**LAPORAN UAS**

**ANALISIS TAHAP PRE PROCESSING DAN ANALISIS ALGORITMA K-MEANS, CANOPY, DAN EXPECTATION-MAXIMIZATION DALAM MENGETAHUI PENYEBARAN WILAYAH GEMPA DENGAN RESIKONYA**

Mata Kuliah Pengantar Data Science



**Dosen :**

Wawan Gunawan, S.Kom, MT

**Disusun Oleh :**

Muhammad Fauzan Zamzami (41522010080)

Vemas Adi Pratama (41522010081)

Okky Tri Wahyuda (41522010114)

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS MERCU BUANA**

**2023**

# Pendahuluan

Gempa bumi merupakan peristiwa alam yang dapat menimbulkan dampak serius terhadap kehidupan dan infrastruktur. Pemahaman mendalam tentang pola dan karakteristik gempa sangat penting untuk mitigasi risiko dan penanganan bencana. Dalam konteks ini, dataset tentang gempa yang diperoleh dari penelitian oleh Ody Octoria Wijaya, sebanyak 9710 data, menjadi sumber informasi berharga untuk penelitian ini. Penelitian ini bertujuan untuk menguji, membandingkan, dan menganalisis kinerja algoritma EM (Expectation-Maximization), Canopy, dan K-Means dalam pengelompokan pola gempa bumi.

Tujuan utama dari penelitian ini adalah untuk menyediakan wawasan yang lebih mendalam tentang keefektifan ketiga algoritma tersebut dalam mengelompokkan data gempa. Dengan pemahaman yang lebih baik tentang pola gempa, diharapkan penelitian ini dapat memberikan kontribusi dalam upaya mitigasi risiko bencana gempa bumi.

Penelitian ini akan dibatasi oleh penggunaan dataset yang dikumpulkan oleh Ody Octoria Wijaya sebagai sumber data utama. Dalam menghadapi tantangan mitigasi risiko gempa, penelitian ini dianggap relevan dan bermanfaat untuk memberikan dasar yang kuat dalam mengidentifikasi pola gempa secara lebih efisien.

Laporan ini akan terstruktur dalam beberapa bagian utama. Pertama, kami akan membahas tahap pre-processing data, yang mencakup pengolahan dan penyiapan data untuk analisis lebih lanjut. Selanjutnya, tahapan analisis dan visualisasi akan memberikan gambaran tentang karakteristik dataset gempa yang digunakan. Bagian selanjutnya akan membandingkan kinerja algoritma EM, Canopy, dan K-Means dalam mengelompokkan pola gempa. Terakhir, penerapan algoritma-algoritma tersebut pada dataset akan disajikan sebagai bagian penting dari penelitian ini. Dengan demikian, diharapkan bahwa penelitian ini dapat memberikan sumbangan positif terhadap pemahaman dan mitigasi risiko terkait gempa bumi.

# Daftar isi

[Pendahuluan 2](#_Toc153279778)

[Daftar isi 3](#_Toc153279779)

[I. Pengolahan Data 4](#_Toc153279780)

[a) Pengumpulan Data dan Tahap Pre Processing 4](#_Toc153279781)

[b) Langkah-langkah Normalisasi Data 6](#_Toc153279782)

[II. Analisis Perilaku Data terhadap Kelas dengan Visualisasi 14](#_Toc153279783)

[III. Perbandingan Algoritma 19](#_Toc153279784)

[CART 19](#_Toc153279785)

[Neural Network 19](#_Toc153279786)

[SGD 19](#_Toc153279787)

[IV. Penerapan 21](#_Toc153279788)

[V. Referensi 21](#_Toc153279789)

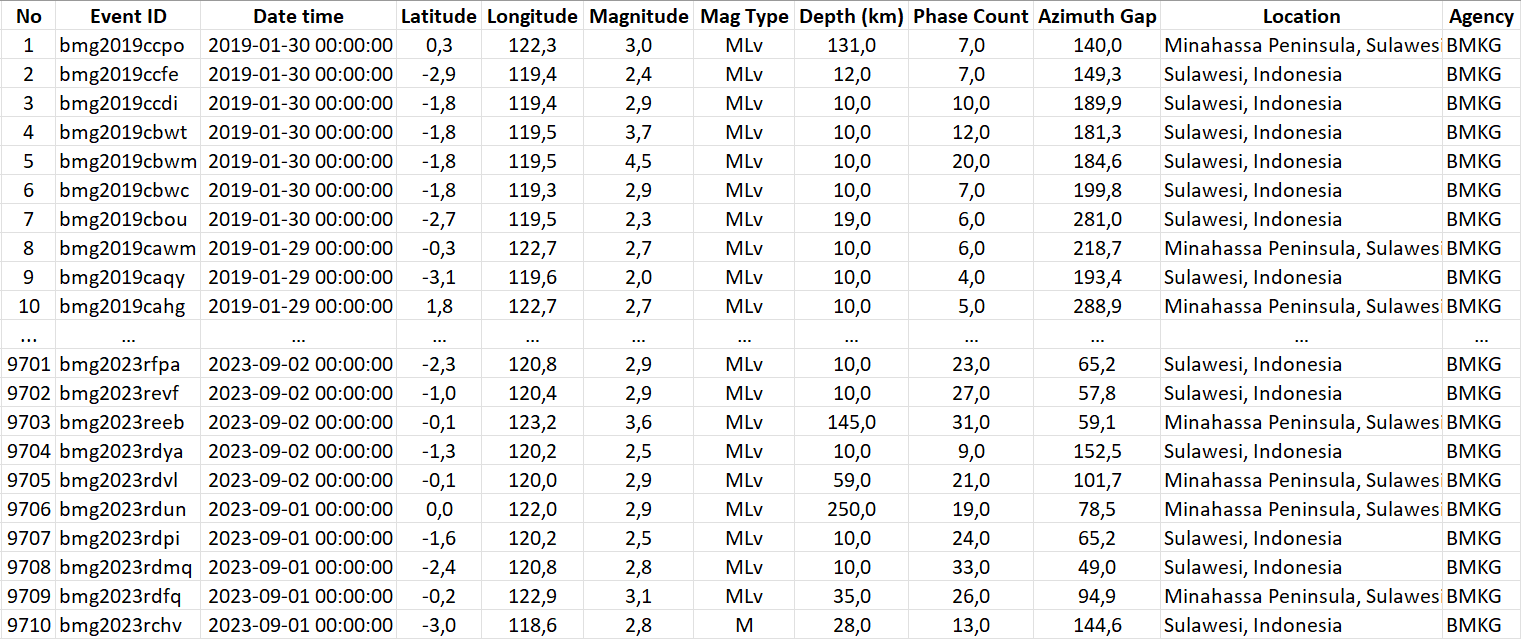
# Pengolahan Data

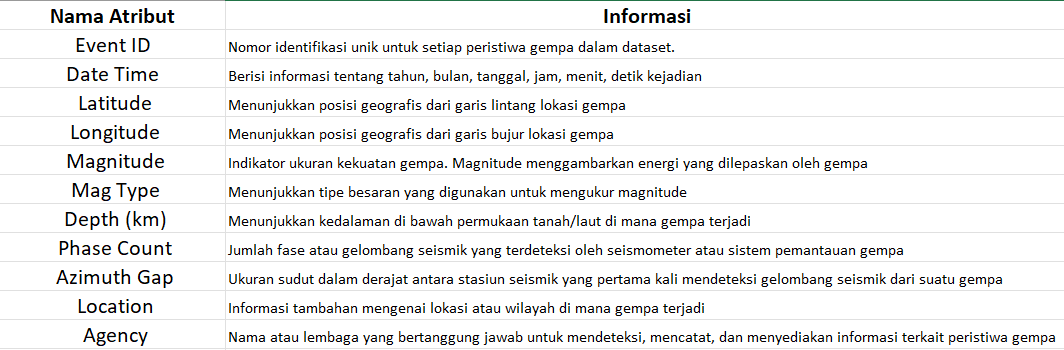
## Pengumpulan Data dan Tahap Pre Processing

Pada tahap ini, data mentah diolah untuk menghasilkan data yang bersih, termasuk pengisian atau penghapusan nilai kosong, penyamarataan format pada setiap atribut, dan penggantian angka yang mungkin typo.

Dataset yang digunakan berasal dari Kumpulan Dataset Gempa oleh Ody Octoria Wijaya yang berisi data-data beratribut mengenai gempa.

Dataset ini terdiri dari 9710 data dengan 11 atribut. Berikut adalah rinciannya:

Gambar 1.2 Dataset Mentah

Gambar 1.3 Atribut Dataset Mentah

Setelahnya, data tersebut menjalani proses preprocessing, di mana terdapat beberapa proses pembersihan data diantaranya :

1. Menghilangkan Atribut yang Tidak Diperlukan

Kami melakukan penghapusan beberapa atribut dalam dataset. Pertama, atribut Event ID dianggap tidak diperlukan karena hanya berfungsi sebagai ID unik untuk setiap kejadian gempa. Kedua, atribut Agency dihapus karena seluruh data memiliki nilai yang sama pada atribut ini, yaitu BMKG. Ketiga, atribut Date Time dianggap tidak memberikan pengaruh atau hubungan yang signifikan dalam mengklusterisasikan model dataset kami.

1. Rename Data yang Memiliki Makna Sama

Pada atribut "Location", kami menemukan 11 label, di antaranya terdapat beberapa yang memiliki makna yang serupa, seperti "Minahassa Peninsula, Indonesia" dengan "Sulawesi, Indonesia". Kami memutuskan untuk menggabungkan label-label tersebut, mengingat bahwa semenanjung Minahassa adalah bagian dari wilayah Sulawesi. Oleh karena itu, kami menganggap keduanya sebagai wilayah "Sulawesi". Hal yang serupa dilakukan untuk wilayah "Kepulauan Talaud", yang merupakan bagian dari provinsi Sulawesi Utara; kami melabelinya dengan "Sulawesi" untuk menyederhanakan klasifikasi.

1. Encode Atribut yang Bertipe Data Category

Kami melakukan Encoding pada atribut Mag Type dan Location. Kami menggunakan One Hot Encoding yang mana merupakan cara untuk mengubah informasi kategori menjadi format yang bisa dimengerti oleh mesin, dengan menandai keberadaan atau ketiadaan suatu kategori menggunakan 1 dan 0.

1. Normalisasi Min-Max pada Nilai Atribut Bertipe Data Numerik

Normalisasi Min-Max bertujuan untuk mengubah semua nilai data ke dalam rentang 0 hingga 1, menyusun data agar lebih terpadat dan merata.

