SIGMA
JURNAL TEKNOLOGI PELITA BANGSA

Vol. 14 No. 4 Desember 2023 p-ISSN: 2407-3903 e-ISSN: 2407-3903

Diterima, 15 September 2023 | Direvisi, 15 November 2023 | Dipublikasikan, 20 Desember 2023

PENERAPAN METODE *CLUSTERING* DALAM PENGELOMPOKKAN DATA CURAH HUJAN DENGAN ALGORITMA *K-MEANS*

Arif Susilo[1], Absul Wahab[2]

[1,2]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]arif.susilo@pelitabangsa.ac.id, [2]abdulwahab@gmail.com

Abstrak

Curah hujan yang tinggi dapat menyebabkan bencana alam seperti banjir, longsor, dan banjir bandang. Fenomena ini semakin sering terjadi dan semakin parah dalam beberapa tahun terakhir, dan Indonesia sebagai negara tropis sangat rentan terhadap bencana alam yang disebabkan oleh cuaca ekstrem, Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) mencatat tingginya angka bencana hidrometeorologi tidak lepas dari perubahan iklim. Koordinator Bidang Analisis Variabilitas Iklim BMKG mengatakan bahwa hasil monitoring BMKG dalam 40 tahun terakhir mengindikasikan curah hujan ekstrem di Indonesia mengalami kecenderungan peningkatan, baik dalam hal frekuensi maupun intensitas. Penyebab dari bencana akibat curah hujan adalah perubahan iklim, pertumbuhan populasi, dan urbanisasi yang meningkatkan risiko bencana alam. Perubahan iklim menyebabkan curah hujan menjadi lebih tidak teratur dan intens. Pertumbuhan populasi menyebabkan peningkatan permukiman di daerah yang rentan terhadap bencana, sehingga risiko bencana semakin tinggi. Urbanisasi juga menyebabkan pengurangan luas daerah alami yang dapat menyerap air hujan, sehingga terjadilah banjir dan longsor. Menurut data yang di dapat dari situs Data Informasi Bencana Indonesia (https://dibi.bnpb.go.id/) terdapat 2690 banjir, 2228 tanah longsor, dan 272 kekeringan yang terjadi di Indonesia dalam kurun waktu 2017 hingga 2018, dan tanah longsor merupakan bencana yang paling banyak terjadi, pada tahun 2020 tercatat telah terjadi bencana tanah longsor di Jawa Barat sebanyak 3,232 kali.

Kata Kunci: Calstering, Algoritma, Curah Hujan, Kemeans.

Abstract

High rainfall can cause natural disasters such as floods, landslides and flash floods. This phenomenon has become more frequent and more severe in recent years, and Indonesia as a tropical country is very vulnerable to natural disasters caused by extreme weather. The Meteorology, Climatology and Geophysics Agency (BMKG) notes that the high number of hydrometeorological disasters cannot be separated from climate change. The BMKG Climate Variability Analysis Coordinator said that BMKG monitoring results in the last 40 years indicated that extreme rainfall in Indonesia had an increasing trend, both in terms of frequency and intensity. The causes of disasters due to rainfall are climate change, population growth and urbanization which increase the risk of natural disasters. Climate change causes rainfall to become more irregular and intense. Population growth causes an increase in settlements in areas that are vulnerable to disasters, so that the risk of disasters becomes higher. Urbanization also causes a reduction in the area of natural areas that can absorb rainwater, resulting in floods and landslides. According to data obtained from the Indonesian Disaster Information Data site (https://dibi.bnpb.go.id/) there were 2690 floods, 2228 landslides and 272 droughts that occurred in Indonesia in the period 2017 to 2018, and landslides is the most common disaster, in 2020 there were 3,232 landslides recorded in West Java.

Keywords: Calstering, Algorithm, Rainfall, Kemeans.

