Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

PREDIKSI POTENSIAL GEMPA BUMI INDONESIA MENGGUNAKAN METODE *RANDOM FOREST* DAN *FEATURE SELECTION*

Henri Tantyoko[1*], Dian Kartika Sari[2], Andreas Rony Wijaya[3]
[1,2] Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
[3] Sains Data, Fakultas Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
Email: [1*]henri@ittelkom-pwt.ac.id, [2]dian.kartika@ittelkom-pwt.ac.id, [3]andreas@ittelkom-pwt.ac.id
(* : coressponding author)

Abstrak- Gempa bumi adalah suatu peristiwa alamiah yang terjadi karena adanya pelepasan energi dari dalam kerak bumi secara tiba-tiba, peristiwa tersebut mengakibatkan getaran dan guncangan pada permukaan bumi. Gempa bumi merupakan salah satu bencana alam yang dapat menyebabkan kerusakan fisik yang besar, dampak ekonomi yang signifikan, dan hilangnya nyawa manusia. Beberapa penyebab gempa bumi antara lain aktivitas tektonik lempeng bumi, pergerakan lempeng tektonik, dan deformasi kerak bumi. Untuk mengurangi jumlah korban jiwa, perlu dilakukan prediksi kapan gempa bumi akan terjadi di suatu wilayah. Salah satu cara untuk memprediksi gempa bumi menggunakan metode Machine Learning yaitu Random Forest (RF), metode ini memanfaatkan beberapa pohon keputusan yang selanjutnya dilakukan voting untuk menentukan keputusan akhir prediksi . Model yang baik adalah model yang menghasilkan kesalahan seminimal mungkin. Oleh karena itu, penulis melakukan skema seleksi fitur untuk mengolah fitur-fitur yang memiliki korelasi yang kuat. Prediksi menggunakan RF dengan seleksi fitur menghasilkan F1 score sebesar 92.23%, yang lebih baik 5.02% dibandingkan tanpa menggunakan seleksi fitur. Metode RF + Seleksi Fitur ini juga jauh lebih baik jika dibandingkan metode machine learning tradisional lainnya seperti SVM, Naïve Bayes, dan Decision Tree.

Kata Kunci: Machine Learning, Random Forest, Seleksi Fitur, Korelasi, F1 Score

Abstract- *Earthquake is a natural event that occurs due to the sudden release of energy from within the earth's crust, this event causes vibrations and shocks on the earth's surface. Earthquake is one of the natural disasters that can cause great physical damage, significant economic impact, and loss of human life. Some of the causes of earthquakes include tectonic activity of the earth's plates, movement of tectonic plates, and deformation of the earth's crust. To reduce the number of fatalities, it is necessary to predict when an earthquake will occur in an area. One way to predict is to use the Machine Learning method, namely Random Forest (RF), this method utilizes several decision trees which are then voted on to determine the final prediction decision. A good model is a model that produces minimal errors. Therefore, the author uses a feature selection scheme to process features that have a strong correlation. Prediction using RF with feature selection produces an F1 score of 92.23%, which is 5.02% better than without using feature selection. This RF + Feature Selection method is also much better than other traditional machine learning methods such as SVM, Naïve Bayes, and Decision Tree.*

***Keywords**: Machine Learning, Random Forest, Feature Selection, Correlation, F1 Score*

1. PENDAHULUAN

Gempa bumi merupakan salah satu bencana alam yang dapat menimbulkan kerusakan yang serius terhadap kehidupan manusia dan infrastruktur. Gempa bumi merupakan salah satu bencana alam yang sering terjadi di Indonesia, sebuah negara yang terletak di wilayah cincin api Pasifik. Indonesia terletak di pertemuan tiga lempeng tektonik utama, yaitu Lempeng Indo-Australia, Lempeng Pasifik, dan Lempeng Eurasia, sehingga menjadi salah satu wilayah dengan aktivitas seismik yang tinggi. Indonesia, sebagai negara dengan kompleksitas tektonik yang tinggi, sering mengalami gempa bumi dengan intensitas yang beragam[1], [2]. Dalam upaya mengurangi dampak buruk yang disebabkan oleh gempa bumi, penting untuk mengembangkan metode prediksi yang dapat memberikan informasi mengenai potensi terjadinya gempa bumi di wilayah Indonesia.

