Analisis Probabilitas Bencana Alam dengan Penerapan Data Mining Menggunakan K-Means dan Linier Regression

Mochamad Althaf Pramasetya Perkasa[1] , Rianto[2]
[1,2] Universitas Siliwangi, Kota Tasikmalaya
Email : 207006080@student.unsil.ac.id

Abstrak

Penggunaan data mining untuk menghitung probabilitas data bencana adalah pendekatan yang dapat membantu dalam mengidentifikasi pola-pola dan hubungan antara variabel-variabel yang berkontribusi terhadap terjadinya bencana alam. Beberapa teknik data mining yang dapat digunakan termasuk K-Means, regresi linear, dan metode lainnya seperti Decision Tree, Naive Bayes, dan Neural Networks. Penelitian ini dilakukan untuk mendapatkan hasil analisa data dari kejadian bencana alam di wilayah indonesia, dengan metode data mining. Teknik yang digunakan pada data mining yaitu k-means dan linear regression. Dataset diambil dari website BNPB, dengan menjumlahkan data kejadian bencana alam di berbagai wilayah Indonesia dari tahun 2019 hingga 2022.

.

Kata Kunci: ***K-means, Linear Regression, Data Mining***

PENDAHULUAN

Bencana alam dan teknologi adalah dua hal yang saling terkait dan mempengaruhi kehidupan manusia. Bencana alam seperti gempa bumi, tsunami, banjir, angin topan, dan kebakaran hutan dapat terjadi secara tiba-tiba dan merusak infrastruktur dan sumber daya manusia. Di sisi lain, teknologi dapat membantu memperkirakan dan memitigasi dampak bencana alam, serta memberikan solusi untuk memulihkan kehidupan manusia setelah terjadi bencana. . Eksplorasi data atau data mining adalah metode pengolahan data yang digunakan untuk mengeksplorasi informasi yang tersimpan dalam data.

Data mining merupakan proses pencarian pola atau informasi yang signifikan dalam data tertentu dengan menggunakan teknik atau metode khusus. Berbagai macam teknik, metode, atau algoritma dalam data mining dapat digunakan tergantung pada tujuan yang ingin dicapai (Mardi, Y. 2017). Dalam proses data mining, terdapat integrasi teknik dari berbagai disiplin ilmu seperti teknologi database dan data warehouse, statistik, machine learning, komputasi tingkat tinggi, pattern recognition, neural network, dan visualisasi data untuk menggali informasi dari data yang terkumpul (Nur, F., Zarlis, M., & Nasution, B. B. 2017).

K-means dan *linear regression* adalah teknik yang diterapkan pada proses data mining untuk menganalisis data kejadian bencana alam. Algoritma *k-means* melibatkan pembentukan partisi klaster awal, di mana setiap data ditempatkan pada klaster yang paling dekat dengan pusat klaster. Setelah itu, algoritma secara iteratif memperbaiki partisi klaster dengan memindahkan setiap data ke klaster lain yang lebih sesuai, hingga tidak terjadi perubahan signifikan pada partisi klaster yang telah dibentuk(Sulistiyawati, A., & Supriyanto, E. 2021).. Pada teknik *linear regression*, diasumsikan terdapat suatu hubungan antara variabel tak bebas yang ingin diprediksi dengan variabel bebas yang saling terkait. Proses prediksi didasarkan pada asumsi bahwa pola pertumbuhan data historis yang terjadi

bersifat linier, meskipun pada kenyataannya tidak selalu benar. Pola pertumbuhan ini kemudian didekati melalui sebuah model yang merepresentasikan hubungan-hubungan yang terdapat dalam keadaan tersebut(Satyahadewi, N., Ediyanto, & Mara, M. N. 2013).

Penelitian ini dilakukan untuk mendapatkan hasil analisa data dari kejadian bencana alam di wilayah indonesia, dengan metode data *mining*. Teknik yang digunakan pada data mining yaitu k-means *dan linear regression*. Teknik *k-means* digunakan untuk membuat klasterisasi, sedangkan *liner regression* untuk membuat hasil prediksi dari data bencana alam(Mardi, Y. 2017).