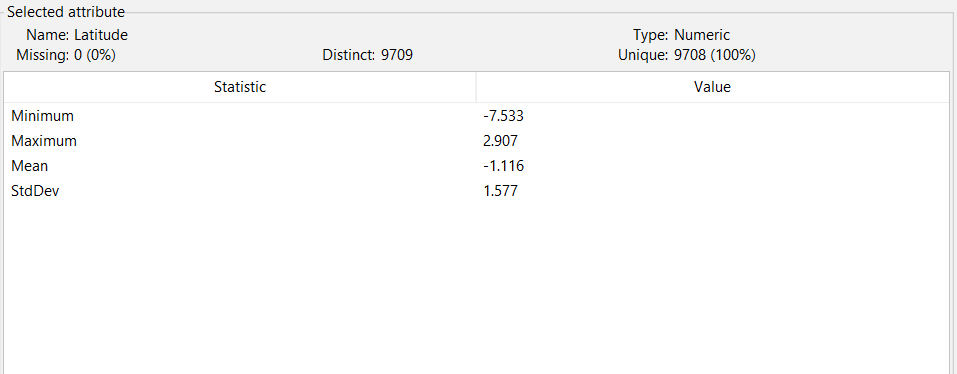
1. Mengkonfirmasi Kebenaran Outlier

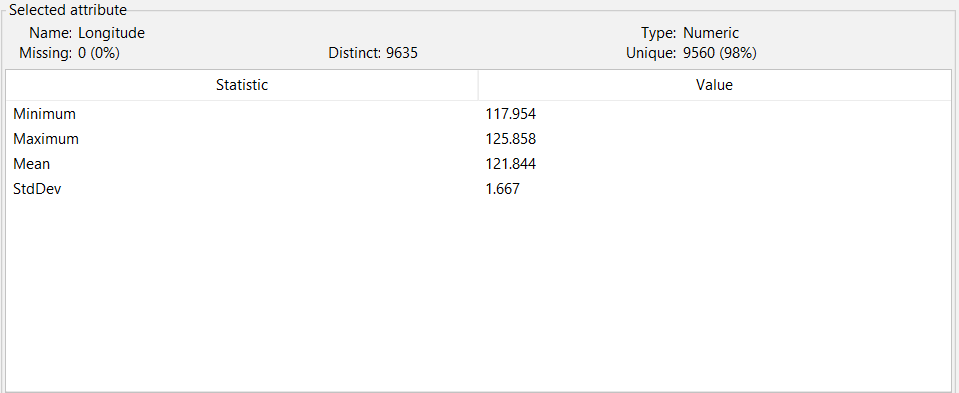
Saat kami menganalisis data mentah, kami menemukan adanya outlier pada beberapa atribut. Salah satu contohnya adalah pada atribut “Depth (km)”, kami menemukan bahwa mayoritas data berada pada angka ratusan hingga ribuan, namun, ada beberapa data minoritas yang hanya memiliki kedalaman dibawah 10km. Setelah kami cari tahu kebenarannya ternyata data data tersebut memang ada, sehingga kami tidak menganggapnya sebuah typo dan kami tidak menghapusnya.



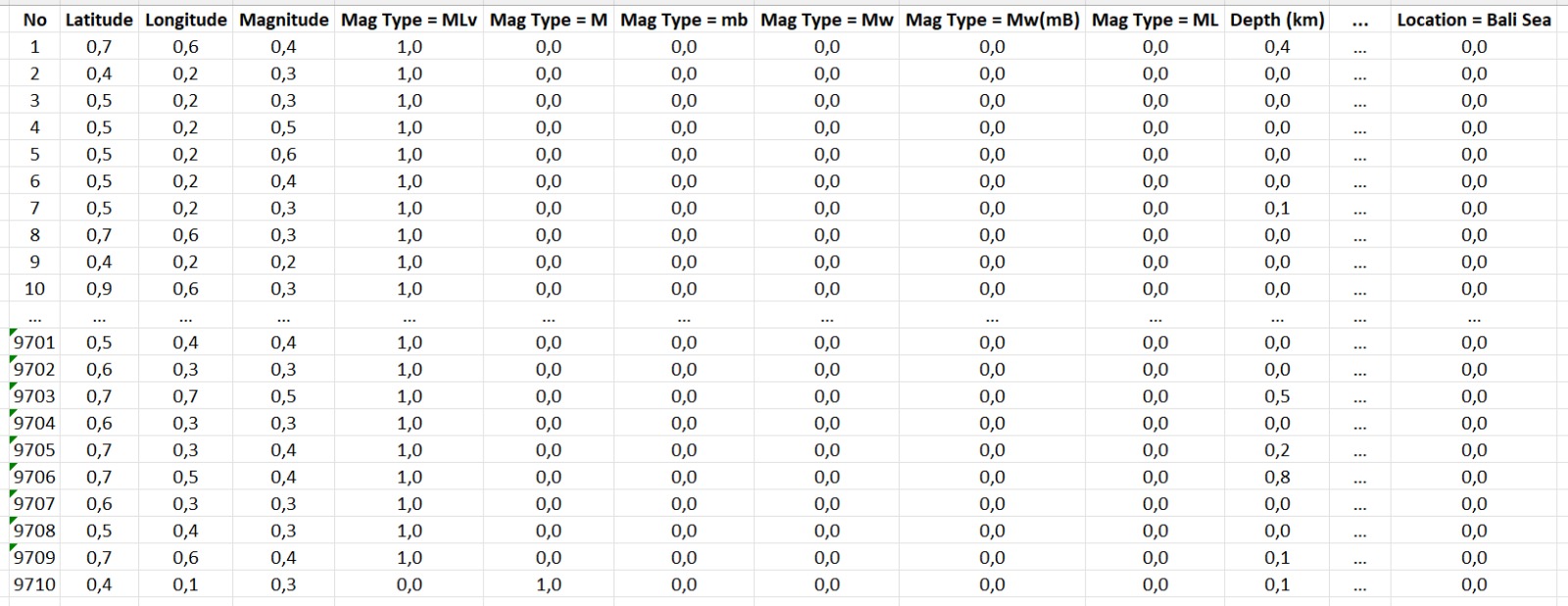
1. Mengecek Kebenaran Data Tiap Atribut

Kami selalu mencari tahu makna setiap atribut dan membandingkannya pada fakta yang ada, contohnya pada atribut Latitude dan Longitude, kami mencocokannya pada kondisi geografis Indonesia yang berada pada 6° LU – 11° LS dan 95° BT – 141° BT.

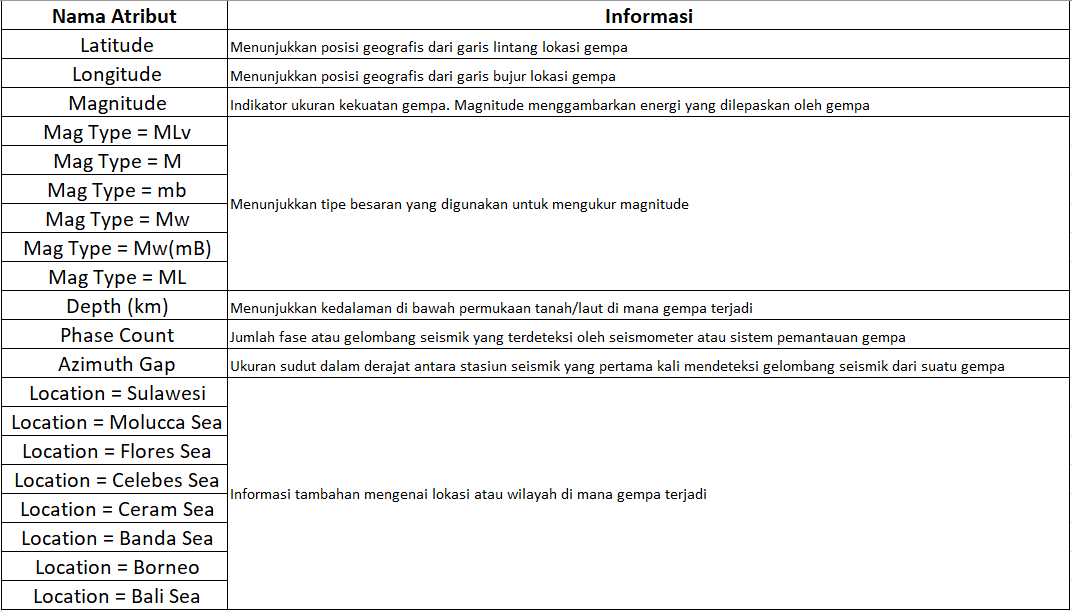




Setelah kami melakukan pre-processing data, terlihat bahwa semua data sudah bertipe data numerik. Selain itu terdapat penambahan atribut akibat dilakukannya One Hot Encoding, atribut mag type terpecah menjadi 6 atribut, dan atribut location terpecah menjadi 8 atribut. Sehingga pada dataset yang bersih terdapat 20 atribut.



Gambar 1.4 Data Setelah Tahap Pre-Processing

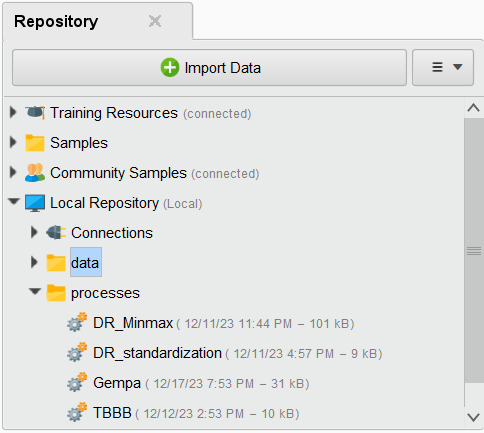


Gambar 1.5 Atribut Data setelah Tahap Pre-Processing

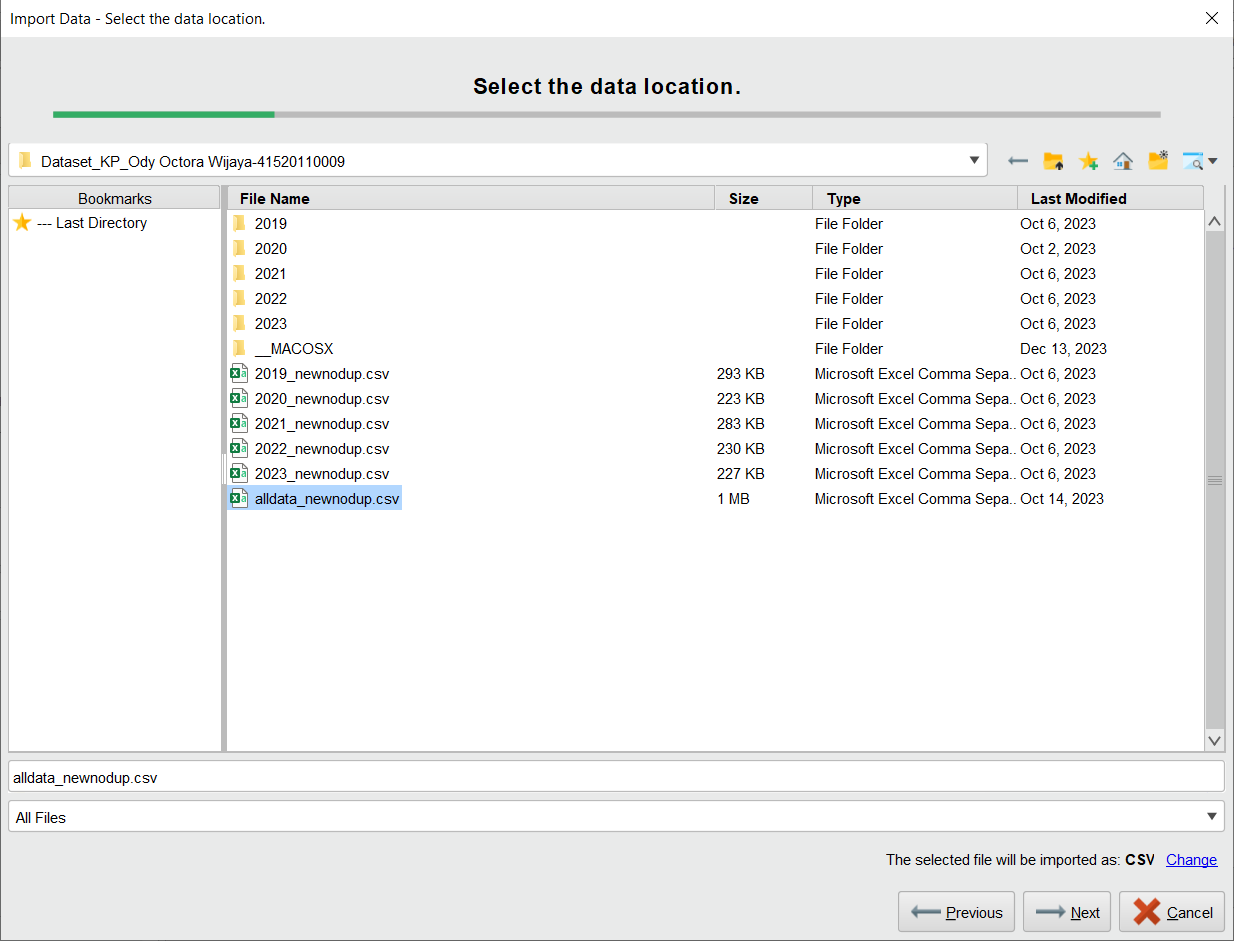
## Langkah-langkah Normalisasi dan Encoding Data

