OPEN ACCESS
CENTIVE 2019
ISBN : 978-602-53004-0-0

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

Memprediksi Ketinggian Tsunami Menggunakan *Random Forest Regressor*

Novantri Prasetya Putra[1], Jerry Lasama[2], Andre Pradika E.P[3],Agi Prasetiadi[4]

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

[1] 18102279@ittelkom-pwt.ac.id
[2] 18102018@ittelkom-pwt.ac.id
[3]18102148@ittelkom-pwt.ac.id
[4]agi@ittelkom-pwt.ac.id

Abstrak

Tsunami memiliki ketinggian yang berbeda-beda, semakin tinggi maka semakin banyak persiapanmenghadapi tsunami untuk menekan jumlah korban. *Machine learning* dapat digunakan untuk menentukan tinggi dari tsunami. Ketika tsunami datang, ketinggian tsunami dapat diprediksi untuk mengetahui titik aman pada bibir pantai. Hal ini akan sangat membantu dalam proses evakuasi sehingga dapat menekan jumlah korban jiwa. Metode yang digunakan yaitu *Random Forest Regressor* untuk dapat menganalisa peubah latitude, longtitude, tahun, bulan, hari, kekuatan gempa, negara, kode negara,dan kode penyebab gempa dengan ketinggian maximum tsunami. Dengan menggunakan metode tersebut, didapatkan negatif mean absolut error sebesar -3,93. Setelah Mean Absolute Error didapatkan, akan ditemukan feature yang memiliki importance terbesar untuk dapat dijadikan predictor.

Kata kunci: *ensemble, machine learning, prediksi ketinggian tsunami*

I. PENDAHULUAN

PADA22 Desember 2018 telah terjadi letusan gunung anak Krakatau yang menyebabkan terjadinya tsunami setinggi 4-6 meter. Peristiwa ini telah menyebabkan banyak nyawa dan kerusakan di seluruh wilayah pesisir[1]. Banyaknya korban jiwa dikarenakan tidak adanya pengumuman akan terjadinya tsunami serta ketidaktahuan akan ketinggian air yang akan datang. Kejadian ini menggerakan kami untuk melakukan penelitian yang dapat memprediksi ketinggian tsunami sehingga warga dapat mengetahui *safe point* bibir pantai. Karakteristik utama dari gempabumi yang menyebabkan tsunami adalah besarnya gempa bumi dalam skala Richter harus besar serta harus memiliki kedalaman fokus dangkal[2].Persiapan awal dilakukan seperti sistem peringatan dini yang dioperasikan di seluruh dunia sebagai langkah mengurangi kerusakan yang diakibatkan oleh tsunami[3].

II. METODE PENELITIAN

A. Data Preprocessing

1) Menghilangkan Missing Value:Missing value adalah nilai yang belumdirekam dalam kumpulan data karena alasan apa pun, atau yang tidak pernah terdata. Diperlukanpembersihan dan persiapan data untuk

membuatnya jelas dan berguna untuk proses ekstraksi pengetahuan. Memproses ulang data adalah pendekatan yang paling umum digunakan untuk melakukan tugas perbaikan ini, dan banyak pilihan tersedia[7]. *Missing values* dapat dikategorikan ke dalamtiga kategori berkaitan dengan mekanisme statistik yang paling menggambarkan mereka yaitu Missing *Compelety At Random* (MCAR) merupakan kumpulan data proteomik, yang sesuai dengan kombinasi dan penyebaran beberapa kesalahan kecil atau fluktuasi stokastik, *MissingAt Random* (MAR) merupakan kelas yang lebihumum daripada MCAR di mana dependensi bersyarat diperhitungkan., dan *Missing Not AtRandom* (MNAR) Merupakan kebalikan dariMAR yaitu memiiliki efek yang di targetkan[8].

*2) Standarisasi Z Score:*Metode *Z score* umum digunakan dalam praktek karena menggunakan skala yang sama dalam membandingkan nilai suatu data. Model *Zscore* membandingkan nilai dari kriteria tertentusuatu data dengan rata-rata data dan dibagi dengan standar deviasinya[9]. Rumus *Z Score* yaitu [10]:

$$Z = \frac{(Xn - X)}{\sigma} \quad (1)$$

B. K Fold Cross Validation

Merupakan prosedur yang popular untuk memperkirakan kinerja pada klasifikasi algoritma atau membandingkan kinerja antara dua klasifikasi algoritma pada set data[11]. Metode ini memerlukan data historis yang acak untuk diatur kedalam „lipatan" berukuran sama dan kemudian menggunakannya sebagai sampel pengujian dalam latihan pengujian serta pelatihan[12]. Keakuratan yang diperoleh dalam setiap iterasi kemudian dirata-rata untuk mendapatkan model akurasi. Satu hal penting yang perlu diperhatikan adalah bahwa data biasanya disusun sebelum dibagi menjadi segmen pada K. Stratifikasi merupakan proses menata ulang data sedemikian rupa sehingga setiap lipatan merupakan perwakilan yang baik dari dari keseluruhan [13].

