KLASIFIKASI AREA GEMPA BUMI MENGGUNAKAN ALGORITMA RANDOM FOREST

Ismail
Jurusan Sistem Komputer, Universitas Gunadarma,
Jl Margonda Raya No 100, Pondok Cina, Depok
ismail_muchsin@staff.gunadarma.ac.id

Abstrak

Salah satu informasi yang dibutuhkan oleh warga negara di dunia adalah informasi mengenai kejadian bencana alam khususnya gempa bumi. Kejadian gempa bumi yang telah terjadi dapat diklasifikasikan dengan menganalisis data gempa bumi pada masa lampau. Pada penelitian ini diimplementasikan penggunaan algoritma Random Forest untuk mengklasifikasikan suatu area apakah termasuk ke dalam kelas gempa bumi atau kelas bukan gempa bumi yang terjadi di dunia. Penelitian ini menghasilkan peta area di seluruh dunia yang terjadi gempa bumi berdasarkan data di masa lampau tahun 1965-2016 dari earthquake dataset kaggle. Penelitian ini menggunakan 7 atribut untuk melakukan klasifikasi antara lain *date, time, latitude, longitude, depth, magnitude,* dan *type. Penelitian ini juga menghitung tabel Confusion Matrix yang dihasilkan dari data aktual dan data prediksi yang telah di proses dalam Random Forest Classifier. Hasil Pengujian Testing Dataset klasifikasian wilayah atau area yang terjadi gempa bumi menghasilkan akurasi sejumlah 99.97%. Hasil penelitian ini diharapkan dapat membantu pihak terkait yang menangani kejadian bencana alam khususnya gempa bumi dengan mengklasifikasikan suatu area termasuk gempa bumi atau bukan gempa bumi berdasarkan atribut yang telah ditentukan.*

***Kata Kunci**: Gempa Bumi, Kaggle Dataset, Klasifikasi, Matriks Confusion , Random Forest*

Abstract

One of the information needed by citizens of the world is information about natural disasters, especially earthquakes. Earthquake events that have occurred can be classified by analyzing earthquake data in the past. In this study, the use of the Random Forest algorithm was implemented to classify an area whether it belongs to an earthquake class or a non-earthquake class that occurs in the world. This research produces a map of areas around the world where earthquakes occurred based on data in the past 1965-2016 from earthquake dataset cluster. This study uses 7 attributes to classify including date, time, latitude, longitude, depth, magnitude, and type. This study also calculates the Confusion Matrix table generated from actual data and predictive data that has been processed in the Random Forest Classifier. The Testing Process Results The classification dataset for areas or areas where an earthquake occurred resulted in an accuracy of 99.97%. The results of this study are expected to help related parties who handle natural disasters, especially earthquakes by classifying an area including earthquakes or non-earthquakes based on predetermined attributes.

***Keywords**: Earthquake, Kaggle Dataset, Classification, Confusion Matrix , Random Forest*

PENDAHULUAN

Informasi mengenai gempa bumi yang tersaji saat ini masih bersifat acak [1], sulit dipahami, dan belum terbukti keabsahannya. Gempa bumi merupakan suatu kejadian yang tidak bisa dihindari, namun dampak bencana alam ini dapat dikurangi atau dapat