Berikut Langkah-Langkah Normalisasi Data :

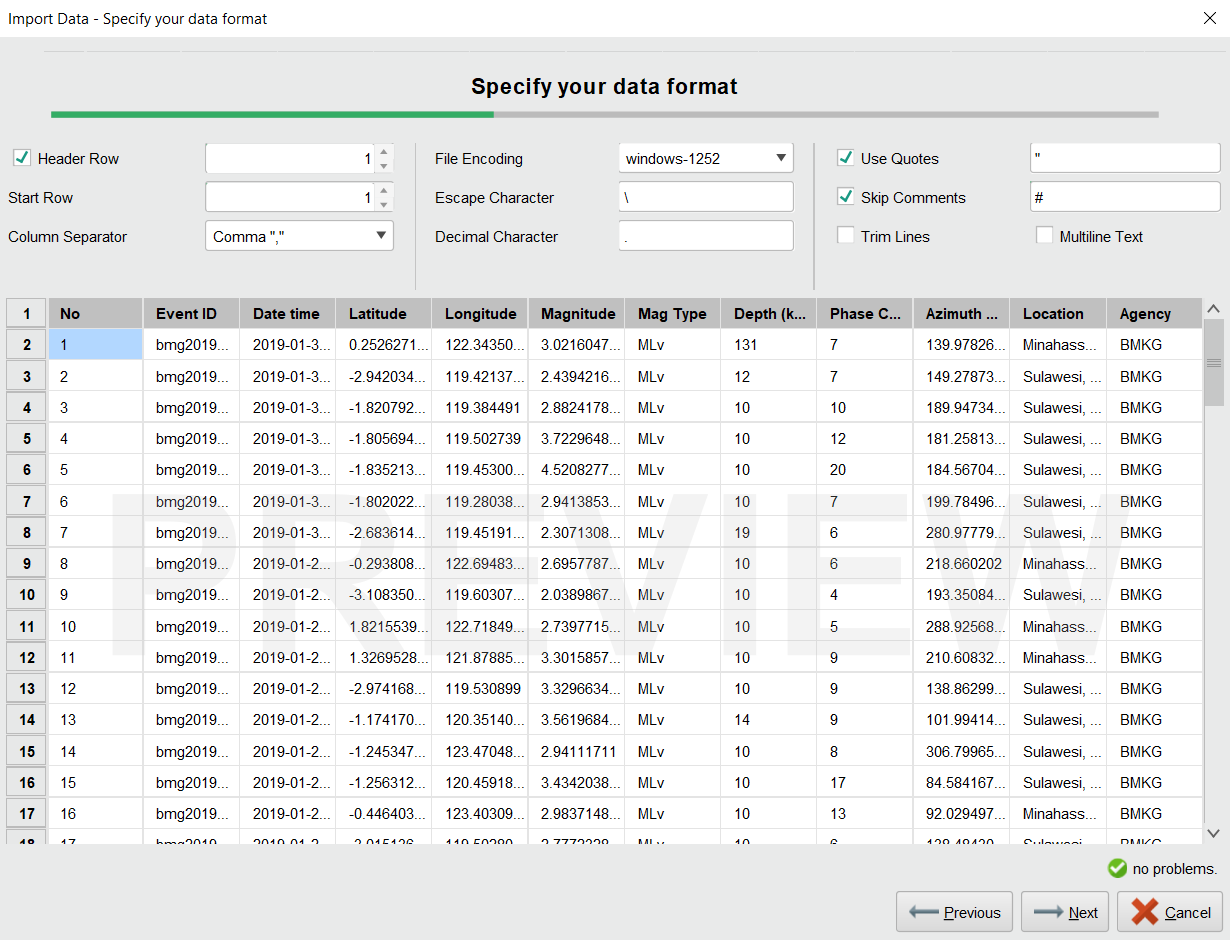
1. Siapkan Dataset, bisa dalam bentuk excel atau csv untuk memudahkan pengimportan. Setelah disiapkan masuk pada aplikasi RapidMiner Studio lalu import data tersebut pada aplikasi.