C. Random Forest Regressor

Random Forest Regressor adalah ensembel dari pohon regresi yang berlainan. Masing-masing memainkan peran pemetaan nonlinier dari ruang input kompleks ke ruang output kontinu. Nonlinier dicapai dengan membagi masalah utama menjadi masalah yang lebih kecil, diselesaikan dengan model yang lebih sederhana. Membagi node di pohon guna menjaga tes yang diterapkan pada sampel data untuk dikirimkan ke arah simpul anak kiri atau kanan. Tesdipilih dengan beberapa kriteria untuk mengelompokkan sampel pelatihan ke dalam kelompok dimana prediksi yang baik dapat dicapai dengan model sederhana. Model-model ini dihitung dari sampel data beranotasi yang mencapai daun dan disimpan di sana. Sementara overfitting kemungkinan terjadi hanya pada pohon keputusan standar, *ensembel* pohon yang dilatih secara acak menikmati kekuatan generalisasi yang tinggi[14].

D. Mean Absolute Error

Mean Absolute Error (MAE) didefinisikan sebagai perbedaan mutlak rata-rata pixelwise antara kebanaran dasar biner G dan peta arti-penting S[15].

Rumus untuk menentukan MAE bisa didapatkan dengan menggunakan rumus *Mean AbsolutePrecentage Error* (MAPE).

Rumus MAPE.

$$MAPE = \frac{|g(x) - y|}{|y|} \quad (2)$$

Rumus MAE[16].

$$MAE = \frac{|g(x) - y|}{|y|} X|y| = |g(x) - y| \quad (3)$$

E. Mean Squared Error

Mean Squared Error (MSE) berfungsi untukmenghitung tingkatan squared error pada prediksi dengan menggunakan rumus[17]:

$$\mathrm{MSE} = |\mathrm{h} - \mathrm{h(n)}|2 \quad (4)$$

III. HASIL PENELITIAN

A. Data Processing

Seringkali data yang didapat banyak berisikan data yang kosong pada bagian tertentu ataupun data yang tidak bisa terbaca oleh software yang digunakan, maka dari itu dilakukan *cleansing data* untuk membersihkan *missing value* yang terdapat pada data yang akandigunakan.

Missing value atau data yang kosong dibersihkandengan men *drop* atau menghapus data yang tidak digunakan. Maka data yang sudah dibersihkan akan lebih mudah diolah dan diproses.

Data yang sudah dibersihkan kemudian distandarisasi menggunakan *Z Score*. Berfungsi untuk membuat data lebih standar.

B. K Fold Cross Validation

Setelah didapatkan data yang lebih standart, proses dilanjutkan dengan *K Fold Cross Validation* untuk memperkirakan kinerja klasifikasi algoritma pada data sehingga mendapatkan nilai keakuratan untuk dijadikan model akurasi.

C. Random Forest Regressor

Kemudian gunakan *Random Forest Regressor* untuk mendapatkan nilai dari *Mean Absolute Error(MAE)* dan *Mean Squared Error (MSE)*.Menghasilkan nilai data MAE sebesar 3.7761 dan nilai data MSE sebesar 44.041.

D. Features Importance

MAE dan MSE sudah didapatkan lalu dilanjutkan dengan mencari *importance* dengan *FeatureImportance.* Seteleh mendapat *importance* dari semua *features*, dapat disimpulkan hasil dari *importance* yangsudah didapat bahwa PRIMARY_MAGNITUDE menjadi predictor atau faktor yang dapat memprediksi dengan keakuratan tertinggi dibanding dengan *features* lain.

E. Hasil

Berdasarkan model yang telah ditrain dengan 10 partisi dan didapatkan mean absolut error sebesar 5,33. Dengan hasil lain.

Membandingkan *feature latitude, longitude, year, month*, dan *year* dengan semua *feature*.

TABEL I
MEAN ABSOLUTE ERROR DAN MEAN SQUARED ERROR PADA SEMUA FEATURE

Metric	Mean	Standard Deviasi
MAE	3,7761	0,8541
MSE	44,041	32,182

TABEL II
FEATURE IMPORTANCE PADA SEMUA FEATURE

Features	Importance
PRIMARY_MAGNITUDE	0.28
YEAR	0.17
LATITUDE	0.11
CAUSE_CODE	0.11
MONTH	0.09
LONGITUDE	0.07
REGION_CODE	0.06
DAY	0.06
COUNTRY	0.05

TABEL III
MEAN ABSOLUTE ERROR DAN MEAN SQUARED ERROR PADA FEATURE LATITUDE, LONGITUDE, TAHUN, BULAN, DAN HARI

Metric	Mean	Standard Deviasi
MAE	3,7761	0,8541
MSE	44,041	32,182

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

TABEL IV
FEATURE IMPORTANCE PADA FEATURE LATITUDE, LOGITUDE, TAHUN, BULAT, DAN HARI

Features	Importance
DAY	0.27
YEAR	0.22
LONGITUDE	0.21
LATITUDE	0.18
MONTH	0.12

IV. PEMBAHASAN

Perbandingan importance antara semua feature dengan feature latitude, longitude, tahun, bulat, dan hari didapatkan bahwa importance PRIMARY_ MAGNITUDE memiliki importance terbesar menjadikan PRIMARY_MAGNITUDE sebagai predictor utama.

