *cluster* 3 (daerah rawan gempa bumi) beranggota 6 daerah, *cluster* 4 (daerah yang harus mewaspadai terjadinya banjir, puting beliung, dan kekeringan) beranggota 1 daerah, *cluster* 5 (daerah yang harus waspada terhadap banjir, tanah longsor, abrasi, dan kebakaran hutan/lahan) beranggota 9 daerah, *cluster* 6 (daerah yang rawan banjir dan tanah longsor) beranggota 1 daerah, dan *cluster* 7 (daerah yang rawan gempa bumi dan erupsi gunung api) beranggota 280 daerah. Hasil *clustering* divisualisasikan menggunakan *Geographical Information System* untuk memetakan daerah berdasarkan jumlah bencana rawan terhadap daerah tersebut.

Penelitian yang dilakukan oleh David Aryo Wicaksono, dkk. [43] bertujuan untuk mengelompokkan daerah rawan bencana alam di Provinsi Sumatera Barat menggunakan algoritma *K-Means* dan visualisasi menggunakan *library* GeoPandas. Atribut yang digunakan adalah bencana gempa bumi dan banjir. Metode *elbow* dan *Silhouette* untuk menentukan jumlah *cluster* yang optimal. Hasil dari penelitian adalah *cluster* 1 (tidak rawan bencana) memuat 7 daerah, *cluster* 2 (rawan bencana) memuat 12 daerah, dan *cluster* 3 (sangat rawan bencana) memuat 1 daerah. Setelah itu, hasil *clustering* divisualisasikan menggunkan *library* GeoPandas untuk memetakan daerah berdasarkan *cluster*-nya.

Penelitian yang dilakukan oleh Andri Dwi Noviandi, dkk. [44] bertujuan untuk menganalisis tingkat kedisiplinan dari warga Bekasi dalam melaksanakan protokol kesehatan Covid-19 menggunakan algoritma *K-Means*. Atribut yang digunakan adalah memakai masker, mencuci tangan, menjaga jarak, dan *hand sanitizer*. Evaluasi model menggunakan *Silhouette Coefficient*. Hasil clustering divisualisasikan menggunakan tools ArcGIS. Hasil penelitian adalah *cluster* 0 (disiplin) memuat 11 wilayah dan *cluster* 1 (tidak disiplin) memuat 6 wilayah. Kualitas dari *cluster* yang dihasilkan adalah baik karena memiliki nilai *Silhouette Coefficient* sebesar 0.926.

Penelitian yang dilakukan oleh Mhd Gading Sadewo, dkk. [45] bertujuan untuk mengelompokkan provinsi di Indonesia berdasarkan jumlah desa atau kelurahan yang memiliki upaya mitigasi bencana alam menggunakan algoritma *K-Means*.

Atribut yang digunakan adalah Sistem Peringatan Dini Tsunami, Perlengkapan Keselamatan, Sistem Peringatan Dini Bencana Alam, dan Jalur Evakuasi. Hasil penelitian adalah *cluster* 1 (tingkat antisipasi tinggi) memuat 3 provinsi, *cluster* 2 (tingkat antisipasi sedang) memuat 9 provinsi, dan *cluster* 3 (tingkat antisipasi rendah) memuat 22 provinsi.

Penelitian yang dilakukan oleh F N Dhewayani, dkk. [46] bertujuan untuk mengelompokkan daerah rawan bencana alam, yaitu kebakaran hutan dan lahan di provinsi Jawa Barat berdasarkan jumlah terjadinya kebakaran hutan dan lahan menggunakan algoritma *K-Means* untuk mengetahui potensi kebakaran dari setiap daerah. Metode *Silhouette* digunakan untuk menentukan jumlah *cluster* yang optimal. Atribut yang digunakan adalah jumlah kebakaran bangunan dan jumlah kebakaran lahan. Hasil *clustering*, yaitu *cluster* 1 (tingkat rawan sangat rendah) yang memuat 108 daerah, *cluster* 2 (tingkat rawan sedang) memuat 13 daerah, *cluster* 3 (tingkat rawan sangat tinggi) memuat 4 daerah, *cluster* 4 (tingkat rawan tinggi) memuat 2 daerah, dan *cluster* 5 (tingkat rawan rendah) memuat 8 daerah. Evaluasi menggunakan *silhouette coefficient* dari *clustering* adalah 0.75 yang berarti struktur dari *cluster* yang dihasilkan kuat.

Tabel 7. Penelitian Terkait

No.	Nama Peneliti	Data	Spesifikasi	Hasil
1.	Suwardi Annas, dkk [42]. (2019)	**Data bencana alam** yang terjadi di 362 **kabupaten di Indonesia** tahun 2016. **Data diperoleh dari Badan Nasional Penanggulangan Bencana.**	***K-Means****, RMSD, Geographical Information System (GIS)*	1. *Cluster* 1 (daerah yang rawan bencana banjir, puting beliung, dan tanah longsor) beranggota 61 daerah. 2. *Cluster* 2 (daerah rawan bencana banjir, dan tanah longsor) beranggota 4 daerah. 3. *Cluste*r 3 (daerah rawan gempa bumi)

No.	Nama Peneliti	Data	Spesifikasi	Hasil
				beranggota 6 daerah. 4. *Cluster* 4 (daerah yang harus mewaspadai terjadinya banjir, puting beliung, dan kekeringan) beranggota 1 daerah. 5. *Cluster* 5 (daerah yang harus waspada terhadap banjir, tanah longsor, abrasi, dan kebakaran hutan atau lahan) beranggota 9 daerah. 6. *Cluster* 6 (daerah yang rawan banjir dan tanah longsor) beranggota 1 daerah. 7. *Cluster* 7 (daerah yang rawan gempa bumi dan erupsi gunung api) beranggota 280 daerah.
2.	David Aryo Wicaksono, dkk [43]. (2023)	Data bencana alam bulanan untuk bencana banjir dan gempa bumi yang terjadi di Provinsi Sumatera Barat dari tahun 2018 sampai tahun 2021 yang	***K-Means,*** ***Elbow Method,*** *Silhouette coeffiecient,* ***GeoPandas***	1. *cluster* 1 (tidak rawan bencana) memuat 7 daerah. 2. *cluster* 2 (rawan bencana) memuat 12 daerah. 3. *cluster* 3 (sangat rawan bencana) memuat 1 daerah.