Prediksi gempa bumi merupakan bidang yang kompleks dan sulit diprediksi secara tepat. Meskipun penelitian dan teknologi terus berkembang, saat ini tidak ada metode yang dapat memberikan prediksi yang akurat dalam menentukan waktu, lokasi, dan kekuatan gempa bumi di suatu wilayah secara spesifik. Prediksi kapan terjadinya gempa bumi di suatu wilayah perlu dilakukan untuk mengurangi jumlah korban jiwa. Salah satu metode yang dapat digunakan untuk prediksi gempa bumi adalah Random Forest. Random Forest adalah sebuah teknik pemodelan yang menggabungkan sejumlah pohon keputusan (*decision trees*) untuk melakukan prediksi. Kelebihan dari Random Forest adalah kemampuannya dalam menangani ketergantungan dan interaksi antara variabel yang kompleks serta mampu mengatasi overfitting. Selain itu, Random Forest dapat menghasilkan kesalahan yang cukup rendah, kinerja yang optimal dalam klasifikasi, mampu menangani data pelatihan dalam jumlah besar dengan efisien, serta metode yang efektif untuk memperkirakan data yang hilang. [3].

Selain itu, dalam pemodelan prediksi gempa bumi, penting untuk melakukan seleksi fitur yang tepat. Seleksi fitur bertujuan untuk mengidentifikasi variabel-variabel yang paling berpengaruh terhadap terjadinya gempa bumi di Indonesia. Dengan memilih fitur-fitur yang relevan dan informatif, dapat meningkatkan kualitas prediksi dan mengurangi kompleksitas model.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Pada penelitian [3] mengamati hasil prediksi rating untuk aplikasi App Store, algoritma yang digunakan yaitu Random Forest, hasilnya adalah algoritma Random Forest memiliki performa yang paling terbaik dari algoritma yang lain dalam membantu menemukan kelemahan pada dataset *Apple's AppStore*. Dengan hasil akurasi 86.23%, *recall* 84.80%, *precision* 84.45%, dan RMSE 0.316. Dalam penelitian [4] tentang cara membandingkan algoritma SGD, Naïve Bayes, dan Random Forest dengan studi kasus kopi Arabika diperoleh hasil akurasi model menggunakan random forest sebesar 96%. Random forest memiliki akurasi paling tinggi dibandingkan metode yang lain. Selain itu penelitan mengenai random forest yang dilakukan oleh [5] memperoleh hasil Random Forest memiliki nilai akurasi sebesar 97,88%. Algoritma random forest memiliki akurasi tertinggi mengalahkan algoritma lain dalam percobaan. Hasil ini dievaluasi menggunakan kurva *Receiver Operating Characteristic* (ROC) untuk mengetahui ketepatan model yang dibangun.

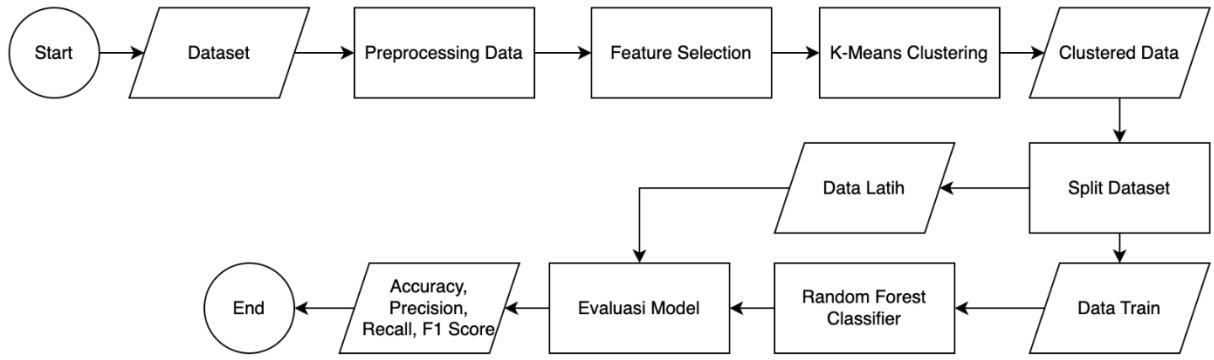
Berdasarkan uraian diatas, penulis akan membuat penelitian mengenai prediksi gempa bumi menggunakan algoritma *random forest*. Diharapkan hasil dari penelitian ini dapat memberikan kontribusi dalam meningkatkan kemampuan prediksi gempa bumi di Indonesia. Dengan menggunakan kombinasi antara Random Forest dan metode seleksi fitur, diharapkan dapat menghasilkan model prediksi yang lebih akurat dan efektif. Informasi tentang potensi gempa bumi di masa depan dapat digunakan untuk membantu pihak berwenang dan masyarakat dalam mengambil langkah-langkah mitigasi yang tepat guna mengurangi dampak buruk yang disebabkan oleh gempa bumi.

