OPEN ACCESS
CENTIVE 2019

ISBN : 978-602-53004-0-0

Conference on Electrical Engineering, Telematics, Industrial Technology, and Creative Media 2019

Prediksi Tsunami Pada Gempa Menggunakan *Random Forest Classifier*

Jerry Lasama[1], Andre Pradika E.P[2], Agi Prasetiadi[3]

Fakultas Teknologi Industri dan Informatika, Institut Teknologi Telkom Purwokerto
Kawasan Pendidikan Telkom, Jl. DI Panjaitan No 128 Purwokerto 53147 Indonesia

[1] 18102018@ittelkom-pwt.ac.id
[2] 18102148@ittelkom-pwt.ac.id
[3] agi@ittelkom-pwt.ac.id

Abstrak

Gempa yang diikuti oleh tsunami memiliki ciri khusus seperti kedalaman, besar, dan lokasi tertentu yang harus dianalisis terlebih dahulu sebelum dinyatakan akan diikuti tsunami atau tidak. Kemajuan teknologi *machine learning* memungkinkan kita melakukan prediksi terjadinya tsunami lebih efisien dibanding sebelumnya. Riset ini memanfaatkan *machine learning* khususnya algoritme *Random Forest Classifier* untuk membuat model yang dapat memprediksi potensi tsunami dengan menggunakan data historis *global significant earthquake* milik NOAA dari tahun 2100 SM yang berisi pola Negara, Kode Region, Lintang, Bujur, Tahun, Bulan, Tanggal, Kedalaman, serta Besarnya gempa. Hasil simulasi model untuk memprediksi tsunami dari gempa pada testing set menunjukan akurasi di atas 75%.

Kata kunci: *gempa, machine learning, random forest, tsunami*

I. PENDAHULUAN

TSUNAMI merupakan gelombang yang dibangkitkan dari pergerakan vertikal kolom air. Pergerakan ini dapat disebabkan oleh aktivitas seismik seperti letusan gunung api, longsoran pada dasar atau atas laut, tubrukan benda langit, atau fenomena meteorologi [1]. Terjadinya tsunami setelah gempa memiliki probabilitas dengan parameter-parameter tertentu [2], contohnya gempa sesar besar ataupun gempa *splay fault* [3]. Berbagai permodelan dilakukan untuk mencari korelasi antara gempa dengan tsunami seperti permodelan simulasi stokastik tsunami [4], bahkan hingga mencari parameter gempa yang berkaitan dengan tsunami [3].

Metode mitigasi bencana tsunami dapat melalui tiga hal yaitu dengan memindahkan seluruh populasi dan infrastruktur yang ada, jauh dari tempat yang sering terjadi tsunami, lalu dengan meningkatkan infrastruktur untuk memitigasi bencana dan mengedukasi masyarakat tentang bencana tersebut, serta membuat suatu sistem peringatan dini dan informasi cepat tanggap [5]. Metode pertama kurang praktikal untuk tempat dengan lokasi yang sudah didiami banyak orang [5], contoh dari metode kedua adalah mengedukasi masyarakat terhadap bencana alam [2]. Penelitian ini memanfaatkan *machine learning* untuk menganalisis kejadian tsunami berdasarkan kode negara, region, lintang, bujur, tahun, bulan, tanggal, kedalaman, serta besarnya gempa.

II. METODE PENELITIAN

A. Data Preprocessing

Preprosesing pada data dibutuhkan untuk mengurangi kompleksitas data sehingga dapat dikomputasi secara optimal oleh algoritma yang digunakan untuk pengamatan[6]. Data dunia nyata dipenuhi dengan hilangnya data pada kolom atau baris pada suatu *dataset*, hilangnya data-data tersebut tentu dapat mengganggu validitas hasil analisa [7].

B. Random Forest Classifier

Random Forest merupakan algoritma yang membuat suatu *bootstrap* pohon klasifikasi dan regresi yang *node*nya dipisahkan oleh algoritma optimasi *Greedy* dengan *Gini impurty* sebagai fungsi untuk meminimalkan *squared-error loss* yang dimiliki[8]. Waktu komputasi yang dibutuhkan algoritma ini untuk mengklasifikasi adalah:

$$T\sqrt{MNlog(N)} \tag{1}$$

dimana T adalah jumlah banyaknya pohon, M adalah banyaknya peubah yang digunakan dalam pemisahan setiap subsampel, dan N adalah banyaknya sampel pada *training set* [9].