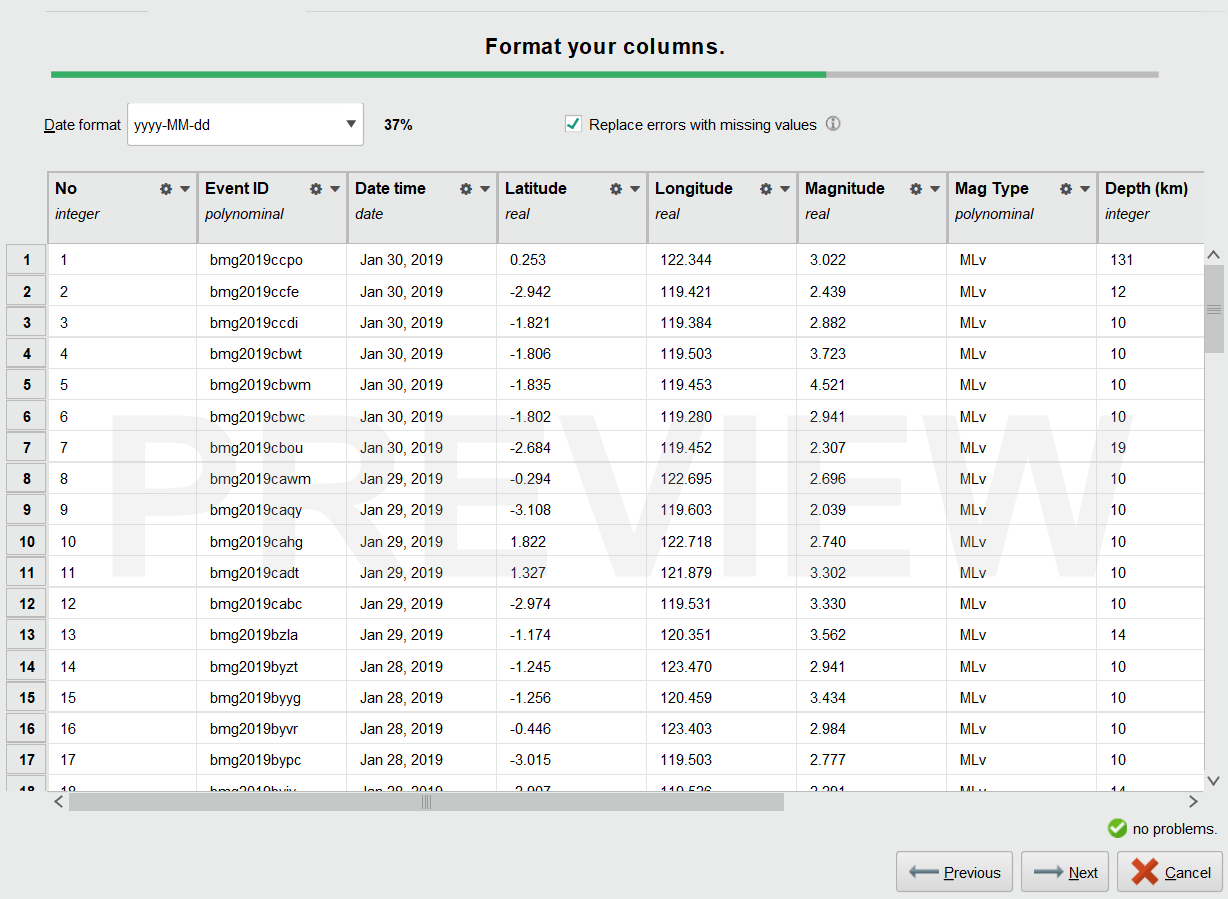
1. Pilih file yang akan diimpor



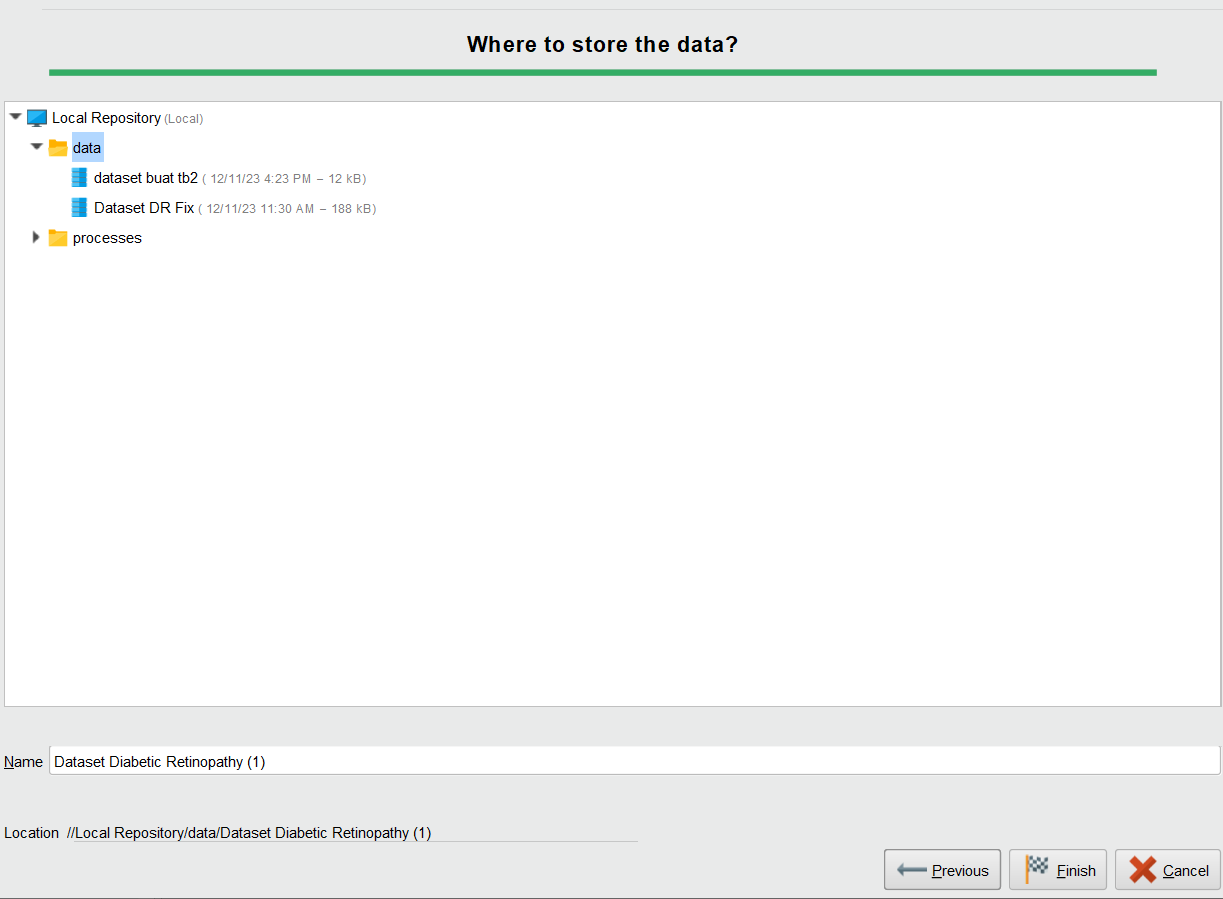
1. Pilih data yang ingin kita olah



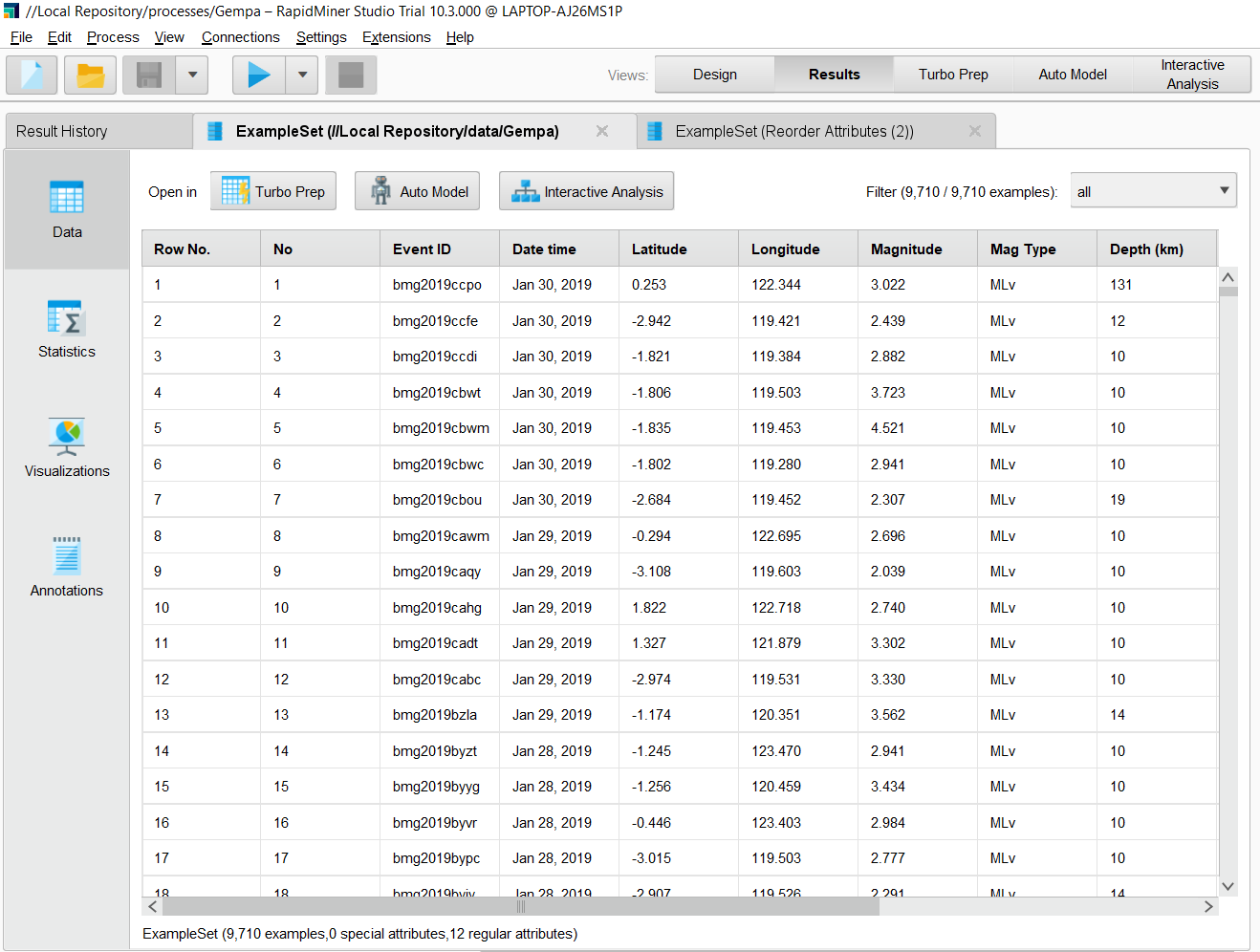
1. Cek kembali jumlah dan kesesuaian data. Periksa juga tipe data atribut dan kelas, atribut dari dataset ini seluruhnya bertipe data integer real, date, dam polynomial. Bila diperlukan kita dapat mengubah role kelas, namun pada dataset ini tidak diperlukan. Selanjutnya, beri ceklist pada Replace errors with missing values.



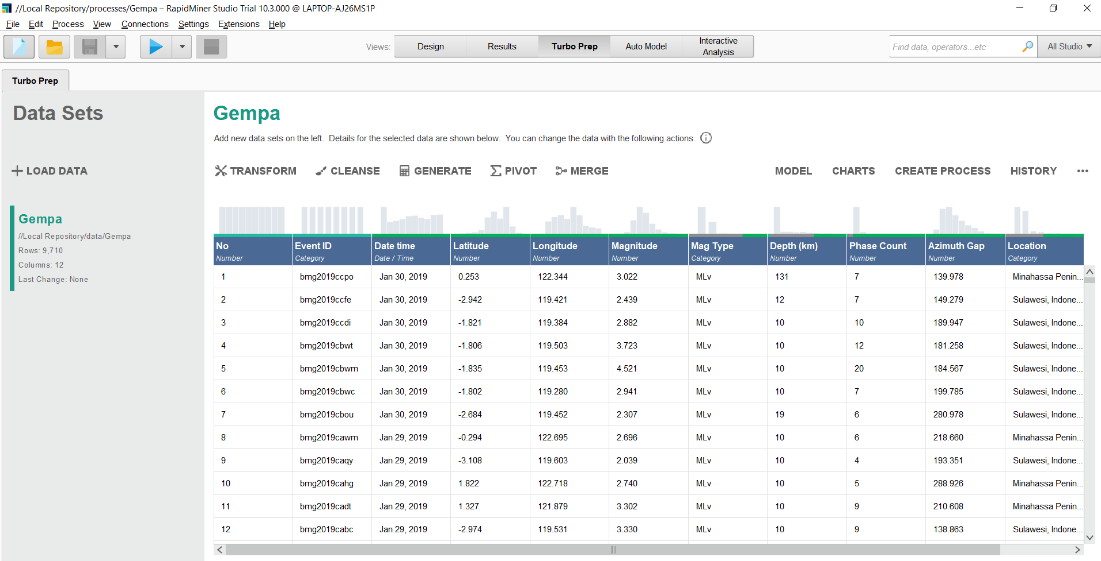
1. Pilih tempat untuk menyimpan dataset tersebut.



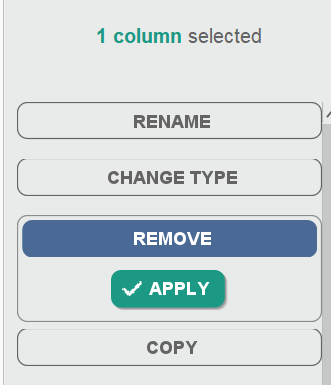
1. Setelah data diimpor, langkah selanjutnya adalah membuat proses-proses dengan Turbo Prep



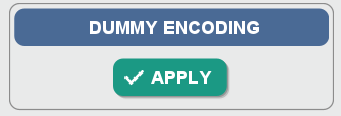
1. Dalam Turbo Prep, terdapat beberapa pilihan, seperti transform, kita dapat mengganti, nama, mengubah tipe, dan menghapus atribut. Pada pilihan cleanse, kita dapat melakukan encoding, menormalisasi dataset, bisa dengan metode auto cleansing, normalisasi manual seperti Min-Max, Standardization, dll.



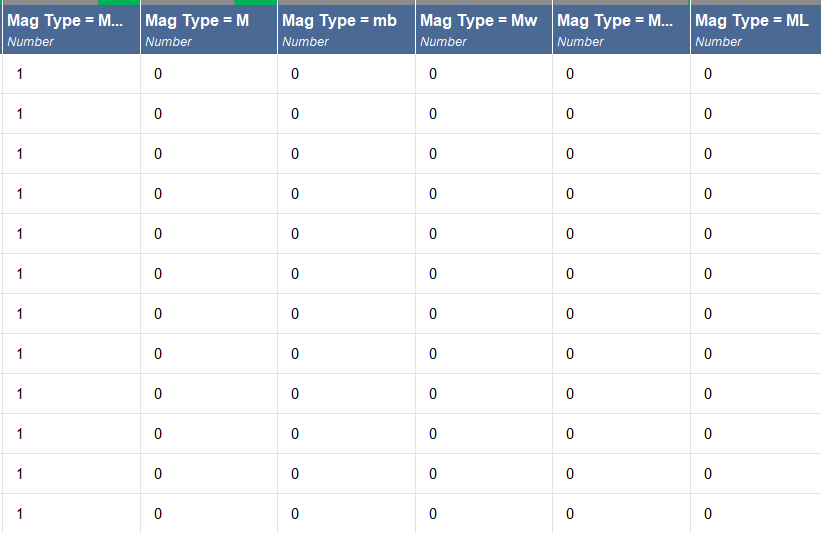
1. Penghapusan atribut dapat dilakukan pada pilihan Transform, lalu pilih atribut yang akan di remove, lalu remove.



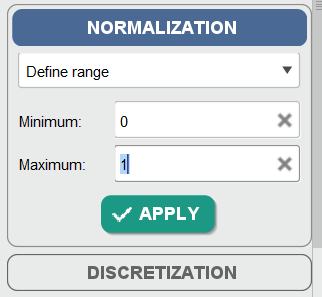
1. Encoding dapat dilakukan pada pilihan Cleanse, lalu pilih Dummy Encoding



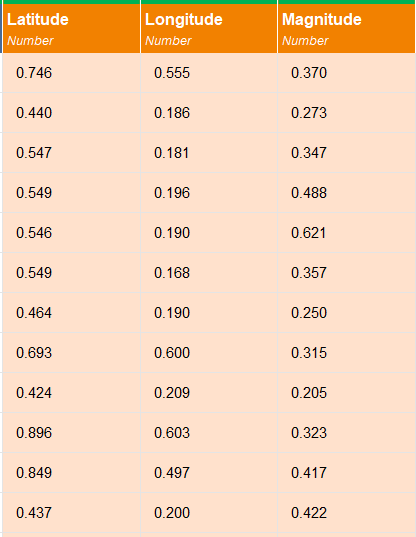
1. Berikut tampilan data setelah di encode



1. Berikut merupakan cara untuk normalisasi Min-Max.



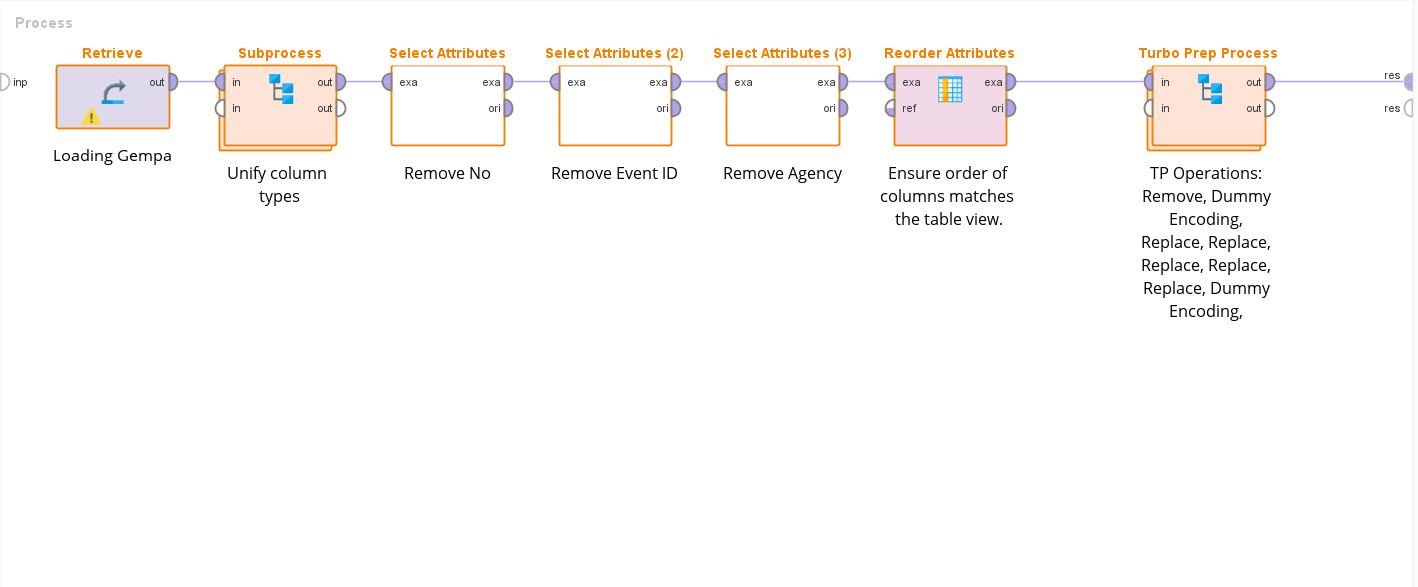
1. Setelah normalisasi, nilai data akan berubah, seperti pada gambar dibawah ini, menggunakan normalisasi Min-Max pada atribut Latitude, Longitude, dan Magnitude, seluruhnya ada pada rentang 0 dan 1.