No.	Nama Peneliti	Data	Spesifikasi	Hasil
		diperoleh dari Badan Pusat Statistik. Lalu, data spasial kota atau kabupaten di Provinsi Sumatera Barat yang diperoleh dari Indonesia Geospasial.		
3.	Andri Dwi Noviandi, dkk [44]. (2021)	Data ketaatan warga Bekasi dalam melakukan protokol kesehatan Covid-19.	***K-Means, Silhouette Coefficient,*** *ArcGIS*	1. *Cluster* 0 (disiplin) sebanyak 11 wilayah. 2. *Cluster* 1 (tidak dfisiplin) sebanyak 6 wilayah. 3. *Silhouette Coefficient* sebesar 0.926.
4.	Mhd Gading Sadewo, dkk [45]. (2018)	Data jumlah desa atau kelurahan yang memiliki upaya mitigasi bencana alam berdasarkan provinsi. Jumlah provinsi yang digunakan adalah 34.	***K-Means***	1. *Cluster* 1 (tingkat antisipasi tinggi) memuat 3 provinsi. 2. *Cluster* 2 (tingkat antisipasi sedang) memuat 9 provinsi. 3. *Cluster* 3 (tingkat antisipasi rendah) memuat 22 provinsi.
5.	F N Dhewayani, dkk [46]. (2022)	Data bencana kebakaran hutan dan lahan yang terjadi di Provinsi Jawa Barat tahun 2019-2021	***K-Means, Silhouette coefficient***	1. *Cluster* 1 (tingkat rawan sangat rendah) yang memuat 108 daerah. 2. *Cluster* 2 (tingkat rawan sedang)

No.	Nama Peneliti	Data	Spesifikasi	Hasil
				memuat 13 daerah. 3. C*luster* 3 (tingkat rawan sangat tinggi) memuat 4 daerah. 4. *Cluster* 4 (tingkat rawan tinggi) memuat 2 daerah. 5. *Cluster* 5 (tingkat rawan rendah) memuat 8 daerah. 6. *Silhouette coefficient* adalah 0.75 yang berarti struktur dari *cluster* yang dihasilkan kuat.

Berdasarkan Tabel 7, dapat diketahui bahwa telah dilakukan penelitian yang membangun model untuk melakukan pengelompokkan daerah berupa 362 kabupaten di Indonesia, daerah di Provinsi Sumatera Barat, wilayah di Kota Bekasi, daerah di Provinsi Jawa Barat, dan Provinsi di Indonesia berdasarkan jumlah dari atribut yang ditentukan menggunakan algoritma *K-Means*. Sementara itu, penelitian yang akan dilakukan adalah mengelompokkan data rawan bencana di Indonesia berdasarkan jumlah bencana alam yang terjadi di kabupaten atau kota yang diperoleh dari BNPB dari tahun 2020 hingga tahun 2022 menggunakan algoritma *K-Means*. Perbedaan antara penelitian yang dilakukan Suwardi Annas dengan penelitian yang akan dilakukan, yaitu data bencana alam yang digunakan pada penelitian oleh Suwardi berasal dari 362 kabupaten, sedangkan pada penelitian ini berasal dari 494 kota atau kabupaten. Selain itu, penelitian oleh Suwardi menggunakan atribut tanah longsor, banjir, banjir bandang, gempa bumi, gelombang pasang laut, puting beliung, erupsi gunung api, dan kebakaran hutan sebagai parameter dalam pengelompokkan data bencana alam, sedangkan penelitian ini menggunakan atribut banjir, tanah longsor, angin puting beliung,

kebakaran hutan dan lahan, kekeringan, serta gelombang pasang atau abrasi. Penelitian ini menggunakan metode *Elbow* dalam menentukan jumlah *cluster* terbaik. Lalu, dari penelitian yang dilakukan untuk mengelompokkan daerah rawan kebakaran di Jawa Barat diketahui bahwa *silhouette coefficient* memberikan kemudahan dalam proses untuk mengevaluasi struktur *cluster*. Sehingga penelitian ini menggunakan teknik evaluasi *silhouette coefficient* dan *davies bouldin index* untuk mengevaluasi *cluster* yang dihasilkan serta hasil *clustering* divisualisasikan menggunakan GeoPandas.

III. METODE PENELITIAN

3.1 Waktu dan Tempat

Penelitian dan pembuatan tugas akhir dilakukan selama empat bulan dimulai dari bulan Oktober 2023 sampai dengan bulan Januari 2024 dan tempat melakukannya adalah Universitas Lampung.

Tabel 8. Jadwal Kegiatan Penelitian

No.	Nama Kegiatan	Waktu														
		Okt				Nov				Des				Jan		
		I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III
1.	Studi Literatur															
2.	Penyediaan Alat dan Bahan															
3.	*K-Means Clustering* dengan metode CRISP-DM															
4.	Analisis Hasil *Clustering* dengan metode CRISP-DM															
5.	Penyusunan Laporan															

3.2 Alat dan Bahan

3.2.1 Alat

Adapun alat berupa perangkat keras (*hardware*) dan perangkat lunak (*software*) yang digunakan pada penelitian ini sebagai berikut :