1. Pendahuluan

Curah hujan yang tinggi dapat menyebabkan bencana alam seperti banjir, longsor, dan banjir bandang. Fenomena ini semakin sering terjadi dan semakin parah dalam beberapa tahun terakhir, dan

Indonesia sebagai negara tropis sangat rentan terhadap bencana alam yang disebabkan oleh cuaca ekstrem, Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) mencatat tingginya angka bencana hidrometeorologi tidak lepas dari perubahan iklim. Koordinator Bidang Analisis Variabilitas Iklim BMKG mengatakan bahwa hasil monitoring BMKG dalam 40 tahun terakhir mengindikasikan curah hujan ekstrem di Indonesia mengalami kecenderungan peningkatan, baik dalam hal frekuensi maupun intensitas.

Penyebab dari bencana akibat curah hujan adalah perubahan iklim, pertumbuhan populasi, dan urbanisasi yang meningkatkan risiko bencana alam. Perubahan iklim menyebabkan curah hujan menjadi lebih tidak teratur dan intens. Pertumbuhan populasi menyebabkan peningkatan permukiman di daerah yang rentan terhadap bencana, sehingga risiko bencana semakin tinggi. Urbanisasi juga menyebabkan pengurangan luas daerah alami yang dapat menyerap air hujan, sehingga terjadilah banjir dan longsor. Menurut data yang di dapat dari situs Data Informasi Bencana Indonesia (https://dibi.bnpb.go.id/) terdapat 2690 banjir, 2228 tanah longsor, dan 272 kekeringan yang terjadi di Indonesia dalam kurun waktu 2017 hingga 2018, dan tanah longsor merupakan bencana yang paling banyak terjadi, pada tahun 2020 tercatat telah terjadi bencana tanah longsor di Jawa Barat sebanyak 3,232 kali. Data curah hujan yang akurat dan terorganisir dengan baik sangat diperlukan dalam membuat keputusan terkait mitigasi bencana. Jawa Barat adalah salah satu provinsi di Indonesia yang memiliki curah hujan yang tinggi dan sering mengalami bencana banjir. Oleh karena itu, pengelompokkan data curah hujan di Jawa Barat menjadi sangat penting dalam mengurangi risiko bencana alam. Data dalam penelitian ini mengacu pada data jumlah intensitas curah hujan di Provinsi Jawa Barat yang diperoleh dari situs resmi http://dataonline.bmkg.go.id/data_iklim pada periode tahun 2019-2022..

2. Landasan Pemikiran

Tinjauan pustaka merupakan salah satu bagian penting yang tidak terpisahkan dari sebuah penelitian. Tinjauan pustaka ini memuat ulasan dan analisis terhadap berbagai literatur terkait yang telah dipublikasi sebelumnya[3]. Penelitian ini menggunakan tinjauan pustaka dari beberapa jurnal penelitian terdahulu mengenai metode *clustering* dan algoritma-algoritma yang digunakan dalam proses klasterisasi atau pengelompokan. Algoritma yang paling banyak digunakan dalam proses klasterisasi adalah algoritma K-Mean, *K-Medoids*, dan Agglomerative Hierarchical *Clustering* (AHC). Metode dan algoritma yang digunakan dalam penelitian ini adalah metode *clustering* dengan menggunakan algoritma *k-means* yang menerapkan perhitungan jarak terdekat atau *euclidean distance* karena dalam proses perhitungan berdasarkan jarak terdekat dapat mengelompokan anggota dengan nilai dan karakteristik similarity yang tinggi dan memisahkan anggota dengan similarity yang rendah.

Clustering atau klasterisasi adalah suatu alat bantu pada data mining yang bertujuan mengelompokkan objek-objek kedalam *cluster-cluster*. *Clustering* merupakan suatu metode pengelompokan berdasarkan ukuran kedekatan. Perbedaan *clustering* dengan grup ialah grup memiliki kelompok yang sama karakteristiknya, namun *clustering* tidak harus sama akan tetapi pengelompokannya berdasarkan kedekatan dari suatu karakteristik *sample* yang ada, salah satunya dengan menggunakan rumus jarak terdekat. Terdapat beberapa metode dalam klasterisasi yang dikelompokan ke dalam beberapa kategori, diantaranya adalah:

1. Metode Berbasis Partisi (*Partitioning Methods*)

Bekerja dengan cara membagi atau mempartisi data ke dalam beberapa kelompok yang meliputi penentuan pusat-pusat *cluster* (*centroid*) berupa rata-rata, modus, atau sebuah objek representatif dari semua objek dalam suatu *cluster* berdasarkan ukuran tertentu. Algoritma yang termasuk dalam metode berbasis partisi diantaranya adalah *K-means*, *K-Medos*, *K-Medoids*, *Fuzzy C-Means*, dan lainnya.