diminimalisir dengan mengenali penyebab gempa bumi dan mempelajari kejadian gempa bumi yang telah terjadi dengan menganalisis data yang ada. Pengolahan data bencana alam gempa bumi yang umum dilakukan untuk menghasilkan sebuah informasi [2] yaitu menggunakan teknik *data mining* [3]. Salah satu teknik dari *data mining* adalah klasifikasi [4][5]. Klasifikasi digunakan untuk menemukan model fungsi [6] dan mendeskripsikan data ke kelas-kelas berdasarkan data di masa lampau. Data yang telah dikumpulkan akan dipelajari dan dianalisis hubungannya sesuai dengan label atau target yang telah ditentukan. Beberapa penelitian terkait implementtasi algoritma klasifikasi *random forest* dilakukan peneliti terdahulu. Penelitian [7] membuat sistem pengklasifikasian untuk penilaian kredit dengan menggunakan *dataset German Credit*. Metode yang diterapkan pada sistem tersebut adalah *Random Over-under sampling Random Forest* yang dapat meningkatkan kinerja akurasi sebesar 14,1% dengan nilai akurasi sebesar 0,901 atau 90,1%. Penelitian [8] membuat suatu sistem yang dapat mendiagnosis kanker payudara dengan menggunakan 2 metode, yaitu *Support Vector Machine* (SVM) dan *Random Forest.* Hasil dari klasifikasi tersebut dibandingkan seberapa akurat hasil dari akurasinya. Penelitian [9] membuat suatu model pengklasifikasian untuk memprediksi curah hujan dengan menggunakan metode *Random Forest* yang menghasilkan akurasi sebesar 99,45%. Penelitian [10] membuat suatu model pengklasifikasian terhadap faktor-faktor yang mempengaruhi tingkat penerimaan konsumen terhadap mobil menggunakan metode *Random Forest*. Penelitian [11] membuat suatu sistem yang dapat memprediksi waktu memperbaiki *bug* dari laporan *bug* dengan menggunakan praproses penyaringan *dataset* dan algoritma *Random Forest* untuk pembangunan pendekatan prediksi. Pada penelitian ini, diimplementasikan penggunaan algoritma *Random Forest* untuk mengklasifikasikan suatu area apakah termasuk ke dalam kelas gempa bumi atau kelas bukan gempa bumi yang terjadi di dunia. Penelitian ini juga menghasilkan peta area di seluruh dunia yang terjadi gempa bumi berdasarkan data di masa lampau tahun 1965-2016 dari *earthquake dataset kaggle*. Penelitian ini menghitung *Confusion Matrix* yang dihasilkan dari data aktual dan data prediksi yang telah di proses dalam *Random Forest Classifier*. Hasil penelitian ini diharapkan dapat membantu lembaga-lembaga yang menaungi kejadian bencana alam khususnya gempa bumi yang ada di dunia dalam mengklasifikasikan suatu area termasuk gempa bumi atau bukan gempa bumi berdasarkan parameter yang telah ditentukan.

METODE PENELITIAN

Tahapan proses pada penelitian ini dimulai dengan data *input* dari data gempa

bumi pada tahun 1965-2016 di dunia, melakukan pemilihan parameter, mengkonversi waktu, memvisualisasikan peta area yang terjadi gempa bumi, melakukan pembentukan *classifier* yang terdiri dari *training set* dan *testing set*, menghitung akurasi dan tahap akhir melakukan pemodelan klasifikasi

Data Gempa Bumi

Dalam penelitian ini akan menggunakan data gempa bumi pada Tabel 1 yang terjadi pada tahun 1965-2016 di dunia [12]. Data penelitian terdiri dari 21 *field,* yaitu *Date, Time, Latitude, Longitude, Type, Depth, Depth Error, Depth Seismic Stations, Magnitude, Magnitude Type, Magnitude Error, Magnitude Seismic Stations, Azimuthal Gap, Horizontal Distance, Horizontal Error, Root Mean Square, ID, Source, Location Source, Magnitude Source,* dan *Status,* serta jumlah *record* sebanyak 23412. Pada Tabel 1 baris yang diberi warna kuning adalah data sampel gempa bumi tahun 1965-2016 di dunia pada tanggal 5 Januari 1965 pada pukul 18.05.58 dengan *latitude* sebesar -20.579 derajat, *longitude* sebesar -173.972 derajat dengan *type Earthquake* pada kedalaman (*depth*) sebesar 20 km dan *depth error* tidak ada. Atribut penelitian ini menggunakan 7 *field* antara lain *date, time, latitude, longitude, depth, magnitude*, dan *type* dari total dataset awal sejumlah 21 *field*. Parameter *type* ini digunakan sebagai variabel target atau *labeling* pada proses pengklasifikasian gempa bumi.