Tabel 9. Alat yang Digunakan

No.	Nama Alat	Spesifikasi	Deskripsi
1.	Laptop	Lenovo , processor Intel Core i5-1035G1, RAM 8 GB, SSD 512 GB, dan Windows 10	Perangkat keras yang digunakan selama melakukan penelitian.
2.	Google Colaboratory		Perangkat lunak berbasis web yang digunakan untuk menuliskan *syntax* dalam bahasa pemrograman python.
4.	Python	Python 3.10.11	Bahasa pemrograman yang digunakan dalam melakukan fase pemahaman data hingga fase evaluasi dalam metode CRISP-DM.
5.	Machine Learning Library		Library yang dibutuhkan dalam mengolah data, melakukan eksplorasi data, membangun model, dan memvisualisasikan hasil dari *clustering*.
	Numpy	*numpy* 1.23.5	Library yang digunakan untuk mencari indeks dari value yang memenuhi kondisi dengan fungsi *where*().
	matplotlib.pyplot		Library yang digunakan untuk melakukan visualisasi data menggunakan piechart.
	Seaborn	*seaborn* 0.12.2	Library yang digunakan untuk melakukan visualisasi data menggunakan diagram *countplot*.
	matplotlib.dates		*Library* yang digunakan untuk memformat agar menampilkan tanggal dalam bentuk bulan

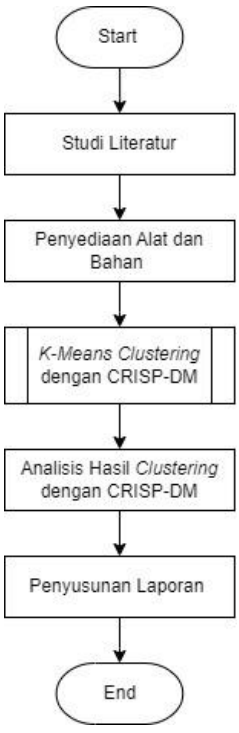
No.	Nama Alat	Spesifikasi	Deskripsi
			dan tahun.
	MinMaxScaler		Library yang digunakan untuk mernormalisasikan data sehingga memiliki nilai dengan rentang *value* 0-1.
	KMeans		Library yang digunakan untuk membangun model *clustering* dengan algoritma *K-Means*.
	silhouette_score		Library yang digunakan untuk menggunakan teknik evaluasi *silhouette score.*
	davies bouldin index		Library yang digunakan untuk menggunakan teknik evaluasi *davies bouldin index.*
	Geopandas	*Geopandas 0.10.2*	Library yang digunakan untuk memvisualisasi data geografis menjadi objek *pandas*.

3.2.2 Bahan

Bahan yang digunakan dalam penelitian adalah data bencana alam yang terjadi di Indonesia dari bulan Januari 2020 hingga bulan Desember tahun 2022 yang diperoleh dari situs Geoportal Data Bencana Indonesia BNPB dan data spasial perbatasan administrasi kota atau kabupaten di Indonesia yang didapatkan dari situs Indonesia Geospasial. Data bencana alam yang digunakan berjumlah 14.785 kasus. Atribut yang terdapat dalam dataset bencana alam, yaitu Kode Identitas Bencana, ID Kabupaten, Tanggal Kejadian, Kejadian, Kabupaten, Provinsi, Fasum Rusak, Rumah Rusak, Rumah Terendam, Terluka, Meninggal, Hilang, dan Penyebab. Jumlah baris data yang terdapat dalam dataset data spasial perbatasan administrasi kota atau kabupaten di Indonesia adalah 515 record. Atribut dalam dataset data spasial, yaitu KAB_KOTA dan *geometry* yang berisi lokasi suatu daerah di permukaan bumi.