2. Metode Berbasis Hirarki (*Hierarchical Methods*)

Bekerja dengan cara mengklaster atau mengelompokan objek-objek data ke dalam sebuah hirarki *cluster* yang bertujuan meringkas dan mempresentasikan data yang digunakan untuk mempermudah dalam proses visualisasi. Algoritma yang termasuk dalam metode berbasis hirarki diantaranya adalah *Balanced Iterative Reducing and Clustering using Hierarchies*, *Chameleon*, dan lainnya.

3. Metode Berbasis Kepadatan (*Density-based Methods*)

Bekerja dengan cara menemukan *cluster* dengan bentuk acak yang tidak teratur atau saling terkait. Metode berbasis kepadatan membentuk *cluster* sesuai dengan kepadatan tinggi anggota kumpulan data di lokasi yang ditentukan. Algoritma yang termasuk dalam metode berbasis kepadatan diantaranya adalah *Density-Based Spatial Clustering of Application with Noise*, *Ordering Points to Identify the Clustering Structure*, dan *Density-based Clustering*.

4. Metode Berbasis Kisi (*Grid-based Methods*)

Metode berbasis kisi menggunakan *approach space driven* (disetir oleh ruang) yaitu mempartisi atau mengelompokan *embedding space* ke dalam sel-sel yang tidak bergantung pada distribusi objek data. Algoritma yang termasuk dalam metode berbasis kisi diantaranya adalah *Statistical Information Grid*, dan *Clustering In Quest*.

Clustering merupakan suatu metode dalam data mining yang *unsupervised*, karena tidak ada satu atribut pun yang digunakan untuk memandu proses pembelajaran, jadi seluruh atribut *input* diperlakukan sama.

Data mining dapat diartikan sebagai proses penambangan data yang menghasilkan sebuah *output* (keluaran) berupa pengetahuan. Mengetahui suatu proses dapat di selesaikan dan dimulai dari sebuah *input*

(data) kemudian di proses sehingga menghasilkan sebuah *output* (keluaran). Tentuya dalam data mining juga mengalami fase yang sama, yang membedakannya adalah pada data mining yang menjadi *input* adalah himpunan data, prosesnya adalah algoritma atau metode dalam data mining itu sendiri, dan *output* nya adalah berupa pengetahuan dalam bentuk pola, decision tree, *cluster* dan lain sebagainya.

Data mining merupakan analisis dan peringkasan yang dilakukan terhadap kumpulan data (dataset) untuk menemukan hubungan antar data yang selama ini tidak diketahui sehingga mudah dimengerti dan berguna bagi pemilik data. Beberapa teknik data mining yang sering digunakan adalah klasterisasi, klasifikasi, prediksi, dan aturan asosiasi. Data Mining adalah teknik yang dilakukan pada basis data besar untuk mengekstraksi pola tersembunyi dengan menggunakan strategi kombinasional dari analisis statistik, pembelajaran mesin, dan teknologi basis data.

3. Metode Penelitian

3.1. Instrumen Penelitian

Untuk mendukung berjalannya penelitian, dibutuhkan bahan dan peralatan agar penelitian ini dapat berjalan dengan baik.

3.2. Bahan penelitian

Bahan dalam penelitian ini adalah data yang berkaitan dengan judul penelitian yaitu data jumlah intensitas curah hujan di Provinsi Jawa Barat yang diperoleh dari situs resmi http://dataonline.bmkg.go.id/data_iklim pada periode tahun 2019-2022.