C. K-Fold Cross Validation

K-Fold Cross Validation membagi data set sebanyak *k* partisi yang tidak tumpang tindih yang nantinya model akan dilatih sebanyak *k*-1 partisi dan partisi ke *k* digunakan untuk validasi, dan sisanya akan diiterasi untuk dilatih dan validasi lagi pada partisi berikutnya [10].

D. Confusion Matrix

Confusion Matrix merangkum performa model dalam mengklasifikai berdimensi sesuai dengan banyaknya kategori yang diklasifikasi terindeks dari nilai benar dari suatu kelas dari objek yang diprediksi[11]. Berikut merupakan contoh dari *confusion matrix* 2 dimensi:

TABEL I
EXAMPLE OF CONFUSION MATRIX

Confusion Matrix	*Predicted Positive*	*Predicted Negative*
Actual Positive	TP	FN
Actual Negative	FP	TN

dengan tingkatan akurasi yang didefinisikan dengan:

$$Accuracy = \sum_{i=1}^{n} N_{ii} / \sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij} \tag{2}$$

akurasi tersebut menunjukkan proporsi dari jumlah prediksi yang benar [12].

E. Precision and Recall F1-Score

Pengukuran akurasi dalam klasifikasi tidak dapat menghitung apabila terjadi suatu ketidakseimbangan antar kelas [13], solusi yang baik untuk mengatasinya adalah dengan mengabaikan nilai yang terkategori *true negative* seperti pada tabel 1, dan menggunakan *precision – recall* [14].

Precision mengukur akurasi dari kelas spesifik yang telah terprediksi, secara formal didefiniskan dengan:

$$Precision_i = N_{ii} / \sum_{k=1}^{n} N_{ki} \tag{3}$$

Recall mengukur kemampuan model dalam memprediksi instansi dari suatu kelas, secara formal didefinisikan dengan

$$Recall_i = N_{ii} / \sum_{k=1}^{n} N_{ik} \quad (4)$$

F1-Score mengukur rerata harmonik dari *Precision* dan *Recall,* secara formal didefinisikan dengan:

$$F - Score_i = \frac{2\ x\ Precision_i\ x\ Recall_i}{Precision_i\ x\ Recall_i} \quad (5)$$

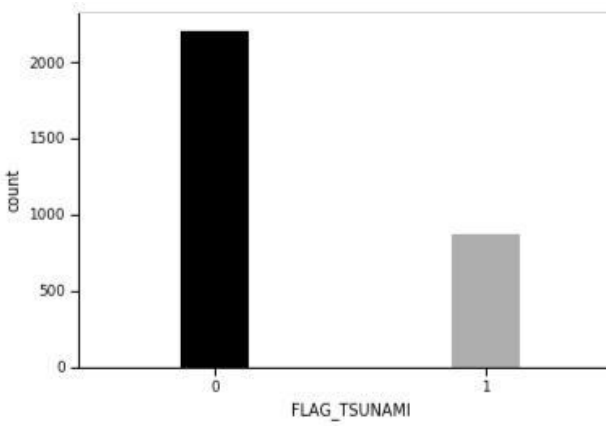
III. HASIL PENELITIAN

A. *Dataset and Preprocessing*

Dataset yang digunakan dalam penelitian ini adalah katalog *Global Significant Earthquake* milik NOAA [11], dengan 6139 baris dan 47 kolom. Data tersebut kemudian dipreproses dengan menseleksi kolom hingga menjadi Negara, Kode Region, Lintang, Bujur, Tahun, Bulan, Tanggal, Kedalaman, dan Besarnya Gempa, sebagai peubah yang menjadi *input* dan *FLAG_TSUNAMI* sebagai peubah yang akan diprediksi. Setelahnya akan dilakukan *pairwise deletion* pada data-data yang hilang rekordnya. Didapatkan 3073 baris dengan 10 kolom dari hasil *pairwise deletion*, lalu pada peubah Negara dan *FLAG_TSUNAMI* dilakukan faktorisasi dimana label berupa text akan dikonversi menjadi bilangan bulat. *Dataset* akan dipisah menjadi 80% untuk *training* dan 20% untuk *testing*. Serta, penseleksian peubah *Country, Region Code, Latitude, Longitude, Month, Date, Focal Depth, Primary Magnitude* menjadi *input,* dan *Flag Tsunami* sebagai target. Kemudian, data *training* akan displit oleh *KFold* menjadi 10 partisi dan mulai dilakukan proses *training* pada model.