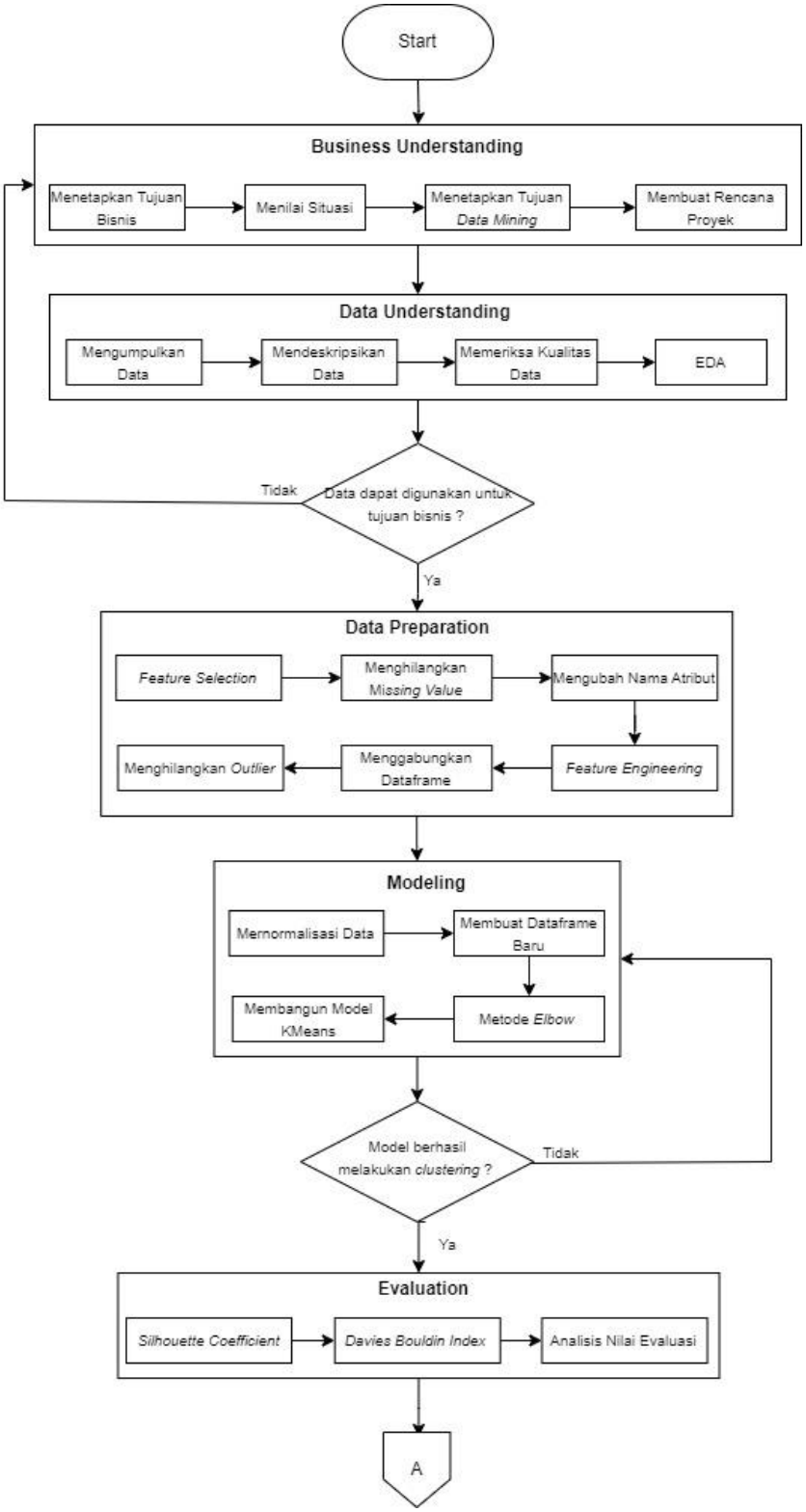
3.3 Tahapan Penelitian

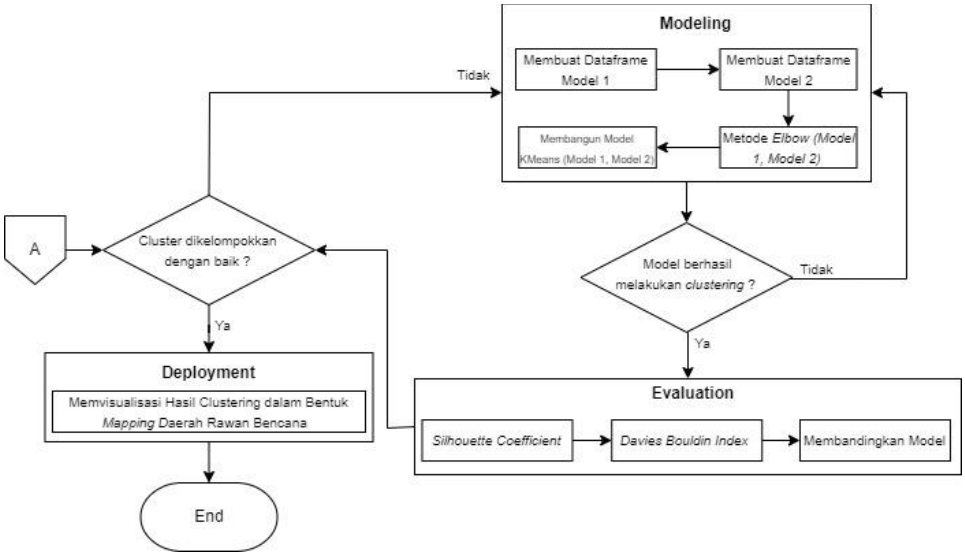
Adapun tahapan pada penelitian ini diawali dengan melakukan studi literatur guna mempelajari ilmu dan penelitian terkait yang telah dilakukan sebelumnya yang bersumber dari buku, jurnal, artikel, dan prosiding. Ilmu yang dipelajari pada tahapan studi literatur untuk menunjang penelitian ini, yaitu *data mining*, *clustering*, algoritma *K-Means*, teknik evaluasi *Sillhouette Coefficient* dan *davies bouldin index*, *metode Cross Industry Standard Process for Data Mining* (*CRISP-DM*), serta visualisasi menggunakan GeoPandas. Lalu, menyediakan alat dan bahan yang digunakan dalam melakukan penelitian. Kemudian, *clustering* data menggunakan metode pengembangan model C*ross Industry Standard Process for Data Mining* (CRISP-DM) yang terdiri atas fase pemahaman bisnis (*business understanding phase)*, fase pemahaman data (*data understanding phase)*, fase persiapan data (*data preparation phase)*, fase pemodelan (*modeling phase)*, fase evaluasi (*evaluation phase*), dan fase penyebaran (*deployment phase*). Setelah itu, menganalisis hasil *clustering* data. Kemudian, menyusun laporan penelitian yang memuat hasil dari penelitian. Tahapan penelitian yang dilakukan dapat dilihat pada diagram alir sebagai berikut.



Gambar 10. Diagram alir tahapan penelitian

Adapun tahapan dari pengembangan model *clustering* menggunakan metode *Cross Industry Standard Process for Data Mining* (CRISP-DM) dapat dilihat pada gambar sebagai berikut.





Gambar 11. *Flowchart Clustering* Data dengan Menggunakan Model Pengembangan CRISP-DM

3.3.1 Fase Pemahaman Bisnis (*Business Understanding Phase*)

Pada fase pemahaman bisnis melakukan beberapa hal yang berfokus untuk memahami tujuan dan persyaratan dari penelitian, antara lain menetapkan tujuan bisnis, menilai situasi, menetapkan tujuan data *mining*, dan membuat rencana proyek. Tujuan bisnis dari dilakukannya penelitian ini adalah untuk mengetahui kota atau kabupaten yang rawan bencana di Indonesia. Hal ini diharapkan dapat membantu pemangku kepentingan dalam mengidentifikasi daerah rawan bencana di Indonesia. Situasi yang sedang berlangsung di Indonesia adalah Indonesia terletak di wilayah cincin api pasifik dan beriklim tropis menyebabkan Indonesia rawan terhadap bencana geologi seperti gempa bumi, tsunami, dan letusan gunung api serta rawan terhadap bencana hidrometeorologi, seperti banjir, angin puting beliung, dan kekeringan. Bencana alam memberikan dampak bagi manusia dan lingkungan. Dampak yang ditimbulkan dapat dikurangi dengan melakukan upaya mitigasi. Dalam mengatur strategi mitigasi membutuhkan informasi mengenai daerah rawan bencana agar upaya mitigasi yang dilakukan tepat dan efektif. Lalu, tujuan *data mining* adalah melakukan pengelompokkan kabupaten atau kota menjadi beberapa *cluster* berdasarkan parameter yang digunakan dalam data bencana guna mengetahui daerah rawan di Indonesia. Oleh karena itu, rencana

proyek yang akan dilaksanakan adalah mengelompokkan data rawan bencana di di berbagai kota dan kabupaten di Indonesia menggunakan algoritma *K-Means clustering*.