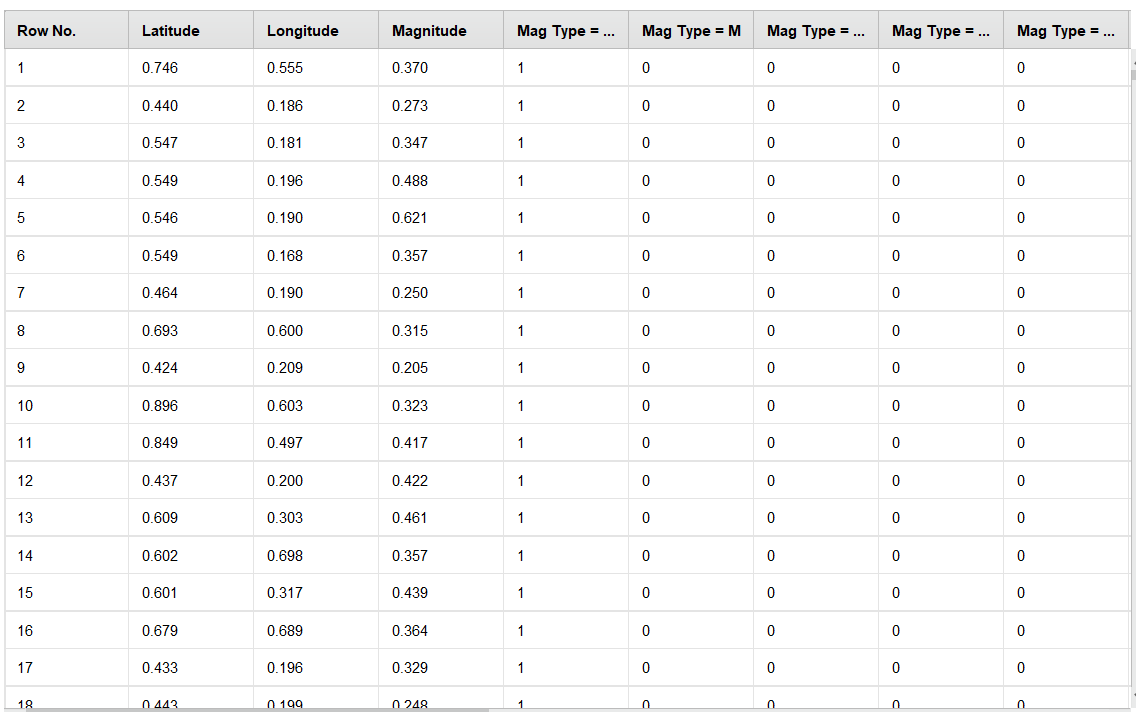
1. Setelah semua atribut dinormalisasi, simpan proses tersebut dengan create process.



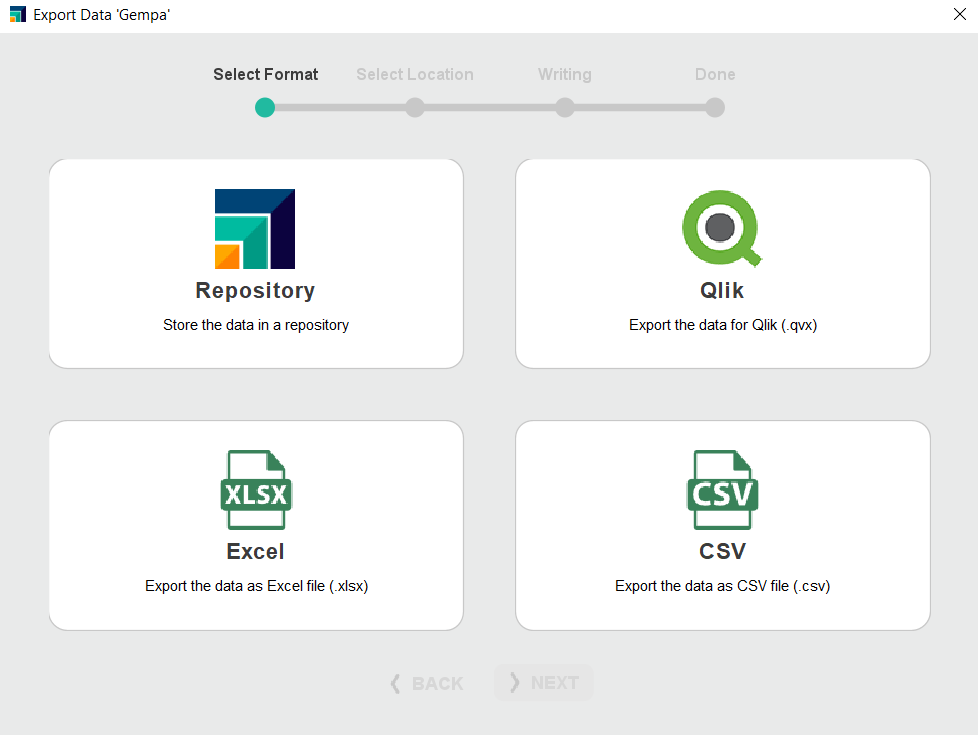
1. Berikut tampilan design pada proses, jika kita ingin menerapkan pada Result, kita dapat memilih semua proses yang ada, kemudian Run.



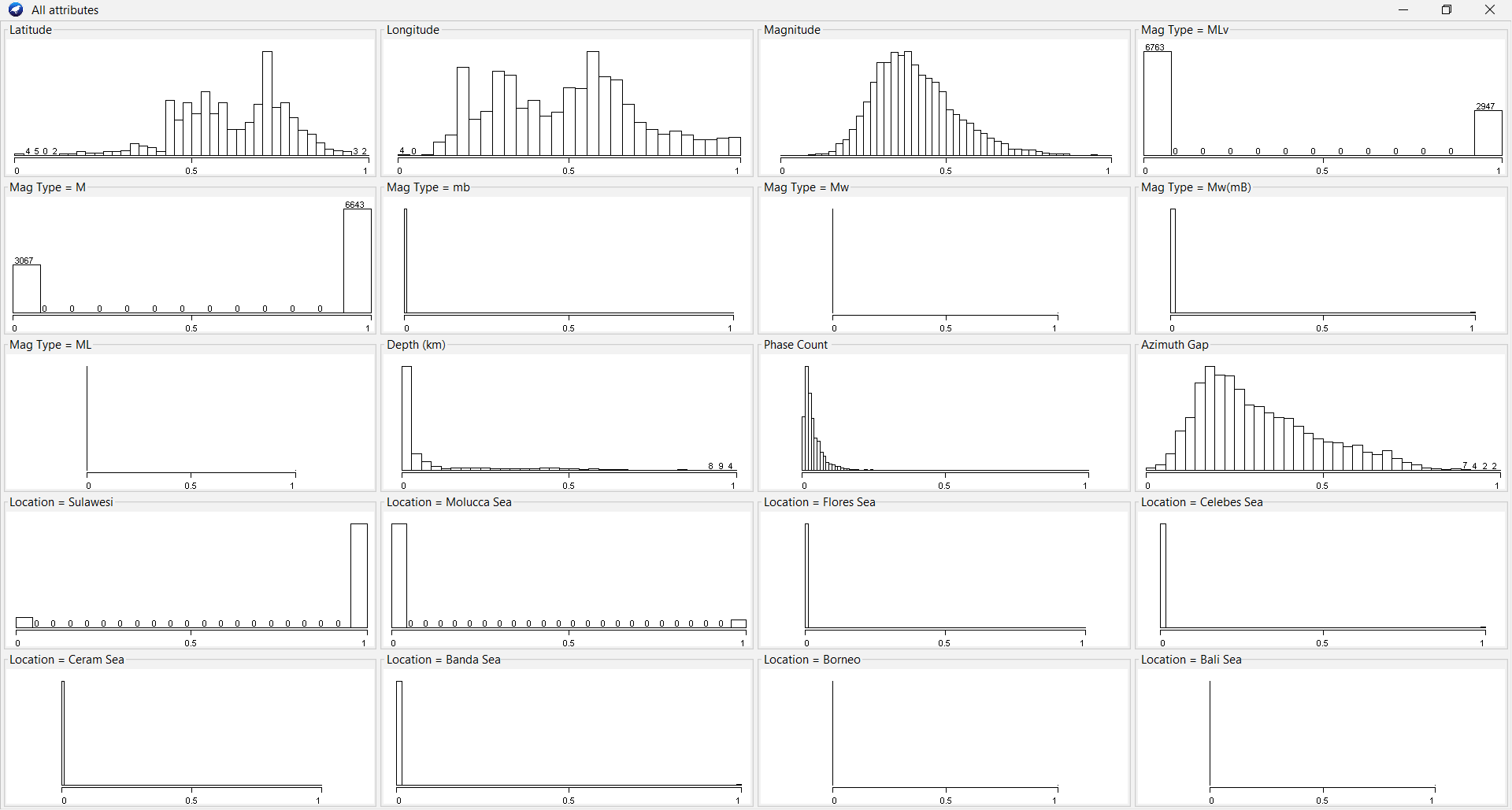
1. Setelah dijalankan, data pada Result akan berubah kedalam bentuk yang telah dinormalisasi.



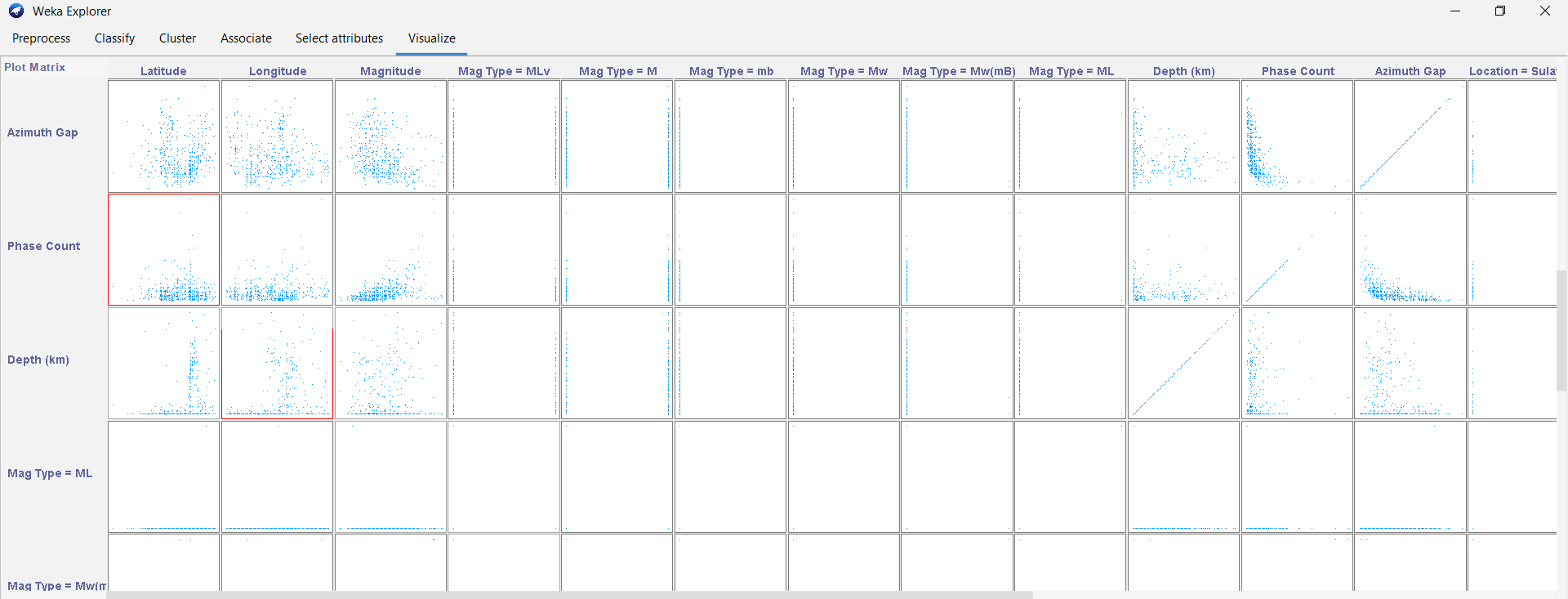
1. Tahap selanjutnya adalah mengekspor file yang sudah dinormalisasi, direkomendasikan menggunakan format .csv agar bisa digenerate pada aplikasi weka saat pengujian dengan algoritma klasifikasi.