Namun, yang mengejutkan, didapatkan data bahwa tahun memiliki *importance* terbesar kedua. Seperti yang diketahui bahwa tahun dianggap sebagai feature yang tidak mempengaruhi ketinggian tsunami. Dapat disimpulkan bahwa mungkin saja bumi memiliki siklus tahunan yang belum bisa dipecahkan atau dimengerti oleh manusia. Mungkin untuk penelitian selanjutnya, kami akan membahas tentang siklus tahunan bumi.

Random Forest Regressor mendapatkan hasildengan membagi data menjadi dua jenis (*training datadan testing data*) dengan metode *K Fold Cross Validation* dimana nilai K diambil sebanyak 10. Hasilkedua data dibandingkan, kemudian *training data* dijadikan masukan sedangkan *testing data* dijadikan untuk menguji atau mengevaluasi output dari algoritma. Kemudian didapatkan data untuk mencari MAE dan MSE hingga *importance* yang dicari.

V. PENUTUP

A. Kesimpulan

Dengan menggunakan algoritma *Random ForestRegressor* dan *K Vold Cross Validation* didapatkanhasil *importance* untuk masing masing features. PRIMARY_MAGNITUDE memiliki *importance* sebesar 0.28 yang membuat *feature* ini menjadi faktor prediksi ketinggian tsunami.

B. Saran

Untuk penulisan selanjutnya dapat menggunakan data lebih dari 1000 data untuk mendapatkan data akurasi yang lebih besar dan menggunakan algoritmalain atau mengoptimasi parameter *Random ForestRegressor* yang ada.

DAFTAR PUSTAKA

[1] R. Nateghi, J. D. Bricker, S. D. Guikema, and A. Bessho, "Statistical analysis of the effectiveness of seawalls and coastal forests in mitigating tsunami impacts in iwate and miyagi prefectures," *PLoS One*, vol. 11, no. 8, pp. 1–21, 2016.
[2] A. Zaytsev, I. Kostenko, A. Kurkin, E.Pelinovsky, and A. C. Yalçiner, "The depth effect of earthquakes on tsunami heights in the sea of Okhotsk," *Turkish J. Earth Sci.*, vol. 25, no. 4, pp. 289–299, 2016.
[3] Y. Igarashi, T. Hori, S. Murata, K. Sato, T. Baba, and M. Okada, "Maximum tsunami height prediction using pressure gauge data by a Gaussian process at Owase in the KiiPeninsula, Japan," *Mar. Geophys. Res.*, vol. 37, no. 4, pp. 361–370, 2016.
[4] S. García, J. Luengo, and F. Herrera, "Preface," Intell. Syst. Ref. Libr., vol. 72, pp. 112–114, 2015.
[5] 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.
[6] S. García, S. Ramírez-Gallego, J. Luengo, J.M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *BigData Anal.*, vol. 1, no. 1, pp. 1–22, 2016.
[7] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," Knowledge-Based Syst., vol. 98, pp. 1–29, 2016.
[8] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies," *J. ProteomeRes.*, vol. 15, no. 4, pp. 1116–1125, 2016.
[9] D. Isynuwardhana, "Aplikasi Z-Score Method Dalam Pembentukan," vol. 17, no. 1, pp. 89–98, 2013.
[10] F. Prastio, S. Martha, and H. Perdana, "Di bandar udara internasional supadio," vol. 08, no. 2, pp. 185–192, 2019.
[11] Wong, Tzu-Tsung, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", 2015.
[12] M. R. Haley, "K-fold cross validation performance comparisons of six naive portfolio selection rules: how naive can you be and still have successful out-of-sample portfolio performance?," Ann. Financ., vol. 13, no. 3, pp. 341–353, 2017.
[13] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, "Proceedings-6th International Advanced Computing Conference, IACC 2016. pp. 78–83, 2016.
[14] Huy Phan, StudentMember,IEEE, Marco Maaß,Student Member, IEEE, Radoslaw Mazur,Member,IEEE, and Alfred Mertins, Senior Member, IEEE, "Random Regression Forests for Acoustic Event Detection and Classification" vol. 23, 2015.
[15] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 5455–5463, 2015.
[16] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," Neurocomputing, vol. 192, pp. 38–48, 2016.
[17] Y. Li, Y. Wang, and T. Jiang, "Norm-adaption penalized least mean square/fourth algorithm for sparse channel estimation,"Signal Processing, vol. 128, pp. 243–251, 2016