3.3.2 Fase Pemahaman Data (*Data Understanding Phase*)

Fase pemahaman data dimulai dari melakukan pengumpulan data yang dibutuhkan dalam melakukan penelitian, mendeskripsikan data, memeriksa kualitas dari data, dan melakukan eksplorasi data. Data yang digunakan diperoleh dari situs Geoportal Data Bencana Indonesia Badan Nasional Penanggulangan Bencana (BNPB), yaitu data bencana alam yang terjadi di Indonesia dari tanggal 1 Januari tahun 2020 sampai dengan tanggal 31 Desember 2022 dan data spasial perbatasan administrasi kota atau kabupaten di Indonesia. Langkah mendeskripsikan data dilakukan dengan memeriksa atribut-atribut, tipe data dari setiap data yang terdapat dalam dataset dan banyak data. Setelah itu, mengeksplorasi data dalam dataset dengan menggunakan visualisasi untuk memahami data, informasi yang belum diketahui sebelumnya, dan hal yang berpotensi menjadi masalah. Kemudian, memeriksa kualitas data dengan memeriksa nilai yang hilang dalam data secara keseluruhan.

3.3.3 Fase Persiapan Data (*Data Preparation Phase*)

Fase persiapan data dimulai dengan melakukan pemilihan atribut yang akan digunakan untuk fase pemodelan dan yang relevan dengan penelitian, menghilangkan *missing value* yang terdapat dalam data, mengubah nama atribut, membuat atribut baru berdasarkan atribut yang telah ada (*feature engineering*), menggabungkan (*merge*) enam dataframe setiap jenis bencana agar menjadi dataframe final, dan menghilangkan *outlier* yang terdapat dalam dataframe final. Fase persiapan data dilakukan menggunakan software Google Colaboratory dan bahasa pemrograman Python dengan library *pandas* dan *numpy*.

3.3.4 Fase Pemodelan (*Modeling Phase*)

Fase pemodelan dimulai dengan melakukan normalisasi data menggunakan *library* M*inMaxScaler* sehingga nilai dari data memiliki range 0 sampai dengan 1. Selanjutnya, memilih teknik data mining untuk melakukan pengelompokkan data dan mengimplementasikan teknik tersebut. Teknik yang digunakan adalah algoritma *K-Means* untuk mengkluster data rawan bencana di Indonesia. Untuk mengetahui jumlah *cluster* yang terbaik digunakan metode *elbow* sehingga membangun model menggunakan jumlah *cluster* yang optimal. Pemodelan dilakukan menggunakan atribut banjir, tanah_longsor, puting_beliung, kebakaran_hutan_lahan, gelombang_pa, dan kekeringan. Parameter *K-Means* yang digunakan dalam membangun model, yaitu n_clusters, init, dan n_init. Atribut untuk pemodelan model pertama adalah banjir, tanah_longsor, puting_beliung, kebakaran_hutan_lahan, dan kekeringan. Pemodelan untuk model kedua menggunakan atribut banjir, tanah_longsor, puting_beliung, kebakaran_hutan_lahan, dan gelombang_pa. Persamaan rumus yang digunakan pada *library MinMaxScaler* sebagai berikut [47].

$$X\_std = [(X - X.min) / (X.max - X.min)] \quad (2.11)$$

Keterangan :

X_std = nilai hasil normalisasi

X = nilai yang akan dinormalisasi

X.min = nilai terendah

X.max = nilai tertinggi

Parameter n_clusters memuat jumlah *cluster* yang akan dibentuk yang diperoleh dari metode *elbow.* Lalu, parameter init untuk menentukan metode inisialisasi *centroid* awal. Proses awal pada metode *k-means*++ adalah memilih secara acak objek pada data untuk dijadikan sebagai *centroid*. Lalu, menghitung jarak setiap objek ke *centroid* yang telah dipilih. Selanjutnya, memilih objek yang memiliki jarak terjauh dari *centroid* terdekat untuk dijadikan sebagai *centroid* baru. Hal ini dilakukan agar kemungkinan dalam menentukan objek sebagai *centroid* berbanding lurus dengan jaraknya dari *centroid* paling dekat. Jumlah dari n_init

bergantung pada jenis metode init yang digunakan. Untuk metode *k-means++* maka n_init adalah 1. Jumlah iterasi maksimal yang dilakukan tidak dibatasi, tetapi berhenti secara otomatis sampai dengan nilai *centroid* dari *cluster* tidak berubah. *Software* yang digunakan untuk melakukan pengklasteran pada fase ini adalah Google Colaboratory dan *library MinMaxScaler* dan *KMeans*.

3.3.5 Fase Evaluasi (*Evaluation Phase*)

Fase evaluasi dilakukan dengan tujuan untuk mengetahui tingkat kekohesifan antara objek dengan *centroid* dalam *cluster*. Teknik evaluasi yang digunakan adalah *silhouette coefficient* dan *davies bouldin index*. Parameter yang digunakan untuk kedua teknik evaluasi adalah x dan labels. X merupakan *array* yang memuat data yang digunakan saat *clustering* dan *labels* merupakan *cluster* dari setiap data. Setelah itu, menganalisis nilai evaluasi. Apabila hasil dari analisis diketahui bahwa *cluster* tidak dikelompokkan dengan baik, maka kembali ke fase pemodelan untuk membangun dua model agar dapat dibandingkan. Lalu, mengevaluasi kedua model dan membandingkan nilai evaluasi dari model pertama dengan model kedua untuk mengetahui model terbaik. Selain itu, fase ini juga memeriksa kembali seluruh tahapan yang telah dilakukan untuk memastikan tidak ada tahapan yang terlewatkan dan telah dilakukan dengan benar sehingga dapat menentukan untuk lanjut ke fase selanjutnya, yaitu fase penyebaran atau kembali ke fase pemodelan.