Berdasarkan penelitian sebelumnya yang dilakukan, terdapat potensi untuk prediksi gempa bumi dengan penambahan seleksi fitur pada pemodelan *machine learning* menggunakan *Random Forest*. Dengan memadukan Random Forest dan metode seleksi fitur yang tepat, penelitian ini memiliki potensi untuk menghasilkan model prediksi yang lebih akurat dan efektif. Seleksi fitur yang cermat dapat membantu mengidentifikasi variabel-variabel yang paling berpengaruh dalam prediksi gempa bumi, sehingga dapat mengurangi kebisingan dan meningkatkan kinerja model.

Hasil dari penelitian ini dapat memberikan kontribusi dalam peningkatan kemampuan prediksi gempa bumi di Indonesia. Informasi tentang potensi gempa bumi di masa depan yang lebih akurat dapat menjadi landasan untuk pengambilan keputusan yang lebih baik dalam mitigasi risiko. Pihak berwenang dan masyarakat dapat menggunakan informasi ini untuk merencanakan dan mengimplementasikan langkah-langkah mitigasi yang tepat, seperti perencanaan bangunan tahan gempa, evakuasi darurat, dan kesadaran masyarakat.

2. METODE PENELITIAN

Bab ini akan membahas metodologi penelitian yang digunakan dalam penelitian prediksi gempa bumi menggunakan algoritma Random Forest dan metode seleksi fitur. Metodologi penelitian ini akan mencakup langkah-langkah yang diambil dalam mengumpulkan data, mengolah data, mengimplementasikan algoritma, dan menganalisis hasil. Penjelasan gambar alur sistem dapat dilihat pada gambar 1.



Gambar 1. Alur kerja sistem

2.1 Dataset

Dataset adalah kumpulan data yang terorganisir dalam format yang terstruktur. Dataset berisi kumpulan informasi yang berkaitan dengan domain Gempa Bumi Indonesia. Dataset digunakan dalam berbagai bidang seperti ilmu komputer, statistik, ilmu data, dan pembelajaran mesin. Mereka digunakan untuk melatih dan menguji model atau algoritma pembelajaran mesin. Dataset dapat ditemukan dalam berbagai sumber, termasuk basis data, perangkat lunak, situs web, repositori publik, dan penelitian ilmiah.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

2.2 *Preprocessing*

Preprocessing data gempa memuat beberapa proses yaitu filter *magnitude* untuk mengambil data yang hanya memiliki *magnitude* lebih dari 4 karena gempa dapat dirasakan jika *magnitude* lebih dari 4 atau kedalaman yang dangkal karena tujuan peneliti untuk memprediksi gempa bumi yang dirasakan di daratan. Selanjutnya data tersebut dilakukan pengubahan dari string ke numerik agar dapat diproses mesin.

2.3 *Feature Selection*

Feature selection merupakan proses memilih subset fitur yang paling relevan dan informatif dari sekumpulan fitur yang tersedia dalam data. Hal ini sangat penting dalam analisis data dan pembangunan model prediktif. Alasan penggunaan *feature selection* yaitu untuk mengurangi dimensi data, meningkatkan kinerja model, Mengurangi kelebihan informasi, dan sebagai efisiensi. [6], [7].