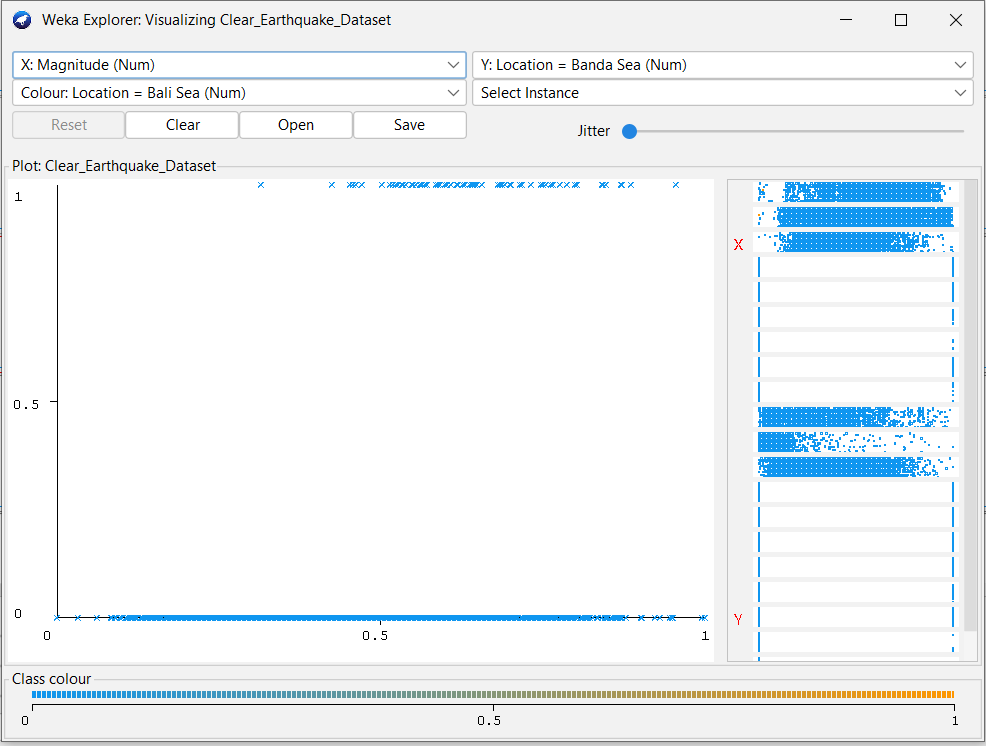
# Analisis Perilaku Data terhadap Kelas dengan Visualisasi

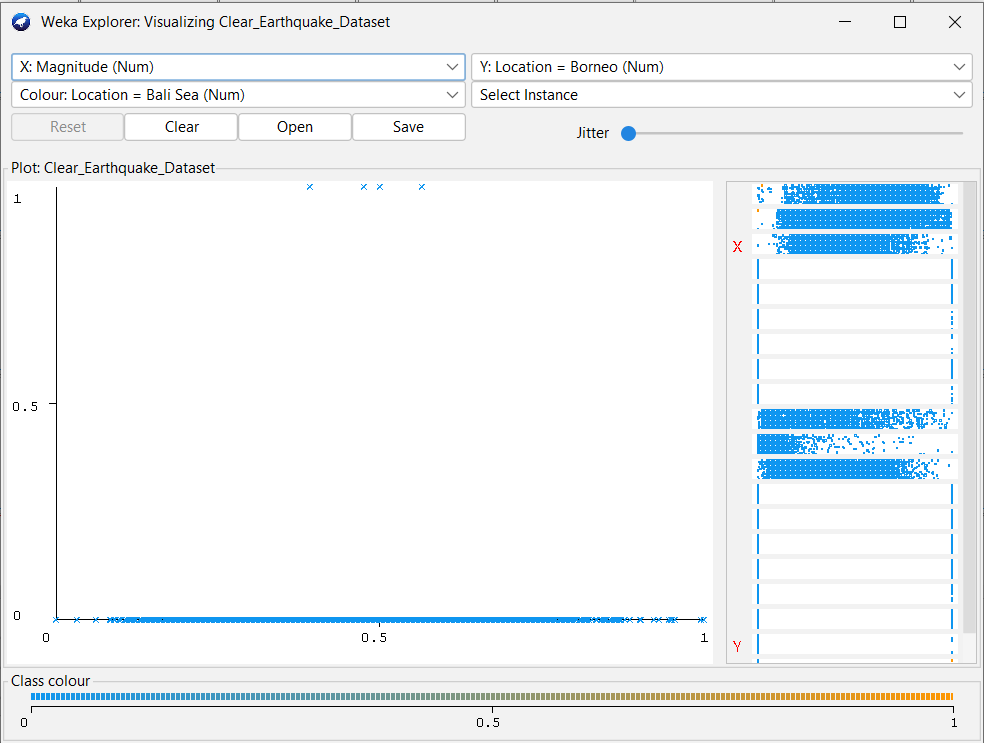


Kumpulan grafik diatas merupakan visualisasi rentang data yang ditunjukkan dengan diagram bar. Terdapat 20 diagram bar yang masing masing menunjukkan pola data tiap atribut.

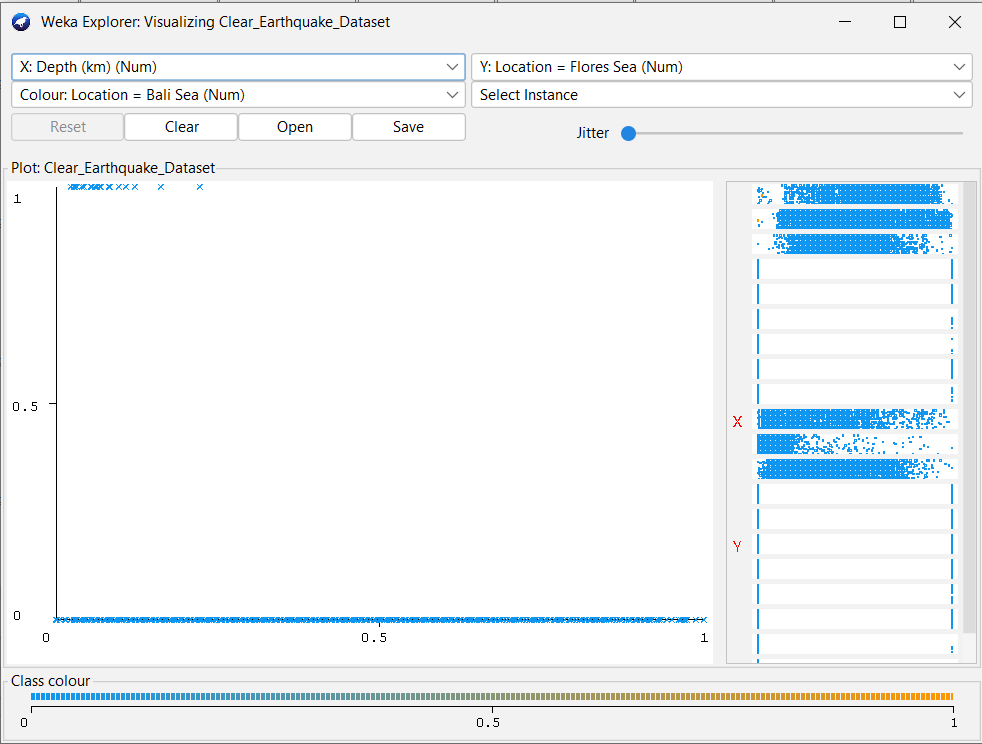


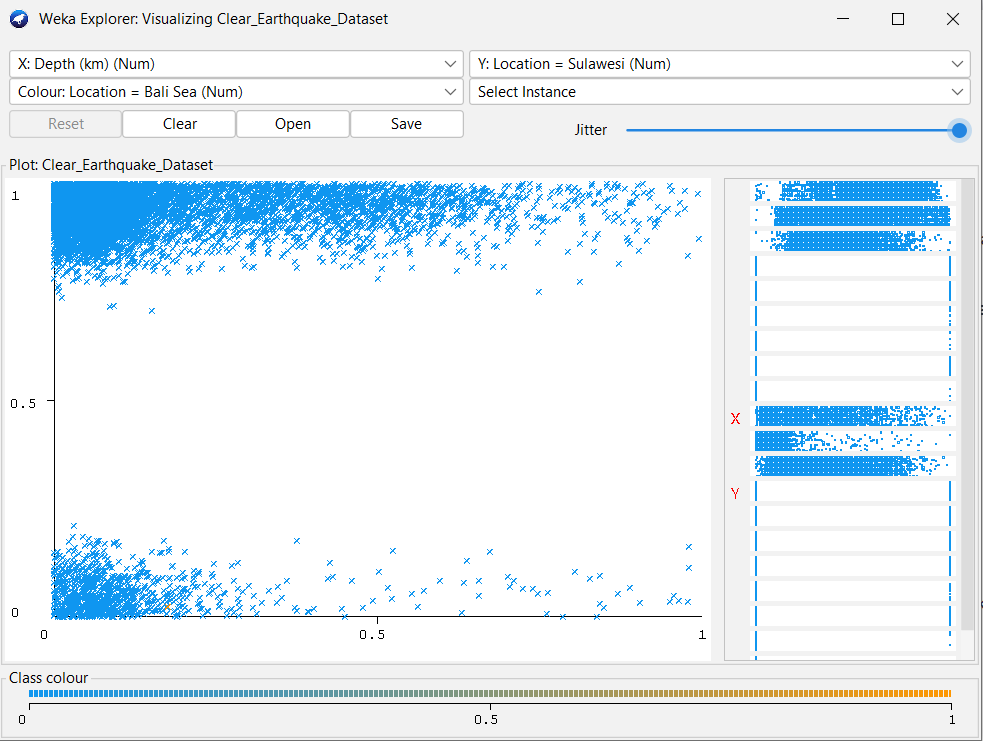
Gambar di atas menampilkan serangkaian visualisasi persebaran berupa titik titik atau biasa yang dinamakan scatter plot atau scatter diagram. Scatter plot menggambarkan korelasi antar atribut. Terdapat 20x20 grafik visualisasi data, masing-masing menunjukkan karakteristik yang unik. Dalam grafik tersebut, teramati beberapa pola penyebaran data, baik yang tercampur maupun yang terpisah.