3.3.6 Fase Penyebaran (*Deployment Phase*)

Pada fase ini hasil *clustering* data rawan bencana di Indonesia dituangkan dalam bentuk *mapping* daerah berdasarkan *cluster* menggunakan GeoPandas sehingga lebih mudah untuk dipahami dan mudah untuk disebarkan. Analisis *mapping* hasil *clustering* menghasilkan *insight* terkait kabupaten atau kota yang rawan terhadap bencana alam di Indonesia dan jumlah terjadinya bencana alam di setiap kabupaten. *Insight* tersebut diharapkan dapat menjadi bahan pertimbangan bagi pemerintah di daerah rawan bencana untuk menentukan upaya mitigasi yang akan dilakukan.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Adapun kesimpulan yang diperoleh berdasarkan hasil *clustering* data rawan bencana di Indonesia menggunakan algoritma *K-Means*, sebagai berikut :

1. Berdasarkan hasil *clustering* data rawan bencana di Indonesia, terbentuk lima *cluster* dari model dengan atribut banjir, tanah longsor, angin puting beliung, kebakaran hutan dan lahan, serta kekeringan. *Cluster* 0 memiliki tingkat kerawanan rendah terhadap seluruh jenis bencana yang menjadi atribut. *Cluster* 1 memiliki tingkat kerawanan tinggi pada bencana banjir, tanah longsor, dan angin puting beliung. Lalu, memiliki tingkat kerawanan rendah pada kebakaran hutan dan lahan serta kekeringan. *Cluster* 2 mempunyai tingkat kerawanan sedang untuk bencana banjir dan kebakaran hutan dan lahan serta tingkat kerawanan rendah terhadap bencana tanah longsor, angin puting beliung, dan kekeringan. C*luster* 3 memiliki tingkat kerawanan sedang pada bencana banjir dan tingkat kerawanan rendah pada bencanan tanah longsor, angin puting beliung, kebakaran hutan dan lahan, serta kekeringan. *Cluster* 4 memiliki tingkat kerawanan tinggi pada bencana tanah longsor dan angin puting beliung, tingkat kerawanan sedang pada bencana banjir, dan tingkat kerawanan rendah pada bencana kebakaran hutan dan lahan serta kekeringan.
2. *Cluster* 0 terdiri atas 392 daerah yang tersebar di seluruh pulau di Indonesia dan paling banyak terdapat di Provinsi Sumatera Utara, yaitu 6 kota dan 24 kabupaten. *Cluster* 1 hanya terdiri atas dua kabupaten, yaitu Kabupaten Bogor dan Cilacap. Daerah yang termasuk dalam *cluster* 2 berjumlah 64 yang tersebar di pulau Sumatera, Jawa, Kalimantan dan Sulawesi dimana paling banyak terdapat di Provinsi Aceh, yaitu 1 kota dan 10 kabupaten. *Cluster* 3 terdiri atas 25 daerah yang tersebar di Pulau Sumatera, Jawa, Bali, Nusa Tenggara Barat,

dan Nusa Tenggara Timur dimana paling banyak terdapat di Provinsi Nusa Tenggara Barat, yaitu 1 kota dan 8 kabupaten. Cluster 4 terdiri atas 11 daerah yang tersebar di Pulau Sumatera dan Jawa dan paling banyak terdapat di Provinsi Jawa Barat, yaitu 1 kota dan 7 kabupaten.

3. *Clustering* data rawan bencana alam di Indonesia menjadi lima *cluster* kerawanan bencana menggunakan algoritma *data mining* K-Means dilakukan dengan mengelompokkan data jumlah terjadinya bencana banjir, tanah longsor, angin puting beliung, kebakaran hutan dan lahan, kekeringan serta gelombang pasang atau abrasi di setiap daerah dan membentuk *cluster* dengan jumlah yang optimal.
4. Berdasarkan hasil evaluasi model menggunakan atribut banjir, tanah longsor, angin puting beliung, kebakaran hutan dan lahan, serta kekeringan, *elbow method* dapat menentukan jumlah *cluster* yang optimal. Hal ini dibuktikan dengan tingkat kekohesifan antara objek dengan *centroid* terbaik terdapat di *cluster* yang dihasilkan oleh model dengan atribut kekeringan dimana nilai *silhouette coefficient* sebesar 0,9 dan nilai *davies bouldin index* sebesar 0,39.
5. *Mapping* data rawan bencana di Indonesia berhasil dibangun menggunakan *library* GeoPandas berdasarkan hasil *clustering* dari model menggunakan atribut banjir, tanah longsor, angin puting beliung, kebakaran hutan dan lahan, serta kekeringan dengan tingkat kecakupan wilayah yang dibangun adalah 96% dari seluruh wilayah di Indonesia.

5.2 Saran

Adapun saran untuk pengembangan penelitian selanjutnya berdasarkan penelitian ini, sebagai berikut :

1. Membangun model menggunakan algoritma *clustering* yang lain dalam melakukan penelitian terkait pengelompokkan data rawan bencana di Indonesia agar dapat mengetahui algoritma yang menghasilkan *cluster* dengan kekohesifan antara objek dengan *centroid* terbaik.
2. Menggunakan metode pendeteksi *outlier* lain, seperti *Gaussian-based methods* atau *Regression-based methods* untuk mendeteksi *outlier* berdasarkan distribusi

dari data agar mengetahui metode terbaik dalam mendeteksi *outlier* dan dihasilkan model dengan *cluster* yang memiliki kepaduan paling baik.