Tahap *Preprocessing* Klasifikasi Gempa

Tahap awal (*preprocessing*) dalam melakukan klasifikasi terdiri atas beberapa tahapan proses antara lain :

1. Melakukan konversi waktu. Konversi waktu yang digunakan pada penelitian ini adalah *unix time* atau *unix epoch time. Unix time* atau *unix epoch time* adalah sistem untuk menggambarkan suatu titik waktu yang merupakan jumlah detik yang telah berlalu sejak 00:00:00 Kamis, 1 Januari 1970, *Universal Time Coordinated* (UTC). Setiap hari diperlakukan seolah-olah mengandung tepat 86400 detik. Pada sistem di mana waktu Unix disimpan sebagai bilangan bulat 32 bit yang telah ditandatangani, nilai terbesar yang dapat direkam adalah 2147483647 (231 - 1), yaitu 03:14:07 Selasa, 19 Januari 2038 UTC. Detik berikutnya, jam akan membungkus ke negatif 2147483648 (−231), yaitu 20:45:52 Jumat, 13 Desember 1901 UTC
2. Melakukan proses visualisasi Peta Area Gempa. Setiap wilayah memiliki titik koordinatnya masing-masing yang terdapat pada peta dunia. Titik koordinat tersebut dilewati dengan 2 garis yang saling berkesinambungan, yaitu garis lintang yang disebut dengan *latitude* dan garis bujur yang disebut dengan *longitude*. Peneliti menggunakan garis lintang (*latitude*) dan garis bujur (*longitude*) untuk melakukan

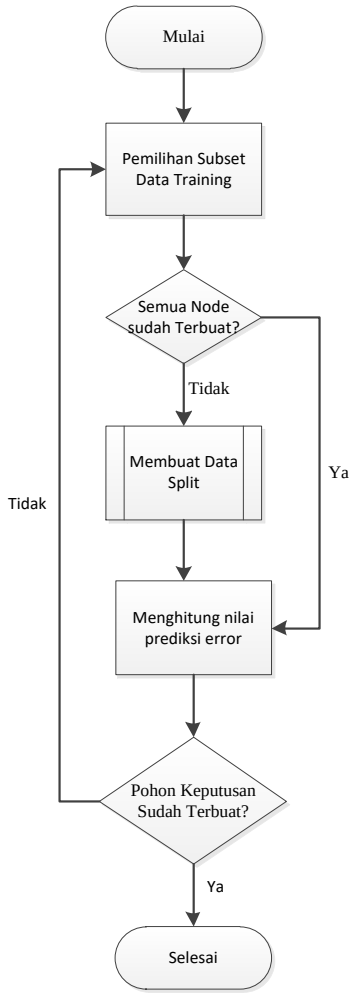
proses visualisasi area gempa pada peta dunia.

Tahap Pembentukan *Classifier*

Tahapan dari pembentukan *Classifier* adalah menentukan variabel dari data *input* yang dijadikan sebagai parameter dalam proses pengklasifikasian dan variabel target (*label*) dengan *base learner*, yaitu *decision tree* seperti dapat dilihat pada Gambar 1.

Tabel 1. Contoh Lima Sampel Data Gempa Bumi Tahun 1965-2016 di Dunia[12]

Date	Time	Latitude	Longitude	Type	Depth	Depth Error
01/02/1965	13:44:18	19.246	145.616	Earthquake	131.06	-
01/04/1965	11:29:49	1.863	127.352	Earthquake	80	-
01/05/1965	18:05:58	-20.579	-173.972	Earthquake	20	-
01/08/1965	18:49:43	-59.076	-23.557	Earthquake	15	-
01/09/1965	13:32:50	11.938	126.427	Earthquake	15	-



Gambar 1. Alur Pembentukan *Classifier* dengan *Random Forest*

Tahap pembentukan *Classifier* penelitian ini terdiri dari :

1. Pemilihan *subset data training* untuk membentuk *node-node* pada *tree*. Jika variabel subset tersebut belum terbuat, maka harus terlebih dahulu melakukan sampling data dengan teknik *bootstrapping* dengan melakukan pengacakan pada dataset kemudian menghitung nilai Gini Index dan selanjutnya mengurutkan variabel subset tersebut dan memilih split terbaik. Pada pembentukan *Random Forest* menggunakan nilai *Gini Index* untuk menentukan pemilah (*split*) yang dijadikan *root/node* berdasarkan tingkat kehomogenan nilai peubah respon (variabel dependen). Setiap *node* akan memilah dan membentuk *node* baru lagi sehingga membentuk *tree* sampai semua variabel yang ada terpenuhi. *Gini index* (S) dapat dihitung menggunakan persamaan 1.