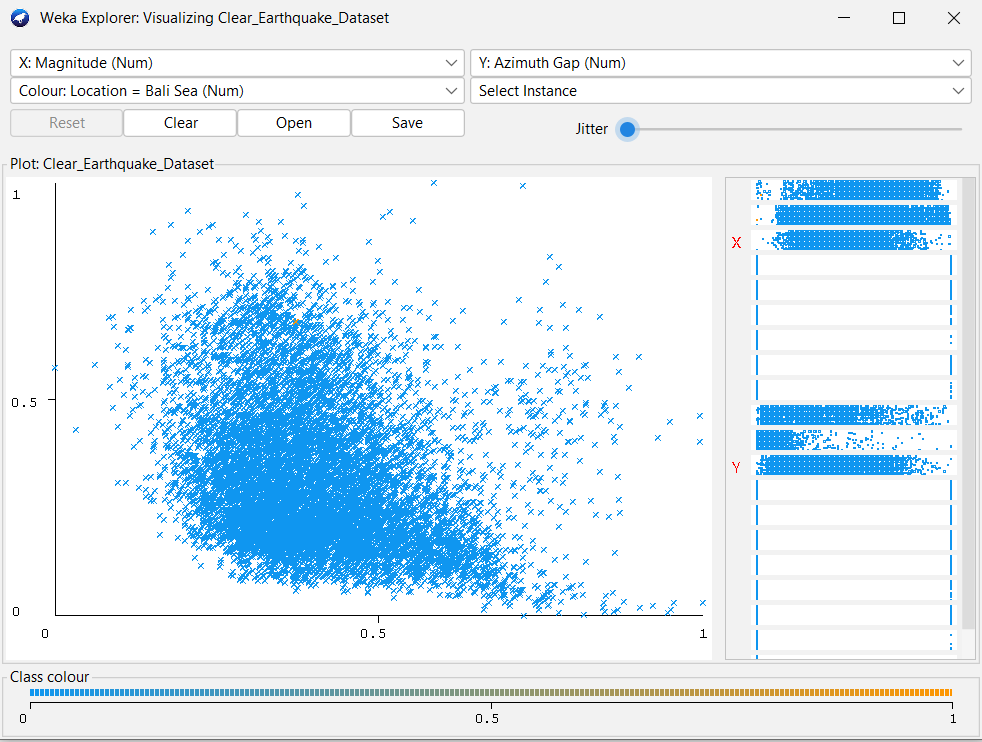


Grafik di atas menggambarkan korelasi antara atribut “Location” dengan atribut “Magnitude”. Dari scatter plot tersebut kita dapat mengetahui perilaku dalam bentuk besarnya gempa berdasarkan wilayah. Salah satu contohnya adalah pada lokasi laut banda, terlihat bahwa rata rata berada pada magnitudo sedang hingga tinggi, sedangkan pada lokasi borneo, terlihat bahwa rata rata besar gempa berkisar antara rendah, ke sedang saja.





Scatter plot diatas menunjukkan sebaran data korelasi antara atribut kedalaman dan juga lokasi. Contohnya, pada gambar pertama menunjukkan bahwa rata-rata gempa yang berada di Laut Flores merupakan gempa dangkal. Sedangkan, gempa yang terjadi di Sulawesi terkonsentrasi pada kedalaman dangkal hingga sedang. Namun, gempa pada kedalaman yang dalam tidak jarang juga terjadi.



Grafik tersebut menunjukkan korelasi antara atribut Magnitude dengan Azimuth Gap. Terlihat bahwa data terkonsentrasi pada kisaran rentang ±0,5. Beberapa lainnya menyebar dan tidak tergabung pada konsentrasi kumpulan mayoritas data.

# Pengujian dan Perbandingan Algoritma

1. **K-Means**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: Clear\_Earthquake\_Dataset

Instances: 9710

Attributes: 20

Latitude

Longitude

Magnitude

Mag Type = MLv

Mag Type = M

Mag Type = mb

Mag Type = Mw

Mag Type = Mw(mB)

Mag Type = ML

Depth (km)

Phase Count

Azimuth Gap

Location = Sulawesi

Location = Molucca Sea

Location = Flores Sea

Location = Celebes Sea

Location = Ceram Sea

Location = Banda Sea

Location = Borneo

Location = Bali Sea

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 24

Within cluster sum of squared errors: 1757.1067324607623

Initial starting points (random):

Cluster 0: 0.334516,0.799603,0.751804,0,1,0,0,0,0,0.023729,0.013015,0.288439,0,0,0,0,0,1,0,0

Cluster 1: 0.730548,0.7344,0.476075,0,1,0,0,0,0,0.369492,0.058568,0.21685,1,0,0,0,0,0,0,0

Cluster 2: 0.313109,0.529864,0.406795,0,1,0,0,0,0,0.023729,0.030369,0.34782,1,0,0,0,0,0,0,0

Cluster 3: 0.68247,0.533725,0.252279,0,1,0,0,0,0,0.288136,0.021692,0.442709,1,0,0,0,0,0,0,0

Cluster 4: 0.845863,0.43352,0.54929,0,1,0,0,0,0,0.118644,0.130152,0.267496,1,0,0,0,0,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1 2 3 4

(9710.0) (757.0) (2285.0) (2775.0) (108.0) (3785.0)

==========================================================================================

Latitude 0.6146 0.6988 0.74 0.5917 0.6442 0.538

Longitude 0.4922 0.8787 0.6298 0.4592 0.6456 0.3516

Magnitude 0.4081 0.5149 0.4326 0.3857 0.7434 0.3789

Mag Type = MLv 0.3035 0.2299 0 0.9993 0 0

Mag Type = M 0.6841 0.7688 0.9996 0 0 0.9979

Mag Type = mb 0.0013 0.0013 0 0 0.1111 0

Mag Type = Mw 0.0004 0 0.0004 0 0.0093 0.0005

Mag Type = Mw(mB) 0.0098 0 0 0 0.8796 0

Mag Type = ML 0.0008 0 0 0.0007 0 0.0016

Depth (km) 0.1185 0.0856 0.2883 0.1024 0.212 0.0317

Phase Count 0.0439 0.0542 0.0418 0.0333 0.4678 0.0389

Azimuth Gap 0.3222 0.2784 0.3116 0.3595 0.0764 0.3171

Location = Sulawesi 0.9056 0 0.9899 0.9787 0.7315 0.9871

Location = Molucca Sea 0.0714 0.8864 0 0 0.2037 0

Location = Flores Sea 0.0041 0 0 0.0068 0 0.0055

Location = Celebes Sea 0.0085 0.0198 0.0101 0.0076 0 0.0063

Location = Ceram Sea 0.0022 0.0238 0 0.0011 0 0

Location = Banda Sea 0.0077 0.07 0 0.0054 0.0648 0

Location = Borneo 0.0004 0 0 0 0 0.0011

Location = Bali Sea 0.0001 0 0 0.0004 0 0

Time taken to build model (full training data) : 0.19 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 757 ( 8%)

1 2285 ( 24%)

2 2775 ( 29%)

3 108 ( 1%)

4 3785 ( 39%)

Hasil dari analisis k-Means clustering menunjukkan bahwa data gempa bumi telah berhasil dikelompokkan menjadi lima cluster. Distribusi instance dalam masing-masing kelompok menggambarkan proporsi relatif dari setiap cluster terhadap total data, dengan Cluster 4 menjadi kelompok terbesar mencakup 39% dari total data. WCSS sebesar 1757.1067324607623 memberikan gambaran sejauh mana data dalam masing-masing cluster tersebar dari centroidnya, di mana nilai yang lebih rendah mengindikasikan kekompakan dan homogenitas cluster. Analisis centroid pada setiap atribut memberikan wawasan tentang karakteristik khusus dari setiap kelompok. Sebagai contoh, Cluster 3 menonjol dengan jumlah instance yang relatif kecil (1.1% dari total data) namun memiliki nilai Magnitude yang tinggi (0.7434)

Hasil evaluasi pada data latihan menunjukkan distribusi clustered instances di dalam kelompok-kelompok hasil analisis k-Means. Terdapat lima cluster yang terbentuk, dan distribusi instance dalam masing-masing cluster memberikan gambaran proporsi relatif dari setiap kelompok. Cluster 4 mencapai proporsi tertinggi dengan 39% dari total data, sementara Cluster 3 memiliki kontribusi paling kecil hanya 1%

1. **Expectation-Maximization**

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N 5 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Relation: Clear\_Earthquake\_Dataset

Instances: 9710

Attributes: 20

Latitude

Longitude

Magnitude

Mag Type = MLv

Mag Type = M

Mag Type = mb

Mag Type = Mw

Mag Type = Mw(mB)

Mag Type = ML

Depth (km)

Phase Count

Azimuth Gap

Location = Sulawesi

Location = Molucca Sea

Location = Flores Sea

Location = Celebes Sea

Location = Ceram Sea

Location = Banda Sea

Location = Borneo

Location = Bali Sea

Test mode: evaluate on training data

=== Clustering model (full training set) ===

EM

==

Number of clusters: 5

Number of iterations performed: 1

Cluster

Attribute 0 1 2 3 4

(0.07) (0.46) (0.03) (0.15) (0.28)

=================================================================

Latitude

mean 0.698 0.5748 0.615 0.7339 0.5952

std. dev. 0.1707 0.1439 0.2516 0.0664 0.1317

Longitude

mean 0.8296 0.4041 0.7433 0.6172 0.4567

std. dev. 0.1812 0.1698 0.2239 0.1242 0.1734

Magnitude

mean 0.5262 0.3836 0.575 0.4378 0.3831

std. dev. 0.1156 0.1171 0.1485 0.1232 0.1042

Mag Type = MLv

mean 0.0474 0 0.6052 0 1

std. dev. 0.2125 0.4598 0.4888 0.0051 0.4598

Mag Type = M

mean 0.9526 0.9973 0.0722 0.9999 0

std. dev. 0.2125 0.0515 0.2589 0.0082 0.4649

Mag Type = mb

mean 0 0 0.0388 0 0

std. dev. 0.0366 0.0366 0.1932 0.0366 0.0366

Mag Type = Mw

mean 0 0.0009 0 0 0

std. dev. 0.0203 0.0298 0.0203 0.0203 0.0203

Mag Type = Mw(mB)

mean 0 0 0.2837 0 0

std. dev. 0.0984 0.0984 0.4508 0.0065 0.0984

Mag Type = ML

mean 0 0.0018 0 0 0

std. dev. 0.0287 0.0421 0.0287 0.0287 0.0287

Depth (km)

mean 0.1111 0.0344 0.118 0.4073 0.1025

std. dev. 0.1827 0.0297 0.1407 0.1826 0.1648

Phase Count

mean 0.0499 0.0371 0.1969 0.0486 0.0324

std. dev. 0.0457 0.032 0.2275 0.0497 0.0332

Azimuth Gap

mean 0.3032 0.3274 0.2075 0.272 0.3598

std. dev. 0.1655 0.1696 0.1589 0.1251 0.1796

Location = Sulawesi

mean 0.0096 1 0.2393 0.9991 0.9996

std. dev. 0.0975 0.2925 0.4267 0.0299 0.0192

Location = Molucca Sea

mean 0.7671 0 0.539 0.0009 0

std. dev. 0.4227 0.2575 0.4985 0.0299 0.2575

Location = Flores Sea

mean 0.0026 0 0.1143 0 0

std. dev. 0.0511 0.0641 0.3182 0.0641 0.0641

Location = Celebes Sea

mean 0.1048 0 0.0393 0 0

std. dev. 0.3063 0.0921 0.1943 0.0921 0.0921

Location = Ceram Sea

mean 0.0315 0 0 0 0

std. dev. 0.1747 0.0465 0 0.0465 0.0465

Location = Banda Sea

mean 0.0784 0 0.0681 0 0

std. dev. 0.2687 0.0876 0.2518 0.0876 0.0876

Location = Borneo

mean 0.006 0 0 0 0

std. dev. 0.0772 0.0203 0.0203 0.0203 0.0203

Location = Bali Sea

mean 0 0 0 0 0.0004

std. dev. 0.0101 0.0101 0.0101 0.0101 0.0192

Time taken to build model (full training data) : 1.67 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 613 ( 6%)

1 1063 ( 11%)

2 445 ( 5%)

3 4921 ( 51%)

4 2668 ( 27%)

Log likelihood: 32.04918

Dalam analisis ini, data gempa bumi telah dikelompokkan ke dalam lima cluster melalui satu iterasi proses Expectation-Maximization. Tabel cluster menyajikan rata-rata dan deviasi standar dari beberapa atribut seperti Latitude, Longitude, Magnitude, Mag Type, Depth, Phase Count, Azimuth Gap, dan Location untuk setiap kelompok. Analisis atribut seperti Latitude, Longitude, Magnitude, Depth, Phase Count, dan Azimuth Gap pada setiap cluster membantu memahami karakteristik spasial dan magnitudo gempa bumi di masing-masing kelompok. Distribusi tipe magnitudo (Mag Type) di setiap cluster memberikan wawasan tentang variasi jenis magnitudo dalam kelompok tersebut. Selain itu, atribut Location menunjukkan sejauh mana kontribusi lokasi tertentu terhadap setiap cluster, memberikan informasi tentang pola spasial lokasi gempa bumi dalam analisis ini.