DAFTAR PUSTAKA

[1] BNPB, “Potensi Ancaman Bencana,” bnpb.go.id. Accessed: Nov. 23, 2023. [Online]. Available: https://www.bnpb.go.id/potensi-ancaman-bencana

[2] BNPB, “Infografis Bencana Tahun 2022,” bnpb.go.id. Accessed: Nov. 25, 2023. [Online]. Available: https://www.bnpb.go.id/infografis/infografis-bencana-tahun-2022

[3] Pemerintah Pusat, “Undang-undang (UU) Nomor 24 Tahun 2007 tentang Penanggulangan Bencana,” Database Peraturan | JDIH BPK. Accessed: Nov. 19, 2023. [Online]. Available: http://peraturan.bpk.go.id/Details/39901/Uu-No-24-Tahun-2007

[4] dibi.bnpb.go.id, “Data Informasi Bencana Indonesia (DIBI).” Accessed: Nov. 25, 2023. [Online]. Available: https://dibi.bnpb.go.id/

[5] T. Yanuarto, S. Pinuji, A. C. Utomo, and I. T. Satrio, *Buku Saku Tanggap Tangkas Tangguh Menghadapi Bencana*. Jakarta Timur: Pusat Data Informasi dan Humas BNPB, 2019.

[6] Ima Nurmalia Permatasari, “Kajian Resiko, Dampak, Kerentanan dan Mitigasi Bencana Abrasi Dibeberapa Pesisir Indonesia,” *J. Ris. Kelaut. Trop. J. Trop. Mar. Res. J-Trop.*, vol. 3, no. 1, p. 56, Apr. 2021, doi: 10.30649/jrkt.v3i1.56.

[7] BNPB, “Sejarah BNPB,” bnpb.go.id. Accessed: Jan. 07, 2024. [Online]. Available: https://www.bnpb.go.id/sejarah-bnpb

[8] BNPB, “Tugas dan Fungsi BNPB,” bnpb.go.id. Accessed: Jan. 07, 2024. [Online]. Available: https://www.bnpb.go.id/tugas-dan-fungsi-bnpb

[9] Badan Nasional Penanggulangan Bencana, “Geoportal Data Bencana Indonesia,” gis.bnpb.go.id. Accessed: Nov. 17, 2023. [Online]. Available: https://gis.bnpb.go.id/

[10] S. Triutomo and T. W. Sudinda, *Indeks Rawan Bencana Indonesia*. Jakarta: Direktorat Pengurangan Risiko Bencana, 2011.

[11] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Yogyakarta: Penerbit Andi, 2020.

[12] A. A. Permana, S. Wahyuddin, L. W. Santoso, A. K. Wardhani, Rahmaddeni, A. J. Wahidin, G. E. Yuliastuti, Elisawati, R. R. Wijayanti, and Abdurrasyid., *Machine Learning*. Padang: PT Global Eksekutif Teknologi, 2023.

[13] G. N. Elwirehardja, T. Suparyanto, and B. Pardamean, *Pengenalan Konsep Machine Learning Untuk Pemula*. Yogyakarta: INSTIPER PRESS, 2023.

[14] A. Jeffares, “Supervised vs Unsupervised Learning in 2 Minutes,” Medium. Accessed: Jan. 18, 2024. [Online]. Available: https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

[15] S. Santoso, *Statistik Multivariat*. Jakarta: Elex Media Komputindo, 2010.

[16] R. Sibarani and C. Chafid, “Algorithma K-Means Clustering Strategi Pemasaran Penerimaan Mahasiswa Baru Universitas Satya Negara Indonesia [Algorithma k-Means Clustering Strategy Marketing Admission Universitas Satya Negara Indonesia],” *Pros. Semin. Nas. CENDEKIAWAN*, pp. 685–690, Oct. 2018, doi: 10.25105/semnas.v0i0.3512.

[17] H. Malikhatin, A. Rusgiyono, and D. A. I. Maruddani, “Penerapan K-Modes Clustering dengan Validasi Dunn Index pada Pengelompokkan Karakteristik Calon TKI Menggunakan R-GUI,” *J. Gaussian*, vol. 10, no. 3, Art. no. 3, Dec. 2021, doi: 10.14710/j.gauss.10.3.359-366.

[18] A. P. Riani, A. Voutama, and T. Ridwan, “Penerapan K-Means Clustering dalam Pengelompokan Hasil Belajar Peserta Didik dengan Metode Elbow,” *J-SISKO TECH J. Teknol. Sist. Inf. Dan Sist. Komput. TGD*, vol. 6, no. 1, p. 164, Jan. 2023, doi: 10.53513/jsk.v6i1.7351.

[19] J. O. Ong, “Implementasi Algoritma K-Means Clustering untuk Menentukan Strategi Marketing President University,” *J. Ilm. Tek. Ind.*, vol. 12, no. 1, Art. no. 1, Jun. 2013, doi: https://doi.org/10.23917/jiti.v12i1.651.

[20] E. H. Wisanta and Y. N. Marlim, “Analisis Algoritma K-Mens untuk Clustering Kepuasan Pelayanan: Mall Pelayanan Publik Pekanbaru,” *Semin. Nas. Inform. SENATIKA*, pp. 222–228, Jun. 2021.

[21] S. Andayani, “Pembentukan cluster dalam Knowledge Discovery in Database dengan Algoritma K-Means,” *SEMNAS Mat. Dan Pendidik. Mat.*, Dec. 2007, [Online]. Available: https://staffnew.uny.ac.id/upload/132162018/penelitian/Pembentukan+cluster+dlm+KDD+dgn+Algoritma+kmeans.pdf

[22] Scikit Learn, “sklearn.cluster.KMeans,” scikit-learn. Accessed: Oct. 13, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[23] E. Wahyuni and Suparman, “A Comparison of Outlier Detection Techniques in Data Mining,” *Proc. 1st STEEEM 2019*, vol. 1, no. 1, pp. 139–147, 2019.