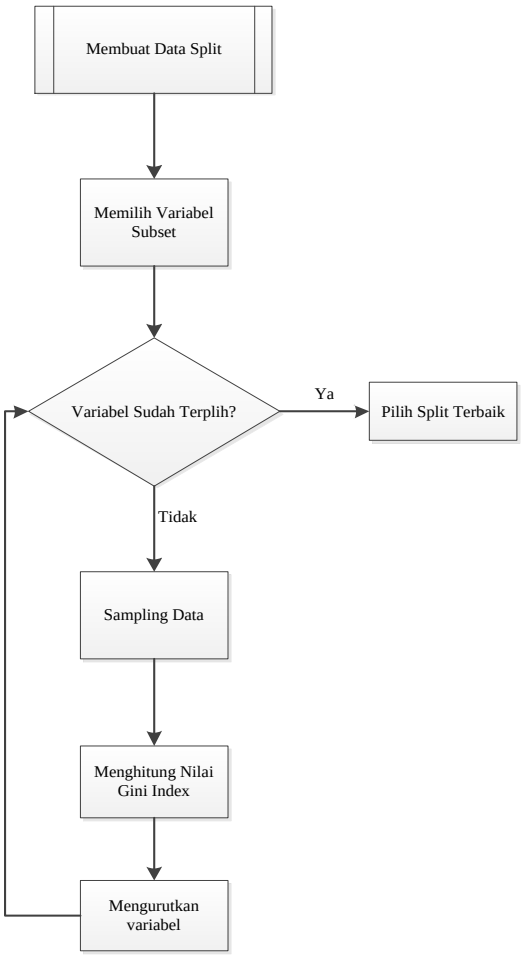
$$\text{Gini(S)} = 1 - \sum_{i=1}^{k} Pi^2 \qquad (1)$$

Nilai *Pi* adalah probabilitas dari *Gini Index* (S) yang dimiliki kelas *I* dan *k* adalah banyaknya nilai atribut yang termasuk ke dalam suatu kelas berdasarkan atribut data. Hasil nilai *Gini index* (S) dapat dilihat pada Tabel 2. Tabel 2 merupakan tabel dari hasil perhitungan nilai *Gini Index* (S) dimana *feature Depth* memiliki nilai *Gini Index* (S) sebesar 0.758921, *feature Latitude* memiliki nilai *Gini Index* (S) sebesar 0.159995, *feature Longitude* memiliki nilai *Gini Index* (S) sebesar 0.080344, dan *feature Magnitude* memiliki nilai *Gini Index* (S) sebesar 0.000740. *Feature name* yang dijadikan *root* adalah *feature Depth*.

2. Membuat *split* antara *data training* dan *data testing*. Pada tahap split data ini *data training* dipilah dan dipisahkan dengan *data testing*. Pencarian *node* akan terus dilakukan sampai semua *node* pada setiap tree terbentuk. Jika semua *node* pada *tree* telah terbuat kemudian dihitung nilai *error prediction*. Pohon-pohon keputusan ini kemudian dikombinasikan setiap pohon atau kelasnya untuk dilakukan *voting*. Pemilihan *voting* terbaik akan dijadikan sebagai hasil dari klasifikasi. Gambar 2 merupakan *flowchart* dalam pembuatan data split.

Tabel 2. Hasil Nilai *Gini Index* (S)

Feature Name	**Nilai *Gini Index* (S)**
Depth	0.758921
Latitude	0.159995
Longitude	0.080344
Magnitude	0.000740



Gambar 2. Flowchart Pembagian Data *Training* dan Data *Testing*

Tahapan pembentukan classifier menggunakan *random forest* dilakukan dengan tahapan :

1. **Pembentukan *Training Set***

   Proses pertama yang dilakukan oleh *Classifier* adalah fase pembelajaran (*learning*), dimana algoritma *Random Forest Classifier* ini dibuat untuk menganalisa data *training* sebagai data masukan lalu direpresentasikan dalam bentuk model klasifikasi dan data *test* yang digunakan untuk memperkirakan akurasi dari model klasifikasi. Data latih yang digunakan pada penelitian ini adalah 18727 atau sekitar 80% dari *dataset* yang ada. Data latih ini digunakan untuk melatih model klasifikasi menghasilkan kemampuan generalisasi yang dapat dipercaya dan tingkat kesalahan yang kecil.