2.4 *K-Means Clustering*

K-Means Clustering adalah salah satu algoritma *unsupervised learning* yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok atau kluster. Tujuan dari K-Means Clustering adalah untuk meminimalkan jarak antara titik data dalam satu *kluster* dengan pusat *kluster* yang sesuai [8]. Berikut pseudocode untuk K-Means Clustering

Tabel 1. Pseudocode K-Means

K-means clustering

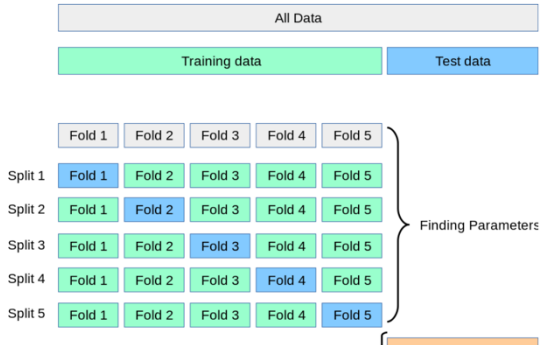
Input : Titik data lokasi dan nilai k centroid
Output: Cluster setiap baris
Step:

1. Inisialisasi secara acak k centroid.
2. Menghubungkan setiap titik data di D dengan centroid terdekat. Membagi titik data menjadi k cluster.
3. Hitung ulang posisi dari centroid.
4. Ulangi langkah 2 dan 3 sampai tidak ada lagi perubahan keanggotaan titik data.

2.5 *Split Dataset*

Split dataset digunakan untuk memisahkan antara data latih dan data uji agar model yang dibangun untuk tujuan pelatihan dan evaluasi model. Tujuan dari split dataset adalah untuk melatih model pada data yang terpisah dari data yang digunakan untuk menguji kinerja model. Hal ini penting untuk menghindari overfitting, yaitu kondisi di mana model terlalu sempurna dalam mempelajari data pelatihan tertentu sehingga tidak dapat generalisasi dengan baik pada data baru.

K-Fold Cross Validation adalah suatu metode validasi yang umum digunakan dalam machine learning untuk mengevaluasi performa model secara objektif. Dalam K-Fold Cross Validation, dataset yang tersedia dibagi menjadi K subset atau "lipatan" yang seimbang secara acak. Kemudian, proses evaluasi dilakukan dengan membagi dataset menjadi K bagian, di mana setiap bagian akan bertindak sebagai data validasi secara bergantian, sementara K-1 bagian lainnya digunakan sebagai data pelatihan [9]. Gambar 2 dibawah ini menunjukan cara splitting data menggunakan k-fold cross validation.



Gambar 2. Kfold Cross Validation

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Gambar 2 menunjukan bahwa data dibagi menjadi 5 fold yang artinya masing-masing fold mempunya skema pembagian data latih dan data uji yang berbeda. Dataset yang sudah menjadi data uji di setiap split tidak akan menjadi data uji lagi pada split berikutnya. Pembagian dataset lebih adil karena data latih sudah mewakili semua skema.

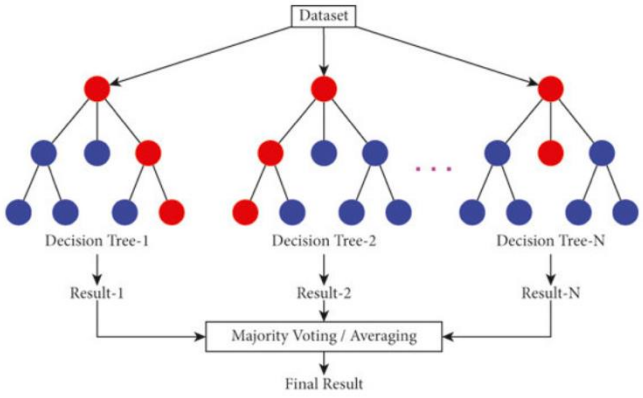
2.6 *Model Random Forest*

Model Random Forest merupakan salah satu teknik *Machine Learning* untuk mengatasi masalah klasifikasi maupun regresi. Cara kerja teknik ini adalah membuat beberapa pohon keputusan yang dibangun pada subset acak dari data pelatihan, dan prediksi akhir didasarkan pada agregat hasil dari pohon-pohon tersebut. Berikut adalah pseudocode algoritma *Random Forest* yang dapat dilihat pada tabel 2.