METODE

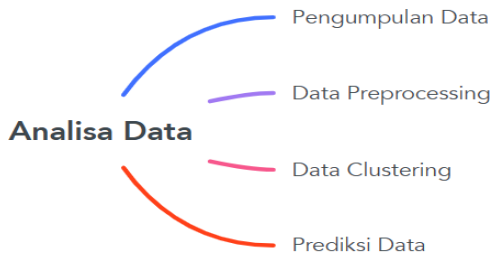
Penelitian dilakukan dengan metode kualitatif, bertujuan untuk mendapatkan hasil analisa dari probabilitas kejadian bencana alam di Indonesia menggunakan data mining.

Tabel 1. Studi literatur

No	Penulis	Fokus Penelitian	Hasil Penelitian
1	Saadi, dkkx	Penggunaan metode k-menas dan liner regresi untuk lokalitas sinyal melalui cahaya	Menunjukkan bahwa penggunaan cahaya tampak sebagai sumber sinyal lokalitas dalam ruangan memiliki potensi yang besar. Metode ini dapat mengatasi beberapa masalah yang sering terjadi dalam sistem lokalitas dalam ruangan berbasis radio, seperti interferensi sinyal dan ketidakstabilan lingkungan**.**
2	Oladele Tinuke Omolewa , dkk	Prediksi hasil kerja akademik mahasiswa menggunakan linear regresi berganda dan pengelompokkan melalui k-means	Penggunaan pengelompokan k-means sebagai langkah pra-pemrosesan dapat membantu dalam mengidentifikasi pola atau kelompok mahasiswa yang serupa dalam hal kinerja akademik. Hal ini dapat memberikan wawasan yang berharga dalam memahami faktor-faktor yang mempengaruhi kinerja akademik mahasiswa.
3	Gbadamosi Babatunde, dkk	Penggunaan linear regresi berganda untuk prediksi hasil produksi pertanian dan k-means untuk pengelompokkan	Menganalisis dampak perubahan iklim pada hasil produksi pertanian menggunakan metode pengelompokan K-means dan regresi linear berganda. Tujuan utama penelitian ini mungkin adalah untuk mengevaluasi hubungan antara perubahan iklim dan produktivitas pertanian, serta mengidentifikasi faktor-faktor yang berkontribusi terhadap hasil produksi pertanian dalam konteks perubahan

			iklim.
4.	Gan Pei Yee, dkk	Pengelompokkan k-menans dan prediksi mneggunakan linear regresi berganda untuk memprediksi pendapatan rumah tangga	Menunjukkan adanya pola pengelompokan pendapatan rumah tangga di Malaysia dengan menggunakan algoritma K-means. Melalui analisis K-means, penelitian ini mengidentifikasi kelompok pendapatan yang berbeda berdasarkan karakteristik yang relevan, seperti pendidikan, pekerjaan, usia, dan faktor sosioekonomi lainnya.
5	Ramadhan, Prihandoko	Penggunaan Data mining untuk prediksi bencana alam	Pengelompokkan data bencana di seluruh provinsi di Indonesia, yang dimiliki oleh BNPD, telah berhasil dilakukan dalam penelitian ini. Selain itu, juga dilakukan prediksi data bencana yang akan terjadi dalam 5 tahun ke depan menggunakan teknik data mining, dengan penerapan Algoritma K-Means dan algoritma Linear Regression. Penggunaan Algoritma K-Means bertujuan untuk melakukan pengelompokkan data bencana berdasarkan provinsi dalam periode 2005-2015.