3.3. Peralatan penelitian

Peralatan dalam penelitian ini membutuhkan beberapa instrumen untuk dapat memaksimalkan proses dan hasil yang ingin dicapai. Pada penelitian ini terdapat kebutuhan perangkat lunak dan perangkat keras.

Prosedur Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data yang menggambarkan jumlah intensitas curah hujan di wilayah Jawa Barat. Data tersebut merupakan data sekunder yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) http://dataonline.bmkg.go.id/ tahun 2022 , Jenis data dalam penelitian ini adalah data kuantitatif, yakni data yang dalam bentuk angka atau bilangan yang dapat dioperasikan dengan perhitungan matematika. Sumber data yang digunakan dalam penilitian ini adalah intensitas curah hujan pada lima stasiun cuaca di Jawa Barat, yaitu:

1. Stasiun Geofisika Bandung
2. Stasiun Klimatologi Bogor
3. Stasiun Meteorologi Citeko
4. Stasiun Meteorologi Kertajati
5. Pos Meteorologi Penggung

3.3 Teknik Analisis Data

Dalam penelitian ini menggunakan normalisasi standar deviasi untuk menormalisasi data awal, normalisasi standar deviasi memiliki keunggulan untuk mengatasi perbedaan skala dan rentang nilai antara atribut-atribut data. Juga dapat membantu mengubah nilai data menjadi skala umum yang memiliki rata-rata nol. Hal ini dapat memudahkan proses pengelompokan data dengan menggunakan algoritma 21

seperti k-means yang menghitung jarak antara data dan pusat cluster. Dengan begitu data yang memiliki skala yang berbeda tidak akan mendominasi atau mengurangi pengaruh data lain yang memiliki skala yang lebih kecil. Normalisasi standar deviasi juga dapat mengurangi dampak dari outlier atau nilai ekstrem yang dapat mengganggu hasil pengelompokan data.

4. Pembahasan

Data yang telah diseleksi kemudian melalui tahapan normalisasi data dengan persamaan standar deviasi yang dilakukan dengan mencari nilai rata-rata dan nilai standar deviasi pada setiap atribut terlebih dahulu.

Nilai rata-rata dan nilai standar deviasi yang telah dihasilkan selanjutnya akan diterapkan pada proses normalisasi data pada tiap atribut dengan persamaan standar deviasi sebagai berikut: $Standar\ Deviasi\ (\sigma)=\sqrt{\Sigma(xi-\mu)2n}$

Keterangan:

σ : Standar Deviasi

xi : data ke-i

μ : nilai rata-rata

n : banyak data 30

dan selanjutnya rumus untuk menormalisasi dengan menggunakan Standar Deviasi adalah sebagai berikut: $x\_baru=x\_lama-\mu\sigma$

Berikut proses tahapan normalisasi dataset awal dengan menggunakan persamaan standar deviasi:

1. Normalisasi Atribut Bulan Januari

$Xnew$=647,1−1239,8382,5= -1,55

2. Normalisasi Atribut Bulan Februari

$Xnew$=852,6−1598,7417,9= -1,79

3. Normalisasi Atribut Bulan Maret

$Xnew$=1062−1324,8212,1= -1,24

4. Normalisasi Atribut Bulan April

$Xnew$=1083,7−1363,8303,0= -0,92

5. Normalisasi Atribut Bulan Mei

$Xnew$=921,1−865,4293,2= 0,19

6. Normalisasi Atribut Bulan Juni

$Xnew$=299,8−560,3338,2= -0,77

7. Normalisasi Atribut Bulan Juli

$Xnew$=192,5−346,5187,9= -0,82

8. Normalisasi Atribut Bulan Agustus

$Xnew$=163,5−350,5356,0= -0,53

9. Normalisasi Atribut Bulan September

$Xnew$=346,1−417,4321,6= -0,22
10. Normalisasi Atribut Bulan Oktober

$Xnew$=996,6−1094,1533,8= -0,18
11. Normalisasi Atribut Bulan November

$Xnew$=1239,7−1076,9252,9= 0,64
12. Normalisasi Atribut Pos Desember

$Xnew$=1053,5−1415,1258,3= -1,40

Simulasi Perhitungan Manual AlgoritmaK-means
Perhitungan manual dengan menggunakan algoritma k-means dan rumus euclidean distance bertujuan untuk mengetahui hasil dari cluster yang dikelompokan 32

berdasarkan jarak terdekat tiap anggota cluster. Hasil tersebut dapat diperoleh dengan melakukan beberapa langkah, yaitu:
1. Menentukan jumlah cluster yang diinginkan. Cluster yang dicari dalam penelitian ini adalah tiga cluster yaitu C1, dan C2
2. Menentukan centroid awal pada dataset secara random atau acak