B. *Modelling Process*

Sebelum model dibuat, dilakukan pemeriksaan terlebih dahulu pada peubah target.



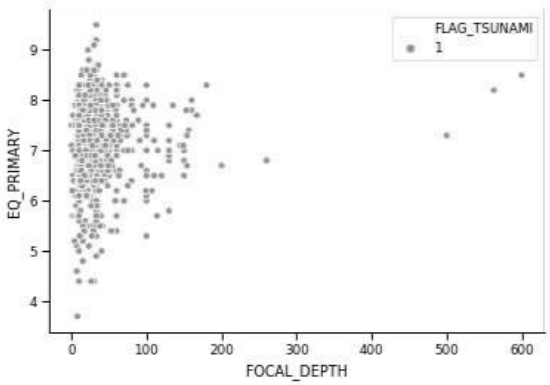
Gambar 1. *Countplot* pada peubah FLAG_TSUNAMI

Pada gambar 1 terlihat adanya ketidakseimbangan antar kelas 0 (Tidak diikuti tsunami) dan kelas 1. Seperti yang ditunjukan pada [15] bahwa *Random Forest Classifier robust* pada ketidakseimbangan data sehingga tidak diperlukan preprosessing lebih lanjut pada peubah target.

Selanjutnya, dilakukan pemeriksaan hubungan antara peubah *input* dengan peubah target, dimulai dari distribusi gempa yang diikuti tsunami.

Gambar 2. Distribusi gempa yang diikuti tsunami di dunia

Dari gambar 2 terlihat jelas bahwa gempa yang diikuti tsunami hanya berada pada pinggir lempeng pulau seperti yang dijelaskan pada [3].



Gambar 3. Kedalaman dan besar gempa terhadap flag tsunami

Gambar 3 menunjukan pemusatan gempa yang diikuti tsunami pada kedalaman 0 hingga 100 dengan kekuatan gempa 6 hingga 8.

C. Model Performance

Berdasarkan model yang telah ditrain dengan 10 partisi *k-fold* akurasi yang didapatkan berada di atas 78%, dengan *confusion matrix* seperti berikut:

TABEL II
CONFUSION MATRIX

Confusion Matrix	Terprediksi Tsunamigenik	Terprediksi Bukan Tsunamigenik
Benar Tsunamigenik	37	35
Benar Bukan Tsunamigenik	18	154

Bersumber pada Tabel 1 didapat akurasi prediksi sebesar 78%, selanjutnya mari kita lihat performa *precision, recall,* dan *f1-score* dari model ini:

TABEL III
PRECISION RECALL F1-SCORE

Kelas	***Precision***	***Recall***	***F1 Score***	***Samples***
Tsunamigenik	0.7755	0.5671	0.6551	134
Bukan Tsunamigenik	0.8878	0.9542	0.9198	481

Dari tabel di atas dapat disimpulkan bahwa rata-rata *precision* adalah 0.8316, *recall* 0.7607, dan *f1-score* 0.7875.

Serta didapatkan *Feature Importances* dari model sebagai berikut:

TABEL IV
FEATURE IMPORTANCE

Feature	***Importance***
LONGITUDE	0.200052
EQ_PRIMARY	0.198399
YEAR	0.125479
LATITUDE	0.120353
FOCAL_DEPTH	0.115496
DAY	0.076829
COUNTRY	0.070139
MONTH	0.052137
REGION_CODE	0.041115

Pada tabel 4 terlihat *LONGITUDE* dan *EQ_PRIMARY* memiliki faktor *importance* tertinggi di sekitar 20%, berikut performa model jika hanya menggunakan kedua peubah tersebut:

TABEL V
CONFUSION MATRIX LONGITUDE + EO PRIMARY

Confusion Matrix	Terprediksi Tsunamigenik	Terprediksi Bukan Tsunamigenik
Benar Tsunamigenik	35	37
Benar Bukan Tsunamigenik	29	142

pada tabel 5 didapat akurasi prediksi sebesar 72%, lalu performa *precision, recall,* dan *f1-score* dari model ini adalah:

TABEL VI
PRECISION RECALL F1-SCORE LONGITUDE + EO PRIMARY

Kelas	***Precision***	***Recall***	***F1 Score***	***Samples***
Tsunamigenik	0.6055	0.4922	0.5432	134
Bukan Tsunamigenik	0.8656	0.9106	0.8875	481