2. **Pembentukan *Testing Set***

   Proses kedua yang dilakukan oleh *Classifier* setelah membagi data latih adalah fase pengujian atau mengevaluasi *output* yang dihasilkan dari algoritma *Random Forest Classifier* untuk mendapatkan hasil akurasi dari proses pengklasifikasian. Data *testing* yang

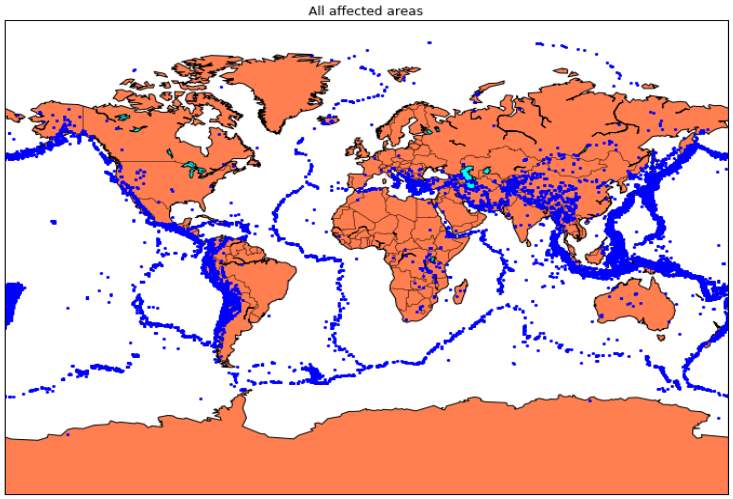
digunakan pada penelitian ini adalah 4682 atau sekitar 20% dari *dataset* yang ada. Data *testing* ini digunakan untuk membandingkan nilai aktual dan nilai prediksi dalam perhitungan tingkat akurasi.

HASIL DAN PEMBAHASAN

Gambar 3 menampilkan hasil dari visualisasi area gempa bumi di dunia. Daerah yang berwarna biru merupakan daerah yang terjadi gempa bumi, daerah yang berwarna merah bata (coral) merupakan daratan, daerah yang berwarna biru muda merupakan danau, dan daerah yang berwarna putih merupakan laut.

Hasil Pengujian Sistem dengan *Confusion Matrix*

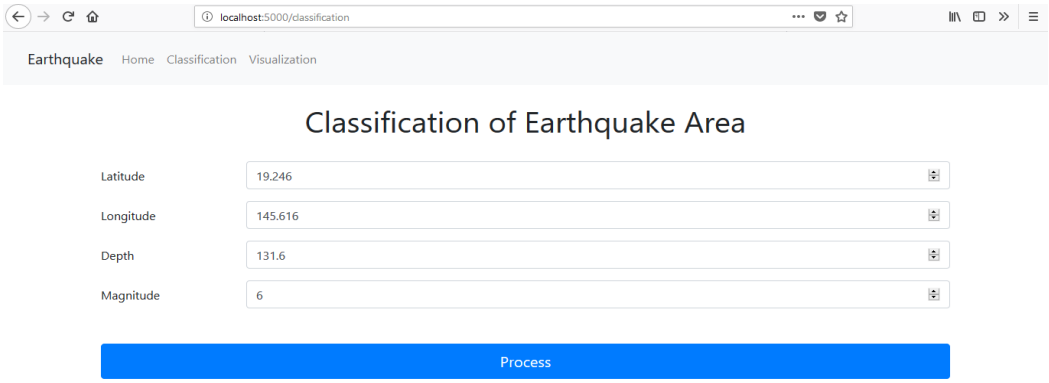
Pada Tabel 3 menampilkan hasil dari pengujian dengan data tes yang berjumlah 5 record. Jenis prediksi (*Predict Type*) memiliki 2 nilai, yaitu 0 atau 1 dimana mendeskripsikan jika jenis prediksi memiliki nilai 0 maka pada *field Description* akan menampilkan "Gempa Bumi" dan sebaliknya jika jenis prediksi memiliki nilai 1 maka pada *field Description* akan menampilkan "Bukan Gempa Bumi".