Tahapan Penelitian



Gambar 1. Diagram Alur Metode Penelitian

Langkah pertama pada penelitian ini yaitu pengumpulan data, pengumpulan data dilakukan dengan menggunakan data resmi dari website BNPB. Data yang diambil yaitu dari tahun 2019 hingga 2022, menggunakan data kejadian bencana alam di wilayah Indonesia. Selanjutnya yaitu preprocessing data, langkah ini yaitu bertujuan untuk membersihkan data-data yang nantinya memudahkan pada saat melakukan tahap data *clustering*, misalnya jika ada data yang berbentuk nominal maka akan digantikan ke dalam bentuk angka. Setelah melakukan pra-proses data, tahapan penelitian selanjutnya yaitu melakukan data *clustering* dengan menggunakan metode k-means, tujuan pengelompokkan data adalah untuk meminimalkan objective function yang telah ditetapkan dalam proses clustering, yang pada dasarnya berusaha untuk meminimalkan variasi dalam satu cluster dan memaksimalkan variasi antar cluster.. Data bencana yang telah dikelompokkan berdasarkan tingkat kemiripannya akan dijadikan bahan untuk memprediksi data kejadian bencana alam dengan menggunakan algoritma linear regresi.

HASIL DAN PEMBAHASAN

Pada bagian hasil dan pembahasan akan diuraikan hasil proses penggunaan data mining, dan analisa data probabilitas yang di dapat dari hasil prediksi menggunakan *linear regression*.

Hasil

A. Pengumpulan Data

Dataset diambil dari website BNPB, dengan menjumlahkan data kejadian bencana alam di berbagai wilayah Indonesia dari tahun 2019 hingga 2022. Untuk data kejadian bencana tersebut sudah termasuk dengan rata-rata jenis bencana alam yang sering terjadi pada satu tahunnya. Wilayah yang ada pada dataset diantaranya Sumatera, Jawa, Kepulauan Riau, Sulawesi, Kalimantan, Papua, Bali, Nusa Tenggara, dan Maluku. Data kejadian bencana alam yang telah dijumlahkan dari tahun 2019 hingga 2022 terdapat pada Tabel 2.

Tabel 2. Dataset kejadian bencana alam

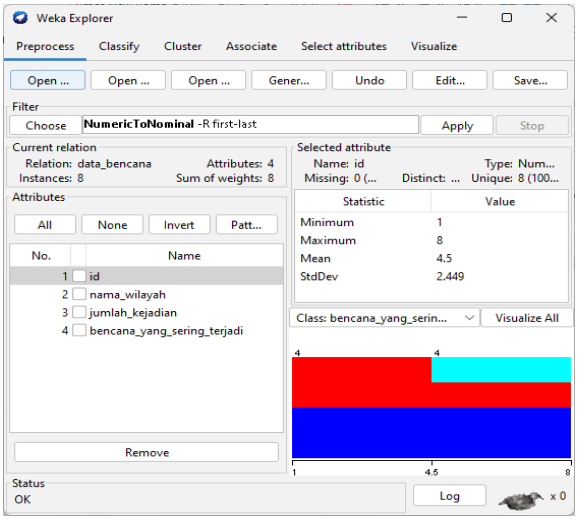
id	Nama Wilayah	Jumlah Kejadian	Bencana yang Sering Terjadi
1	Sumatera	3.131	Tanah Longsor
2	Jawa	8.752	Banjir
3	Kepulauan Riau	182	Tanah Longsor
4	Sulawesi	763	Banjir
5	Kalimantan	926	Puting Beliung
6	Papua	230	Tanah longsor
7	Bali, Nusa Tenggara	252	Banjir
8	Maluku	252	Tanah Longsor

Tabel 3. Atribut pada dataset kejadian bencan alam

No	Atribut	Keterangan
1	id	Nomor yang diberikan pada setiap wilayah
2	Nama Wilayah	Nama wilayah yang ada di Indonesia
3	Jumlah Kejadian	Jumlah kejadian bencana alam dari tahun 2019-2022
4	Bencana yang Sering Terjadi	Jenis bencana alam yang sering terjadi pada tahun 2019-2022