Dari tabel di atas dapat disimpulkan bahwa rata-rata *precision* adalah 0.7452, *recall* 0.7245, dan *f1-score* 0.7337. Meskipun terjadi sedikit penurunan akurasi, didapatkan performa *training* yang lebih cepat karena *feature* yang digunakan semakin sedikit, berikut merupakan perbedaan kecepatan eksekusi:

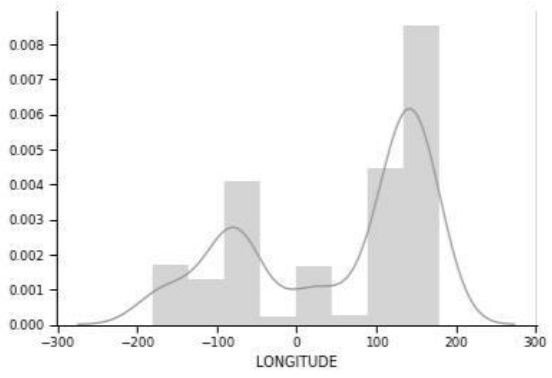
TABEL VII
PERFORMANCE DIFFERENCE

Numbers of Features	***Running Time***
9	683 milisekon
2	475 milisekon

Terlihat adanya selisih pada *running time* sebanyak 208 milisekon.

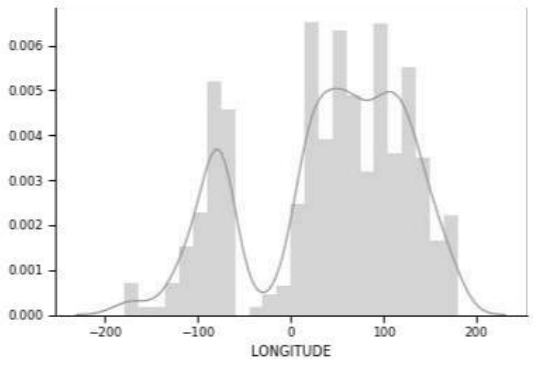
IV. PEMBAHASAN

Tabel 4 menunjukan peubah *LONGITUDE* dan *EQ_PRIMARY* merupakan prediktor yang sangat kuat, hal tersebut diperjelas pada tabel 5,6 dan 7 dimana pengurangan akurasi pada model utama yang menggunakan 9 peubah dengan model yang menggunakan 2 peubah sebanyak 6% serta running time yang berkurang 200 milisekon untuk 3000 data.



Gambar 4. Distribusi *longitude* terhadap *flag tsunami* 1

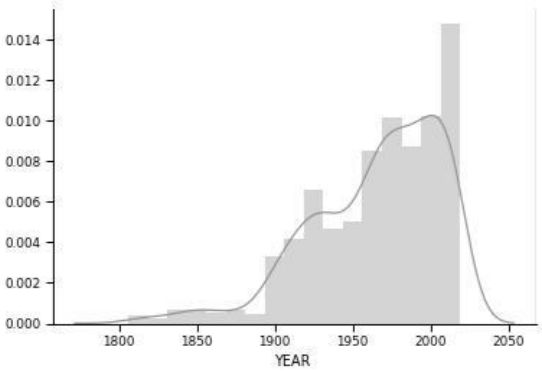
Jelas bahwa peubah *EQ_PRIMARY* dan *FOCAL_DEPTH* berkorelasi dengan gempa tsunamigenik seperti yang dijelaskan pada[3], untuk longitude, polanya dapat dilihat melalui gambar berikut:



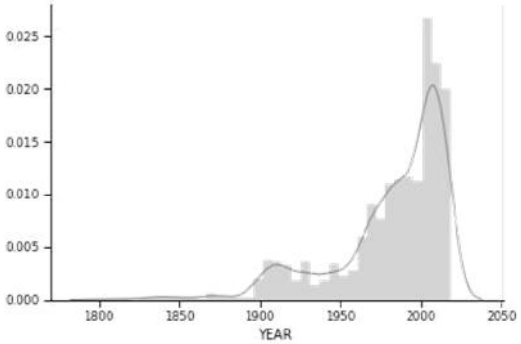
Gambar 5. Distribusi *longitude* terhadap *flag tsunami* 0

Pada gambar 2 terlihat bahwa gempa tsunamigenik hanya terjadi di pinggir lempeng pulau, dan memiliki pola lintang dan bujur tertentu, pada gambar 4 dan 5 terlihat jelas bahwa gempa tsunamigenik berpusat pada bujur -200 hingga -100 dan 100 hingga 200, sehingga model dapat dengan mudah mencari pola didalam peubah *LONGITUDE* dan *LATITUDE*, yang kemudian akan dijadikan *feature* dengan tingkat *importances* tinggi. Perlu dilakukan penelitian lebih lanjut terhadap korelasi antara peubah *LONGITUDE* dan *LATITUDE* dengan gempa tsunamigenik untuk menemukan alasan lain.