[24] Scikit Learn, “Novelty and Outlier Detection,” scikit-learn. Accessed: Oct. 14, 2023. [Online]. Available: https://scikit-learn/stable/modules/outlier_detection.html

[25] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, “An Outliers Detection and Elimination Framework in Classification Task of Data Mining,” *Decis. Anal. J.*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.

[26] B. Priya. C, “How to Detect Outliers in Machine Learning – 4 Methods for Outlier Detection,” freeCodeCamp.org. Accessed: Nov. 20, 2023. [Online]. Available: https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/

[27] K. Babitz, “Introduction to K-Means Clustering with scikit-learn in Python,” datacamp.com. Accessed: Jan. 07, 2024. [Online]. Available: https://www.datacamp.com/tutorial/k-means-clustering-python

[28] A. Gupta, “Elbow Method for Optimal Value of K in KMeans,” geeksforgeeks.org. Accessed: Dec. 07, 2023. [Online]. Available: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

[29] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method for Identification of The Best Customer Profile Cluster,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[30] R. Buyya, R. N. Calheiros, and A. V. Dastjerdi, *Big Data Principles and Paradigms*. Elsevier, 2016. doi: 10.1016/C2015-0-04136-3.

[31] D. W. D. Rahmawati, I. Cholisoddin, and N. Santoso, “Optimasi K-Means untuk Pengelompokan Data Kinerja Akademik Dosen menggunakan Particle Swarm Optimization,” *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 3, no. 4, Art. no. 4, Feb. 2019.

[32] Scikit Learn, “sklearn.metrics.silhouette_score,” scikit-learn. Accessed: Oct. 15, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.metrics.silhouette_score.html

[33] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. in Wiley series in probability and mathematical statistics. Hoboken, N.J: Wiley, 2005.

[34] M. Orisa, “Optimasi Cluster pada Algoritma K-Means,” *Pros. SENIATI*, vol. 6, no. 2, pp. 430–437, Jul. 2022, doi: 10.36040/seniati.v6i2.5034.

[35] Scikit Learn, “sklearn.metrics.davies_bouldin_score,” scikit-learn. Accessed: Oct. 16, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[36] IBM, “CRISP-DM Help Overview,” ibm.com. Accessed: Nov. 15, 2023. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

[37] Kusrini and E. taufiq luthfi, *Algoritma Data Mining*. Yogyakarta: Penerbit Andi, 2009.

[38] F. Sembiring, *Buku Ajar Dasar Pemrograman (Python)*. Jawa Barat: Nusa Putra Press, 2021.

[39] Google Research, “Google Colaboratory,” research.google.com. Accessed: Dec. 03, 2023. [Online]. Available: https://research.google.com/colaboratory/faq.html

[40] Google Research, “Google Colaboratory,” colab.research.google.com. Accessed: Dec. 03, 2023. [Online]. Available: https://colab.research.google.com/

[41] GeoPandas, “About GeoPandas,” geopandas.org. Accessed: Dec. 03, 2023. [Online]. Available: https://geopandas.org/en/stable/about.html

[42] S. Annas and Z. Rais, “K-Means and GIS for Mapping Natural Disaster Prone Areas in Indonesia,” in *Proceedings of the Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, 12 October 2019, Bandung, West Java, Indonesia*, Bandung, Indonesia: EAI, 2020. doi: 10.4108/eai.12-10-2019.2296336.

[43] D. A. Wicaksono and Y. A. Susetyo, “Clustering Zonasi Daerah Rawan Bencana Alam di Provinsi Sumatera Barat Menggunakan Algoritma K-

Means dan Library GeoPandas," *J. Indones. Manaj. Inform. Dan Komun.*, vol. 4, no. 2, Art. no. 2, May 2023, doi: 10.35870/jimik.v4i2.225.

[44] A. D. Noviandi, T. N. Padillah, and Y. Umaidah, "Clustering Tingkat Kedisiplinan Warga Bekasi dalam Menjalankan Protokol Kesehatan di Masa Pandemi Covid-19 dengan Algoritme K-Means," vol. 7, no. 4, pp. 681–688, Aug. 2021, doi: 10.5281/ZENODO.5336446.

[45] M. G. Sadewo, A. P. Windarto, and A. Wanto, "Penerapan Algoritma Clustering Dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/ Mitigasi Bencana Alam Menurut Provinsi dengan K-Means," *KOMIK Konf. Nas. Teknol. Inf. Dan Komput.*, vol. 2, no. 1, Art. no. 1, Oct. 2018, doi: 10.30865/komik.v2i1.943.

[46] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," *J. Teknol. Dan Inf.*, vol. 12, no. 1, pp. 64–77, Mar. 2022, doi: 10.34010/jati.v12i1.6674.

[47] "sklearn.preprocessing.MinMaxScaler," scikit-learn. Accessed: Jan. 21, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[48] N. Muhamad, "WRI 2022: Indonesia Negara Paling Rawan Bencana Kedua di Dunia | Databoks," databoks.katadata.co.id. Accessed: Dec. 03, 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/10/17/wri-2022-indonesia-negara-paling-rawan-bencana-kedua-di-dunia

[49] Indonesia Geospasial Tech, "Batas Administrasi Provinsi, Kabupaten/Kota, Kecamatan, Desa Tahun 2019," indonesia-geospasial.com. Accessed: Nov. 27, 2023. [Online]. Available: https://www.indonesia-geospasial.com/2020/04/download-shapefile-shp-batas.html

[50] T. I. Hermanto and Y. Muhyidin, “Analisis Sebaran Titik Rawan Bencana dengan K-Means Clustering dalam Penanganan Bencana,” *J. Sains Komput. Inform. J-SAKTI*, vol. 5, pp. 406–416, Mar. 2021.

[51] M. T. Furqon and L. Muflikhah, “Clustering The Potential Risk of Tsunami Using Density-Based Spatial Clustering of Application with Noise (DBSCAN),” *J. Environ. Eng. Sustain. Technol.*, vol. 3, no. 1, Art. no. 1, May 2016, doi: 10.21776/ub.jeest.2016.003.01.1.