Tabel 2. Pseudocode Random Forest

Pseudocode Random Forest
Input : Data numerik **Output**: Hasil Prediksi **Step:** 1. Algoritma memilih sampel acak dari dataset yang disediakan 2. Membuat decision tree untuk setiap sampel yang dipilih. Kemudian akan didapatkan hasil prediksi dari setiap *decision tree* yang dibuat. 3. Melakukan proses voting terhadap semua keputusan decision tree yang sudah dibuat. 4. Algoritma akan memilih hasil prediksi yang paling banyak dipilih sebagai keputusan prediksi akhir

Pseudocode pada tabel 2 dapat diilustrasikan dengan gambar 3. Input berupa dataset dan output berupa voting dari beberapa pohon keputusan.



Gambar 3. Visualisasi Cara Kerja Random Forest

2.7 Evaluasi Model

F1 score adalah salah satu metrik evaluasi model yang umum digunakan untuk mengevaluasi performa model klasifikasi. F1 score menggabungkan presisi (precision) dan recall untuk memberikan gambaran keseluruhan tentang seberapa baik model dapat mengklasifikasikan dengan benar kelas positif, misalnya, kelas minoritas atau kelas yang dianggap penting [10]. Berikut adalah rumus F1 Score dapat dilihat pada rumus dibawah ini.

$$F1\ score\ =\ 2\ *\ \frac{Recall*Precission}{Recall+Precission} \quad (1)$$

$$Recall\ =\ \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (2)$$

$$Recall\ =\ \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (3)$$

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3. HASIL DAN PEMBAHASAN

Detail hasil dan penelitian ini diuraikan dalam beberapa poin yang merupakan penjabaran di metode penelitian pada gambar 1.

3.1 Dataset

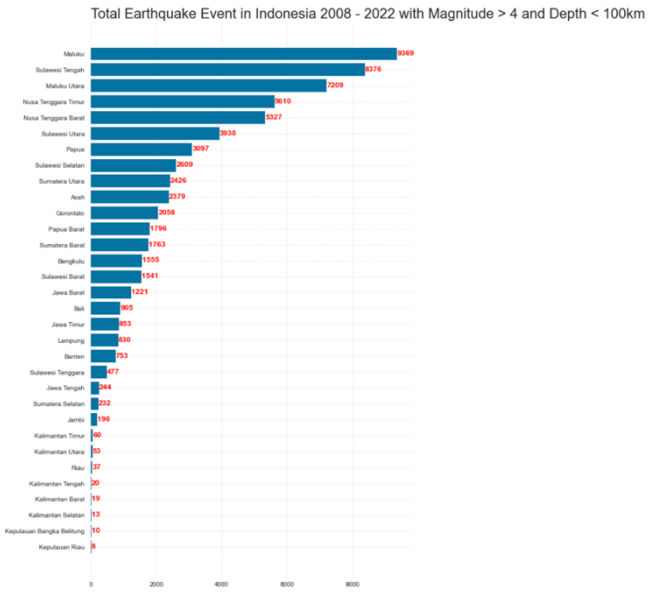
Dataset yang digunakan merupakan hasil dari *scrapping* dari website resmi BMKG Indonesia, total dataset berjumlah 77.257 data dengan parameter date, time, latitude, longitude, kedalaman, *magnitude*, tsunami dan region. Data diambil dari tahun 2001 sampai 2022. Gambar dataset dapat dilihat pada gambar

2022/03/30	23:44:03.579	0.19 S	119.79 E	10	2.8	-	Minahassa Peninsula, Sulawesi
2022/03/30	23:34:51.306	3.38 S	128.82 E	10	3.3	-	Seram, Indonesia
2022/03/30	23:31:43.655	0.00 S	124.46 E	24	3.4	-	Southern Molucca Sea
2022/03/30	18:00:57.335	1.23 N	128.46 E	10	4.1	-	Halmahera, Indonesia
2022/03/30	17:37:58.817	7.13 S	115.29 E	18	3.4	-	Bali Sea
2022/03/30	16:54:04.532	4.96 S	102.82 E	21	3.9	-	Southern Sumatra, Indonesia
2022/03/30	15:30:28.970	2.55 N	98.63 E	10	2.2	-	Northern Sumatra, Indonesia
2022/03/30	14:30:18.268	0.05 S	124.51 E	11	4.8	-	Southern Molucca Sea
2022/03/30	11:35:32.712	0.02 N	124.44 E	28	3.9	-	Minahassa Peninsula, Sulawesi
2022/03/30	11:21:34.869	1.03 S	128.80 E	10	3.6	-	Halmahera, Indonesia
2022/03/30	09:18:16.971	9.31 S	114.28 E	10	3.2	-	South of Bali, Indonesia
2022/03/30	08:49:57.609	8.83 S	110.21 E	10	3.7	-	Java, Indonesia