Dari hasil tersebut, terdapat lima cluster yang masing-masing memiliki kontribusi berbeda terhadap total data. Cluster 3 menjadi kelompok terbesar dengan 4921 instances, menyumbang sekitar 51% dari total data, sementara Cluster 4 menyumbang sekitar 27%. Cluster 0, 1, dan 2 masing-masing memiliki kontribusi sekitar 6%, 11%, dan 5% dari total data. Nilai log likelihood sebesar 32.04918 memberikan indikasi bahwa model Expectation-Maximization yang digunakan dapat dengan baik menjelaskan distribusi data dalam kelompok-kelompok tersebut.

1. **Canopy**

=== Run information ===

Scheme: weka.clusterers.Canopy -N 5 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Relation: Clear\_Earthquake\_Dataset

Instances: 9710

Attributes: 20

Latitude

Longitude

Magnitude

Mag Type = MLv

Mag Type = M

Mag Type = mb

Mag Type = Mw

Mag Type = Mw(mB)

Mag Type = ML

Depth (km)

Phase Count

Azimuth Gap

Location = Sulawesi

Location = Molucca Sea

Location = Flores Sea

Location = Celebes Sea

Location = Ceram Sea

Location = Banda Sea

Location = Borneo

Location = Bali Sea

Test mode: evaluate on training data

=== Clustering model (full training set) ===

Canopy clustering

=================

Number of canopies (cluster centers) found: 5

T2 radius: 1.200

T1 radius: 1.500

Cluster 0: 0.613742,0.456626,0.397319,0,1,0,0,0,0,0.12696,0.039854,0.313674,1,0,0,0,0,0,0,0,{5988} <0,1,2>

Cluster 1: 0.595344,0.456654,0.383359,1,0,0,0,0,0,0.102426,0.03259,0.359463,1,0,0,0,0,0,0,0,{2714} <0,1,3>

Cluster 2: 0.731781,0.882283,0.503456,0,1,0,0,0,0,0.081137,0.054545,0.263981,0,1,0,0,0,0,0,0,{496} <0,2,3>

Cluster 3: 0.745545,0.897515,0.483261,1,0,0,0,0,0,0.084941,0.052347,0.267267,0,1,0,0,0,0,0,0,{174} <1,2,3>

Cluster 4: 0.670286,0.556964,0.751974,0,0,0,0,1,0,0.242643,0.515672,0.07111,1,0,0,0,0,0,0,0,{69} <4>

Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 6077 ( 63%)

1 2761 ( 28%)

2 581 ( 6%)

3 193 ( 2%)

4 98 ( 1%)

Dalam analisis menggunakan algoritma Canopy clustering, terdapat lima cluster (canopies) yang terbentuk dengan T2 radius sekitar 1.200 dan T1 radius sekitar 1.500. Setiap cluster memiliki representasi centroidnya, seperti pada contoh Cluster 0 dengan nilai Latitude sekitar 0.613742, Longitude sekitar 0.456626, dan Magnitude sekitar 0.397319, mencerminkan karakteristik rata-rata dari instance-instance dalam cluster tersebut. Distribusi instance dalam masing-masing cluster adalah sebagai berikut: Cluster 0 dengan 6077 instances (63% dari total data), Cluster 1 dengan 2761 instances (28% dari total data), Cluster 2 dengan 581 instances (6% dari total data), Cluster 3 dengan 193 instances (2% dari total data), dan Cluster 4 dengan 98 instances (1% dari total data). Proses membangun model Canopy clustering menggunakan seluruh data latihan memerlukan waktu sekitar 0.07 detik. Hasil ini memberikan pemahaman awal tentang pembentukan cluster dan karakteristik masing-masing kelompok dalam dataset gempa bumi.

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspek** | K-Means | EM | Canopy |
| **General Explanation** | K-Means Clustering adalah algoritma unsupervised learning yang dipakai untuk mengelompokkan dataset yang belum dilabel ke dalam kluster yang berbeda. Simbol K pada K-means Clustering menandakan jumlah kluster yang digunakan. | EM (Expectation-Maximization) adalah sebuah algoritma yang digunakan untuk melakukan clustering atau pengelompokan data dengan pendekatan probabilistik. | Canopy Clustering merupakan salah satu algoritma clustering yang digunakan untuk mengelompokkan data menjadi beberapa grup berdasarkan jarak antara data. |
| **Kategori** | Clustering | | |
| **Dataset Uji** | Dataset Gempa oleh Ody Octoria Wijaya | | |
| **Aplikasi Normalisasi** | RapidMiner Studio | | |
| **Metode Normalisasi** | Min-Max | | |
| **Metode Encoding** | One Hot Encoding | | |
| **Software Uji** | Waikato Environment for Knowledge Analysis | | |
| **Jumlah Kluster** | 5 kali Kluster | | |
| **Langkah Uji** | 1. Buka WEKA 2. Impor dataset normalisasi 3. Cek persebaran data 4. Pilih opsi Cluster 5. Lakukan uji menggunakan Algoritma : 6. K-Means : SimpleKMeans 7. EM ( Expectation-Maximization ): EM 8. Canopy : Canopy 9. Uji dengan jumlah 5 kali cluster 10. Ringkasan Uji | | |
| **Cluster 0** | 757 (8%) | 613 (6%) | 6077 (63%) |
| **Cluster 1** | 2285 (24%) | 1063 (11%) | 2761 (28%) |
| **Cluster 2** | 2775 (29%) | 445 (5%) | 581 (6%) |
| **Cluster 3** | 108 (1%) | 4921 (51%) | 193 (2%) |
| **Cluster 4** | 3785 (39%) | 2668 (27%) | 98 (1%) |

# Penerapan

1. Algoritma K-Means

Analisis clustering menggunakan metode k-Means pada data gempa bumi menghasilkan kelompok-kelompok dengan distribusi instance yang memberikan gambaran proporsi dan ukuran relatif dari masing-masing cluster, di mana Cluster 4 menonjol sebagai kelompok terbesar. WCSS, yang memiliki nilai 1757.1067324607623, memberikan informasi tentang sejauh mana instance di dalam masing-masing cluster mendekati centroidnya; semakin rendah nilai WCSS, semakin kompak dan homogen cluster tersebut. Centroid untuk setiap atribut dalam masing-masing kelompok memberikan wawasan tentang karakteristik khusus, seperti Cluster 3 yang ditandai dengan Magnitude yang relatif tinggi (0.7434), sementara Cluster 0 memiliki Longitude yang tinggi (0.8787).

Evaluasi pada data latihan menghasilkan lima cluster dengan distribusi instance sebagai berikut: Cluster 0 (8%), Cluster 1 (24%), Cluster 2 (29%), Cluster 3 (1%), dan Cluster 4 (39%). Penerapan hasil ini dapat memberikan pemahaman cepat tentang proporsi dan ukuran relatif dari masing-masing kelompok dalam dataset gempa bumi. Sebagai contoh, Cluster 4, sebagai kelompok terbesar, menunjukkan pola atau karakteristik yang dominan dalam data.

1. Algoritma EM

Dalam cluster ke-0, teramati bahwa gempa-gempa bumi cenderung terkonsentrasi pada lintang rata-rata sekitar 0.698 dengan variasi yang signifikan, tercermin dari deviasi standar sekitar 0.1707. Magnitude rata-rata di dalam kelompok ini mencapai sekitar 0.5262, dengan variasi magnitudo yang relatif terbatas, ditunjukkan oleh deviasi standar sekitar 0.1156. Selanjutnya, tipe magnitudo Mw(mB) menunjukkan rata-rata sekitar 0.2837 dengan variasi yang lebih luas, tercermin dari deviasi standar sekitar 0.4508. Secara geografis, lokasi yang dominan dalam cluster ini adalah Molucca Sea, dengan nilai rata-rata sekitar 0.7671 dan deviasi standar sekitar 0.4227, memberikan gambaran terinci tentang pola spasial gempa bumi di dalam kelompok ini.

Distribusi jumlah instance di setiap cluster memberikan wawasan penting tentang ukuran dan proporsi relatif dari masing-masing kelompok dalam analisis clustering. Cluster 3, yang memiliki 4921 instances dan menyumbang sekitar 51% dari total data, menonjol sebagai kelompok dengan kontribusi paling besar, menunjukkan kemungkinan adanya pola yang signifikan dalam data gempa bumi. Meskipun Cluster 0, 1, 2, dan 4 memiliki proporsi yang lebih kecil, mereka tetap memberikan informasi berharga tentang kelompok-kelompok lain dalam data. Selain itu, nilai log likelihood yang tinggi (32.04918) mengindikasikan bahwa model Expectation-Maximization (EM) mampu dengan baik menjelaskan distribusi data dalam cluster-cluster yang dihasilkan

1. Algoritma Canopy

Deskripsi cluster hasil analisis Canopy menunjukkan variasi dan karakteristik yang berbeda dalam dataset gempa bumi. Cluster 0, yang mencakup 63% dari total instances dengan 6077 data, mungkin mencerminkan kelompok gempa bumi dengan ciri-ciri tertentu, seperti koordinat geografis dan magnitudo khusus. Cluster 1 (28% dari total data) mungkin menunjukkan kelompok gempa bumi dengan karakteristik yang berbeda dari Cluster 0, mungkin terkait dengan lokasi atau tipe gempa yang berbeda. Cluster 2 (6% dari total data) mungkin mencakup gempa bumi dengan atribut yang signifikan dan unik dibandingkan dengan cluster lainnya. Cluster 3 (2% dari total data) mungkin mencerminkan kelompok gempa bumi dengan karakteristik khusus yang membedakannya dari cluster lainnya. Cluster 4 (1% dari total data) mungkin mencerminkan kelompok gempa bumi dengan atribut yang sangat berbeda dari kelompok lainnya

Hasil penelitian ini memiliki dampak signifikan dalam memahami pola gempa bumi dan dapat dijadikan landasan untuk pengembangan strategi mitigasi risiko bencana gempa. Melalui pemrosesan data yang cermat dan penggunaan algoritma clustering seperti K-Means, EM, dan Canopy, kita berhasil mengelompokkan pola gempa ke dalam klaster yang memiliki karakteristik berbeda. Pemahaman ini dapat membantu pihak berwenang dalam mengidentifikasi pola gempa secara lebih efisien, memahami variasi geografis dan magnitudo, serta mengidentifikasi wilayah yang lebih rentan terhadap gempa bumi. Selanjutnya, data hasil pengelompokan ini dapat digunakan untuk mengembangkan model prediktif, memperkuat peringatan dini, dan merancang rencana mitigasi bencana yang lebih spesifik. Dengan menyatukan informasi ini, penelitian ini berpotensi memberikan kontribusi positif dalam upaya melindungi masyarakat dan infrastruktur dari risiko gempa bumi yang serius.

# Referensi

Fan, Z., & Xu, X. (2018). “ Application and visualization of typical clustering algorithms in seismic data analysis” . Procedia Computer Science, 151, 171-178. https://doi.org/10.1016/j.procs.2019.04.026

I. H. Rifa, H. Pratiwi, and R. Respatiwulan, "CLUSTERING OF EARTHQUAKE RISK IN INDONESIA USING K-MEDOIDS AND K-MEANS ALGORITHMS," MEDIA STATISTIKA, vol. 13, no. 2, pp. 194-205, Dec.2020.

https://doi.org/10.14710/medstat.13.2.194-205

Ferin Reviantika , Chita Nauly Harahap, Yufis Azhar, “Analisis Gempa Bumi Pada Pulau Jawa Menggunakan Clustering Algoritma K-Means", Jurnal Dinamika Informatika Volume 9, No 1, Februari 2020.

https://jdi.upy.ac.id/index.php/jdi/article/view/70

Nurfidah Dwitiyanti\* , Erlin Windia Ambarsari, Noni Selvia, “Algoritma K-Means untuk Mengelompokkan Hotel di Sekitar Wilayah Indonesia yang Rentan Gempa Bumi”, Vol 4, No 2, November 2023.

https://doi.org/10.30865/resolusi.v4i1.1477

https://www.antaranews.com/berita/3181177/gempa-dangkal-3-km-guncang-muna-barat-sulawesi-tenggara