Selanjutnya adalah memeriksa pengaruh peubah *YEAR* yang memiliki *feature importance* sebesar 10%, dengan melihat distribusinya :



Gambar 6. Distribusi *year* terhadap *flag tsunami* 1



Gambar 7. Distribusi *year* terhadap *flag tsunami* 0

Terlihat adanya perbedaan pada distribusinya dimana pada gambar 6 distribusi memusat pada tahun 1900 hingga 2000, sedangkan pada gambar 7 distribusi terpusat pada tahun 2000. Kedua hal tersebut menunjukan adanya siklus tahunan pada gempa tsunamigenik yang membutuhkan penelitian lebih lanjut.

V. PENUTUP

A. Kesimpulan

Model yang dibuat dapat memprediksi terjadinya tsunami dari gempa yang terjadi dengan tingkat akurasi diatas 75% dan dapat ditingkatkan seiring bertambahnya ketersediaan data yang ada.

B. Saran

Penelitian berikutnya dapat menggunakan metode-metode lain untuk memproses ketidakseimbangan pada kelas *Flag_Tsunami*, mengoptimasi parameter pada model, serta menguji algoritma lain yang dapat memprediksi lebih akurat. Dan juga dapat meneliti lebih lanjut hubungan antara lintang dan bujur terhadap gempa tsunamigenik, serta siklus gempa tahunan gempa tsunamigenik.

DAFTAR PUSTAKA

[1] B. Mambu, G. H. Tamuntuan, and G. Pasau. "Simulasi Ketinggian dan Waktu Tiba Gelombang Tsunami di Tahuna Sebagai Upaya Mitigasi Bencana," *J. MIPA*, vol. 8, no. 1, p. 13, 2019.
[2] D. Santoro, M. Yamin, and M. Mahrus, "Penyuluhan Tentang Mitigasi Bencana Tsunami Berbasis Hutan Mangrove Di Desa Ketapang Raya Kecamatan Keruak Lombok Timur," no. 1982, 2019.
[3] I. Van Zelst, S. Brizzi, and E. Van Rijsingen, "Investigating global correlations between tsunami, earthquake, and subduction zone characteristics."
[4] K. Goda, T. Yasuda, N. Mori, and T. Maruyama, *New scaling relationships of earthquake source parameters for stochastic tsunami simulation*, vol. 58, no. 3. 2016.
[5] D. Melgar and Y. Bock, "Journal of Geophysical Research: Solid Earth Kinematic earthquake source inversion and tsunami runup," *J. Geophys. Res. Solid Earth*, pp. 1-26, 2015.
[6] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017.
[7] K. M. Lang and T. D. Little, "Principled missing data treatments," *Prev. Sci.*, vol. 19, no. 3, pp. 284–294, 2018.
[8] S. Wager, "Comments on: A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 261–263, 2016.
[9] M. Belgiu and L. Dra, "ISPRS Journal of Photogrammetry and Remote Sensing Random forest in remote sensing: A review of applications and future directions ˇ gut," vol. 114, pp. 24–31, 2016.
[10] K. J. Grimm, G. L. Mazza, P. Davoudzadeh, K. J. Grimm, G. L. Mazza, and P. Davoudzadeh, "Model Selection in Finite Mixture Models: A k-Fold Cross-Validation Approach Model Selection in Finite Mixture Models: A k - Fold Cross-Validation Approach," *Struct. Equ. Model. A Multidiscip. J.*, vol. 00, no. 00, pp. 1–11, 2016.
[11] E. Martin, *S Sample Complexity Description of the Learning System Cross-References Recommended Reading Search Engines: Applications of ML Motivation and Background*. 2017.
[12] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci. (Ny).*, vol. 340–341, pp. 250–261, 2016.
[13] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp.249-259, 2018.
[14] P. A. Flach and M. Kull, "Precision-Recall-Gain curves: PR analysis done right," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 838–846, 2015.
[15] D. J. DIttman, T. M. Khoshgoftaar, and A. Napolitano, "The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data," *Proc. - 2015 IEEE 16th Int. Conf. Inf. Reuse Integr. IRI 2015*, pp. 457–463, 2015.