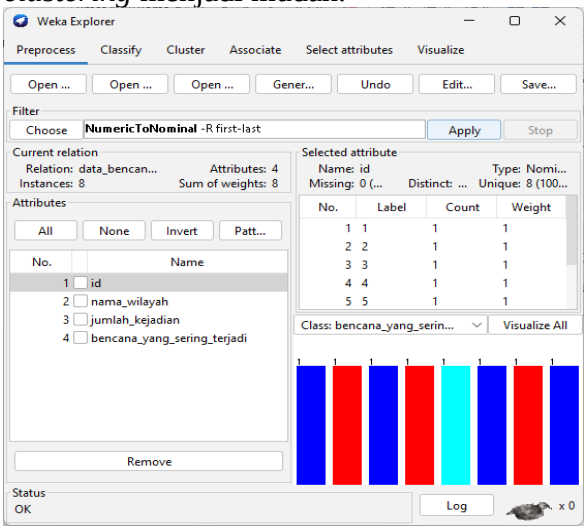
B. *Preprocessing* Data

Preprocessing data dilakukan dengan menggunakan aplikasi *Weka*, untuk langkah pertama yaitu import dataset yang akan dilakukan preprocessing. Jika dataset berhasil terdeteksi langkah selanjutnya yaitu mengubah tipe data dari masing-masing atribut untuk melakukan proses pembersihan data. Tujuan preprocessing data ini yaitu mengganti dan mengisi nilai dataset yang hilang pada masing-masing atribut. Dapat dilihat pada Gambar 2 dan 3, penggunaan aplikasi *Weka* untuk preprocessing data kejadian bencana alam.



Gambar 2. Tampilan pada aplikasi *Weka* jika dataset sudah terbaca

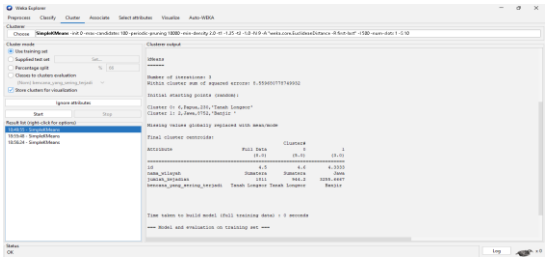
Gambar 2 merupakan tampilan jika file dataset yang di *import* telah berhasil, setelah file data kejadian bencana alam terdeteksi pada aplikasi *Weka* selanjutnya dilakukan praproses data menggunakan fitur yang ada pada *Weka*. Tahapan ini dilakukan untuk membuat dataset yang digunakan menjadi bersih agar proses data *clustering* menjadi mudah.



Gambar 3. Hasil *preprocessing* data menggunakan *Weka*

C. Data *Clustering*

Data *clustering* digunanakan untuk mengelompokkan atribut-atribut yang memiliki kesamaan, clustering disini menggunakan algoritma *k-means*. Metode data clustering menggunakan aplikasi Weka, untuk menetapkan klaster-klaster yang akan digunakan untuk tahapan prediksi data.



Gambar 4. Tampilan hasil *clustering* menggunakan aplikasi *Weka*

Dapat dilihat pada Gambar diatas (Gambar 4) dibuat 2 klaster untuk mengelompokkan wilayah yang memiliki jenis bencana alam yang sama. Klaster pertama mengelompokkan wilayah yang memiliki tingkat kerawanan bencana alam tanah longsor dan klaster kedua mengelompokkan wilayah dengan tingkat kerawanan bencana alam banjir. Wilayah Sumatera yaitu daerah yang memiliki tingkat terjadinya bencana alam yang tinggi pada klaster 1 sedangkan pada klaster 2 wilayah Jawa memiliki tingkat terjadinya bencana alam yang tinggi.

D. Prediksi Data

Prediksi data bencana ini bertujuan untuk mendapatkan angka probabilitas yang didapat dari model *linear regression*. Data yang nantinya akan dijadikan bahan untuk diprediksi diantaranya dataset bencana alam, kemungkinan terjadinya banjir, tanah longsor, dan putting beliung.

	id	nama_wilayah	jumlah_kejadian	bencana_yang_sering_terjadi	jenis_bencana_0	jenis_bencana_1	jenis_bencana_2
0	1	Sumatera	3131	Tanah Longsor	1	0	0
1	2	Jawa	8752	Banjir	0	1	0
2	3	Kepulauan Riau	182	Tanah Longsor	1	0	0
3	4	Sulawesi	763	Banjir	0	1	0
4	5	Kalimantan	926	Puting Beliung	0	0	1

```
# Memisahkan fitur dan target
X = df_encoded[['jumlah_kejadian']]
y = df_encoded[['jenis_bencana_1', 'jenis_bencana_0', 'jenis_bencana_2']]
```

```
# Melatih model regresi linear
model = LinearRegression()
model.fit(X, y)
```

```
LinearRegression()
```