Gambar 4. Gambar Dataset

3.2. Visualisasi Data

Tujuan peneliti untuk mendapatkan output berupa prediksi lokasi gempa bumi yang berdampak bagi masyarakat disekitar lokasi kejadian. Karakteristik Gempa Bumi berdasarkan magnitude dapat dilihat pada gambar 5 dengan sumber dari BMKG. Gambar menunjukan total kejadian gempa bumi di Indonesia kurun waktu 2008 – 2022 dengan magnitude lebih dari 4 dan kedalaman kurang dari 100km.

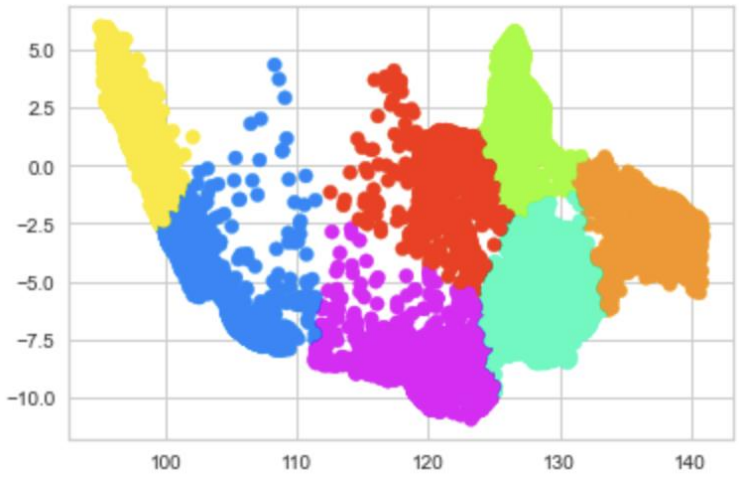


Dari data pada gambar 5 hasil yang dapat diambil adalah Provinsi Maluku mempunyai intensitas yang paling sering terjadi gempa bumi yaitu sebanyak 9.369 kejadian. Daerah yang paling jarang mengalami gempa bumi yaitu di Kepulauan Riau.

Gambar 5. Sebaran Data Gempa Bumi Indonesia

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Data gempa bumi divisualisasikan ke dalam tujuh cluster yang optimal melalui data *latitude* dan *longitude*. Persebaran data gempa bumi dapat dilihat di gambar 6.



Gambar 6. Persebaran data gempa bumi di Indonesia

3.3. K-Means Clustering

Tujuan k-means clustering pada penelitian ini untuk mengelompokan data gempa menjadi empat yaitu sangat berbahaya, berbahaya, medium, tidak berbahaya. Hasil *clustering* dapat dilihat pada tabel 3.

Tabel 3. Hasil clustering gempa bumi Indonesia

Date	Time	Latitude	Longitude	Magnitude	Kedalaman	Tsunami	Lokasi	Cluster
2022/03/30	18:00:57	1.77	-7.65	6.7	400 m	-	Maluku	Berbahaya
2022/03/29	06:15:41	-2.56	1.89	3.2	186m	-	Sulsel	Tidak Berbahaya
2022/03/28	12:23:31	-1.78	-15.22	4.6	386 m	-	Sumbar	Medium
2022/03/27	14:45:47	0.77	-7.65	6.4	89 m	-	Maluku	Berbahaya

3.4. Random Forest Model Klasifikasi

Pembangunan model *Random Forest* dilakukan setelah proses *clustering*. Pembangunan model dilakukan dengan menyertakan beberapa parameter banyak pohon untuk menentukan model yang paling baik diantara parameter lain. Hasil F1-Score dengan beberapa parameter dan perbandingan dengan model *Machine Learning* lainnya dapat dilihat pada tabel.