Daftar Pustaka

[1] A. Susilo et al., "Coronavirus Disease 2019 : Tinjauan Literatur Terkini Coronavirus Disease 2019 : Review of Current Literatures," J. Penyakit Dalam Indones., vol. 7, no. 1, pp. 45–67, 2020.

[2] ABC, "Inilah Strategi Sejumlah Negara Untuk Menangani Pandemik Global Virus Corona," 17 Maret 2020, 2020. https://www.tempo.co/abc/5397/inilah-strategi-sejumlah-negara-untuk-menangani-pandemik-global-virus-corona (accessed Jun. 01, 2020).

[3] K. Alfi, "Jabar Tegaskan Kebijakan Penanganan Covid-19 Sejalan dengan Pusat," 06 April 2020, 2020. https://news.detik.com/berita/d-4966614/jabar-tegaskan-kebijakan-penanganan-covid-19-sejalan-dengan-pusat (accessed Jun. 02, 2020).

[4] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," Procedia Comput. Sci., vol. 167, pp. 1821–1828, 2020, doi: 10.1016/j.procs.2020.03.201.

[5] M. Syarifuddin, "ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN," Inti Nusa Mandiri, vol. 15, no. 1, pp. 1–6, 2020.

[6] O. Dwiraswati and K. N. Siregar, "ANALISIS SENTIMEN PADA TWITTER TERHADAP PENGGUNAAN ANTIBIOTIK DI INDONESIA DENGAN NAIVE BAYES CLASSIFIER," Media Inf., vol. 15, no. 1, pp. 1–9, 2019.

[7] B. M. Pintoko and K. M. L, "Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier," e-Proceeding Eng., vol. 5, no. 3, pp. 8121–8130, 2018.

[8] S. Afrizal, H. N. Irmanda, and N. Falih, "Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit," vol. 4221, pp. 157–168, 2019. 71

[9] N. Ruhyana, "Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil / Genap Pada Twitter Dengan Metode Klasifikasi Naive Bayes," J. IKRA-ITH Inform., vol. 3, no. 1, pp. 94–99, 2019.

[10] B. Liu, Sentiment Analysis and Opinion Mining. 2012.

[11] R. Rosdiana, T. Eddy, S. Zawiyah, and N. Y. U. Muhammad, "Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar," pp. 87–93, 2019.

[8] Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Eecis, 7(1), 59–64. https://doi.org/10.1038/hdy.2009.180

[9] Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, 2(3), 207–217.

[9] Tampubolon, K., Saragih, H., & Reza, B. (2013). Implementasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-Alat Kesehatan, 93–106.

[10] Via, Y. V., Nugroho, B., & Syafrizal, A. (2015). Sistem Pendukung Keputusan Klasifikasi Tingkat Keganasan Kanker Payudara Dengan Metode Naive Bayes Classifier, X, 2–7.

[11] Virgana, Pauziah, U., & Sonny, M. (2014). Kajian Algoritma Naïve Bayes Dalam Pemilihan Penerimaan Beasiswa Tingkat SMA. Jurnal Administrasi Bisnis.

[12] Wati, M., & Hadi, D. A. (2016). Implementasi Algoritma Naive Bayesian Dalam Penentuan Penerima Program Bantuan Pemerintah, 3(1), 22–26.

[13] Wulan Sari, B., & Prabowo, D. (2017). Penentuan Kelayakan Penerima Bantuan Renovasi Rumah Warga Miskin Menggunakan Naïve Bayes, 39(5), 561–563.