Gambar 5. Pembuatan model *linear regression* untuk prediksi data bencana

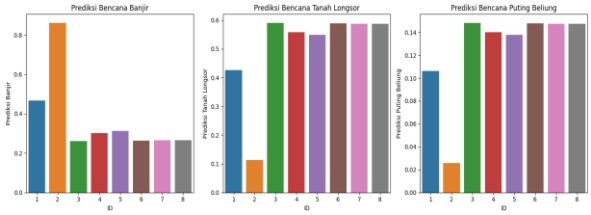
Pembahasan

Analisis data hasil prediksi bencana menggunakan regresi linear dilakukan untuk memprediksi jumlah kejadian bencana di wilayah Jawa, Sumatera, Kepulauan Riau, Sulawesi, Kalimantan, Papua, Bali, Nusa Tenggara, dan Maluku. Regresi linear digunakan untuk membangun model matematis yang menghubungkan variabel dependen (jumlah kejadian bencana). Dengan melakukan analisis ini, dapat diperoleh pemahaman yang lebih baik tentang faktor-faktor yang berkontribusi terhadap jumlah kejadian bencana di wilayah-wilayah tersebut.

Tabel 3. Probabilitas peningkatan bencana

id	nama_wilayah	Prediksi Banjir	Prediksi Tanah Longsor	Prediksi Puting Beliung
1	Sumatera	0.467547551204491	0.426350598960708 54	0.1061018498348005
2	Jawa	0.8616458734169484	0.11272689953505893	0.02562722704799275
3	Kepulauan Riau	0.2607879083999123	0.5908900562825802	0.14832203531750754
4	Sulawesi	0.3015228532861314	0.5584731608251345	0.1400039858887341
5	Kalimantan	0.3129510736242617	0.5493785756967977	0.13767035067894054
6	Papua	0.2641532738982574	0.5882118962447878	0.1476348298569548
7	Bali_Nusa Tenggara	0.2656957330849989	0.5869844062274663	0.14731986068753483
8	Maluku	0.2656957330849989	0.5869844062274663	0.14731986068753483

Pada gambar diatas (Gambar 4) dapat dilihat jika probabilitas peningkatan terjadinya bencana alam yang sering dialami pada wilayah-wilayah di Indonesia meningkat. Bencana alam tersebut diantaranya banjir, tanah longsor dan puting beliung. Probabilitas jenis bencana tersebut diambil dari klaster yang telah dibuat sebelumnya, dengan menambahkan satu klaster baru yaitu klaster puting beliung.



Gambar 6. Diagram probabilitas bencana alam yang sering terjadi

Gambar 7. Merupakan visualisasi dari peningkatan probabilitas dari bencana alam yang sering terjadi di wilayah Indonesia. Pada klaster banjir dapat dilihat jika wilayah Jawa memiliki probabilitas peningkatan bencana yang tinggi, kemudian pada klaster tanah longsor dan puting beliung probabilitas yang paling tinggi peningkatannya dimiliki oleh wilayah Kepulauaan Riau. Angka yang tertera pada bagian kurva x merupakan id dari masing-masing wilayah. Probabilitas ini diambil hanya dari data-data jumlah kejadiannya saja, untuk hasilnya tidak bisa dijadikan landasan sebagai hasil yang akurat. Visualisasi grafik dibuat menggunakan pemrograman bahasa python.