Gambar 3. Hasil Visualisasi Area Gempa

Tabel 3. Hasil Pengujian dengan *Confusion Matrix*

	Latitude	Longitude	Depth	Magnitude	Predict Type	Description
0	19.246000	145.616000	131.6	6.00	1	Gempa Bumi
1	37.302167	-116.408333	1.2	5.62	0	Bukan Gempa Bumi
2	37.295333	-116.455667	1.2	5.63	0	Bukan Gempa Bumi
3	37.231500	-116.473667	1.4	5.52	0	Bukan Gempa Bumi
4	7.420000	106.030000	10.0	5.20	1	Gempa Bumi



Gambar 4. Hasil Pengujian *Testing Dataset*

Hasil Pengujian *Testing Dataset*

Pengujian dilakukan dengan memasukkan data-data yang ada pada data tes. Data yang dimasukkan antara lain *latitude*, *longitude*, *depth*, dan *magnitude*. Pada pengujian pertama menggunakan *latitude* sebesar 19.246, *longitude* sebesar 145.616, *depth* sebesar 131.6, dan *magnitude* sebesar 6 seperti dapat dilihat pada Gambar 4.

KESIMPULAN DAN SARAN

Implementasi algoritma *Random Forest* pada pengklasifikasian terhadap area atau wilayah yang terjadi gempa bumi di dunia dengan menggunakan data lampau pada tahun 1965-2016 telah berhasil dilakukan. Pemvisualisasian peta area yang terjadi gempa bumi di dunia dari data lampau pada tahun 1965-2016 berhasil dilakukan dalam proses pengklasifikasian dengan menggunakan algoritma *Random Forest.*

Perhitungan *Confusion Matrix* telah berhasil dihitung pada proses pengklasifikasian dengan algoritma *Random Forest* menghasilkan nilai persentase akurasi sebesar 99.97%.

Hasil dari penelitian ini diharapkan dapat membantu lembaga-lembaga yang menaungi kejadian bencana alam khususnya gempa bumi. Pengembangan lebih lanjut dapat dilakukan untuk penelitian selanjutnya dengan penambahan atribut dalam melakukan klasifikasi sehingga akurasi klasifikasi dapat meningkat.

DAFTAR PUSTAKA

[1] G. Otari and R. Kulkarni, "A review of application of data mining in earthquake prediction," *International Journal of Computer Science and Information Technologies*, vol.3, no.2, pp.3570-3574, 2012.

[2] A.S.N. Alarifi, N.S.N. Alarifi, S. Al-Humidan, "Earthquakes magnitude predication using artificial neural network in northern Red Sea area", *Journal of King Saud University-Science*, 24(4), 301-313, 2012.

[3] Han, J., dan Kamber, M, "*Data Mining Concept and Tehniques*", San Fransisco: Morgan Kauffman, 2016.

[4] Iswari, N.M.S, "*Penggunaan Teknik Data Mining untuk Manajemen Resiko Sistem Informasi Rumah Sakit*", ULTIMATICS. Vol. 3, No. 2, pp. 16–22, 2015.

[5] E. Buulolo, N. Silalahi, Fadlina and R. Rahim, "C4.5 Algorithm To Predict the Impact of the", *International Journal of Engineering Research & Technology (IJERT)*, vol. 6, no. 2, pp.10-15, 2017.

[6] S. Mangalathu, H.V. Burton, "Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions", *International Journal of Disaster Risk Reduction*, IJDRR 101111, 6 March 2019.

[7] A. Syukron & A.Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit", *Jurnal Informatika*, 5(2), 175–185, 2018.

[8] H. Aliady, N.J. Tuasikal, & E. Widodo, "Implementasi Support Vector Machine (SVM) dan Random Forest", *Sentika*, 23–24, 2018.

[9] A. Primajaya & B.N. Sari, "Random Forest Algorithm for Prediction of Precipitation", *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, 1(1), 27–31, 2018.

[10] Y.S. Nugroho dan N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest", *Jurnal Teknik Elektro*, 9(1), 24–29, 2016.

[11] N.F. Azhar, & Rochimah, " Memprediksi Waktu Memperbaiki Bug dari Laporan Bug Menggunakan Klasifikasi Random Forest", *Jurnal Sistem Dan Informatika*, Vol. 11(No. 1), 156–164, 2016.

[12] Earthquake Kaggle Dataset, U. G, "Significant Earthquakes, 1965-2016" Available on: https://www.kaggle.com/usgs/earthquake-database#database.csv. Tanggal akses: 05 Juli 2019.