Tabel 4. Perbandingan nilai F1-Score klasifikasi Random Forest

Model Machine Learning	Banyak Pohon	F1-Score (%)
Random Forest	50	82.53
Random Forest	60	83.32
Random Forest	70	82.24
Random Forest	80	84.48
Random Forest	90	85.39
Random Forest	**100**	**87.21**
Naïve Bayes	-	85.89
Decision Tree	-	86.24
SVM	-	87.11

Model Random Forest dengan memakai 100 parameter pohon dapat menghasilkan F1-Score sebesar 87,21% , hasil tersebut membuat unggul jika dibandingkan dengan model Machine Learning lainnya seperti Naïve Bayes, Decision Tree, dan SVM. Banyaknya pohon yang digunakan untuk memberikan keputusan sangat mempengaruhi baik

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 83-89
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

atau buruknya model. Pengambilan parameter yang optimal dituntut untuk tidak terlalu banyak keputusannya dan tidak terlalu sedikit dalam memberikan keputusan.

3.5. Feature Selection

Feature Selection dilakukan untuk menyeleksi kolom yang kurang memiliki hubungan antara kolom lain, kolom tersebut dapat memperburuk kinerja model yang sudah dibangun [11]. Hasil dari seleksi fitur didapatkan F1-Score seperti yang terdapat pada tabel.

Tabel 5. *Feature Selection* untuk klasifikasi random forest dengan 100 pohon

Fitur	F1-Score (%)
[Tanggal, Latitude, Longitude, Magnitude, Lokasi]	**93.23**
[Tanggal, Latitude, Longitude, Magnitude, Kedalaman]	91.44
[Latitude, Longitude, Magnitude, Lokasi]	90.12
[Latitude, Longitude, Magnitude, Tsunami, Lokasi]	89.78
[Tanggal, Waktu, Latitude, Longitude, Magnitude, Tsunami, Lokasi]	87.21
[Latitude, Longitude, Magnitude, Tsunami, Lokasi]	86.95

Pengambilan kolom tanggal, latitude, longitude, magnitude dan lokasi saja dapat meningkatkan performansi model jika tanpa adanya *feature selection*. Dari percobaan pada tabel diatas membuktikan bahwa menggunakan fitur yang tepat dan efisien dapat meningkatkan performansi model.

4. KESIMPULAN

Model Random Forest menghasilkan F1-Score terbaik sebesar 92.23% jika menggunakan tambahan proses *feature selection*. Kedekatan antar fitur-fitur diteliti dalam penelitian ini sehingga proses tersebut terbukti meningkatkan performansi model karena mengambil parameter yang tepat untuk proses pembangunan Machine Learning. Random Forest memiliki kemampuan untuk menghasilkan model yang akurat dalam berbagai tugas. Ini karena Random Forest menggabungkan prediksi dari banyak pohon keputusan yang dibangun secara acak, dan kemudian mengambil hasil mayoritas. Kombinasi prediksi ini membantu mengurangi overfitting dan menghasilkan prediksi yang lebih baik.

DAFTAR PUSTAKA

[1] Tupan Tupan Tupan, Noorika Retno Widuri, and Rulina Rachmawati Rachmawati, "Analisis Bibliometrik Publikasi Ilmiah Tentang Prediksi Gempa Bumi Berbasis Data Scopus Periode Tahun," *Libraria Jurnal Perpustakaan*, vol. 8, no. 1, 2020.

[2] A. Noor, "Perbandingan Algoritma Support Vector Machine Biasa dan Support Vector Machine Berbasis Particle Swarm Optimization untuk Prediksi Gempa," 2018.

[3] F. Yulian Pamuji, V. Puspaning Ramadhan, and R. Artikel, "Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy Info Artikel ABSTRAK," vol. 7, pp. 46–50, 2021, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[4] V. R. Sari, F. Firdausi, and Y. Azhar, "EDUMATIC: Jurnal Pendidikan Informatika Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.

[5] W. Apriliah *et al.*, "SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[6] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052031.

[7] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, "A review of machine learning methods of feature selection and classification for autism spectrum disorder," *Brain Sciences*, vol. 10, no. 12. MDPI AG, pp. 1–23, Dec. 01, 2020. doi: 10.3390/brainsci10120949.

[8] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[9] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Comput Stat*, vol. 36, no. 3, pp. 2009–2031, Sep. 2021, doi: 10.1007/s00180-020-00999-9.

[10] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach Learn*, vol. 110, no. 3, pp. 451–456, Mar. 2021, doi: 10.1007/s10994-021-05964-1.

[11] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit Health*, vol. 6, 2020, doi: 10.1177/2055207620914777.