SIMPULAN (PENUTUP)

Hasil dari penelitian ini yaitu didapatkannya probabilitas kejadian yang sering terjadi pada beberapa wilayah di Indonesia. setelah melakukan data *clustering*,didapat 3 klaster yaitu cluster wilayah yang sering terjadinya bencana banjir, tanah longsor, dan puting beliung.Ketiga kluster tersebut diantaranya klaster banjir, tanah longsor, dan putting beliung. Masing-masing klaster memiliki anggota, pada klaster banjir(Jawa, Sulawesi, Bali, dan Nusa Tenggara), tanah longsor(Sumatera, Kepulauan Riau, Papua, dan Maluku), puting beliung hanya memiliki satu anggota yaitu wilayah Kalimantan .Probabilitas tersebut dihitung menggunakan permodelan *linear regression*, menggubakan pemrograman bahada python. Dari grafik atau tabel yang diberikan dari hasil perhitungan menggunakaan linear regression, wilayah mengalami probabilitas peningkatan pada bencana alam banjir adalah wilayah Jawa. Untuk peningkatan probabilitas bencana alam tanah longsor dan puting beliung dialami oleh wilayah Kepulauan Riau. Perhitungan ini tidak memberikan efek yang signifikan pada kejadian aktualnya, karena data yang diambil hanya dari sudut pandang jumlah kejadiannya saja. Untuk perhitungan yang lebih efektif diperlukan data-data yang lebih kompleks.

DAFTAR PUSTAKA

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Elektrotechnik Und Informationstechnik*, *2*(2), 213–219. https://doi.org/10.22202/jei.2016.v2i2.1465

Nur, F., Zarlis, M., & Nasution, B. B. (2017). PENERAPAN ALGORITMA K-MEANS PADA SISWA BARU SEKOLAHMENENGAH KEJURUAN UNTUK CLUSTERING JURUSAN. *InfoTekJar : Jurnal Nasional Informatika Dan Teknologi Jaringan*. https://doi.org/10.30743/infotekjar.v1i2.70

Sulistiyawati, A., & Supriyanto, E. (2021). Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan. *Jurnal Tekno Kompak*, *15*(2), 25. https://doi.org/10.33365/jtk.v15i2.1162

Satyahadewi, N., Ediyanto, & Mara, M. N. (2013). PENGKLASIFIKASIAN KARAKTERISTIK DENGAN METODE K-MEANS CLUSTER ANALYSIS. *BIMASTER*, *2*(02). http://jurnal.untan.ac.id/index.php/jbmstr/article/download/3033/2998

Saadi, M., Saeed, Z., Ahmad, T., Saleem, M., & Wuttisittikulkij, L. (2019). Visible light-based indoor localization using*k-means*clustering and linear regression. *Transactions on Emerging*

Telecommunications Technologies, *30*(2), e3480. https://doi.org/10.1002/ett.3480

Omolewa, O. R., Oladele, A. T., Adeyinka, A. A., & Ogundokun, R. O. (2019). Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions. *Journal of Engineering and Applied Sciences*, *14*(22), 8254–8260. https://doi.org/10.36478/jeasci.2019.8254.8260

Babatunde, G., Emmanuel, A. A., Oluwaseun, O. R., Bunmi, O. B., & Precious, A. E. (2019). Impact of Climatic Change on Agricultural Product Yield Using K-Means and Multiple Linear Regressions. *International Journal of Education and Management Engineering*, *9*(3), 16–26. https://doi.org/10.5815/ijeme.2019.03.02

Yee, G. P., Rusiman, M. S., Ismail, S., Suparman, S., Hamzah, F. M., & Shafi, M. (2023). K-means clustering analysis and multiple linear regression model on household income in Malaysia. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *12(2)*, *731*, *12*(2), 731. https://doi.org/10.11591/ijai.v12.i2.pp731-738

Ramadhan, M. F., & Prihandoko, P. (2019). PENERAPAN DATA MINING UNTUK ANALISIS DATA BENCANA MILIK BNPB MENGGUNAKAN ALGORITMA K-MEANS DAN LINEAR REGRESSION. *Jurnal Ilmiah Informatika Komputer*, *22*(1). https://ejournal.gunadarma.ac.id/index.php/infokom/article/view/1535

Yadav, A. K., Malik, H., & Chandel, S. S. (2014). Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models. *Renewable & Sustainable Energy Reviews*, *31*, 509–519. https://doi.org/10.1016/j.rser.2013.12.008

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, *28*(3), 591. https://doi.org/10.2307/1910133