Jurnal SEICT
Vol. 1, No. 1 (2020), pp 10-16

Implementasi Metode XGBoost dan *Feature Importance* untuk Klasifikasi pada Kebakaran Hutan dan Lahan

Ichwanul Muslim Karo Karo
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
ichwanulkarokaro@telkomuniversity.ac.id

Abstract
Forest and land fires in Indonesia have become a problem of the annual environmental crisis. The largest distribution of forest fires occurred in the island of Sumatra. One of the measures to prevent and minimize the risk of forest fires is to classify the types of hotspots in the land, so that a priority scale can be obtained for fire suppression. This study aims to classify the types of hotspots using the XGBoost method and feature importance in Sumatra. Hotspot data obtained from Globalforestwatch.com. The process of subtracting variables from the data obtained has a very significant impact on the classification model. There are six and or seven variables that are very influential in determining hotspots, these variables also produce the best classification model. XGBoost and feature importance has an accuracy of 89.52%. Sensitivity (SE), Specificity (SP), and Matthews Correlation Coefficient (MCC). Were 91.32%, 93.16% and 92.75%, respectively. This method is also better than the results of previous studies.

Keywords: Hotspot, XGBoost, *feature importance*

Abstrak
Kebakaran hutan dan lahan di Indonesia telah menjadi masalah krisis lingkungan tahunan. Sebaran kebakaran hutan terbesar terjadi dipulau Sumatera. Salah satu upaya tindakan dalam pencegahan dan meminimalisasikan resiko kebakaran hutan dengan cara mengklasifikasikan jenis titik panas di lahan, sehingga di dapat skala prioritas dalam pemadaman titik api. Penelitian ini bertujuan mengklasifikasikan type titik panas dengan metode XGBoost dan *feature importance* yang terdapat di pulau Sumatera. Data titik panas diperoleh dari Globalforestwatch.com. Proses mengurangi variabel dari data yang diperoleh menghasil dampak yang sangat signifikan pada model klasifikasi. Terapat enam dan atau tujuh variabel yang sangat berpengaruh dalam menentukan titik panas, variabel tersebut jugalah yang menghasilkan model klasfikasi terbaik. XGBoost dan *feature importance* menghasilkan akurasi sebesar 89.52%. Sensitivity (SE), Specificity (SP), dan Matthews Correlation Coefficient (MCC).secara berturut turut 91.32 %, 93.16 % dan 92.75 %. Metode ini juga lebih baik dibandingkan dengan hasil penelitian sebelumnya.

Kata kunci: titik panas, XGBoost, *feature importance*

1. Pendahuluan

Hutan merupatan merupakan paru paru dunia. Data mencatat luas hutan Indonesia adalah 94,1 juta Ha atau 50,1% dari total daratan. Luas hutan Indonesia menempati posisi nomor 9 sebagai hutan terluas di Dunia. Sejak tahun 2001 hingga 2019, Indonesia kehilangan 26.8 Mha hutan dengan penurunan 17% setara dengan 10.9 Gt emisi CO_2 [1]. Kebakaran hutan dan lahan masih terus terjadi di Indonesia, terutama di Sumatera, Riau dan Kalimantan hingga Papua. Efek kebakaran hutan dan lahan yang terjadi akhir-akhir ini juga cukup menghawatirkan. Sebaran asap yang ditimbulkan sudah amat meluas, mencapai sebagian besar wilayah Sumatera dan Kalimantan, bahkan negara tetangga sempat merasakan dampaknya. Sepanjang tahun 2019, menurut data Kementerian Lingkungan Hidup dan Kehutanan, luas kebakaran hutan dan lahan di Indonesia mencapai 328.722 hektar. Di Kalimantan Tengah tercatat seluas 44.769 hektar, Kalimantan Barat 25.900 hektar, Kalimantan Selatan 19.490 hektar, Sumatera Selatan 11.826 hektar, Jambi 11.022 hektar dan Riau 49.266 hektar. Berdasarkan data tersebut jumlah kebakaran terparah terjadi di pulau Sumatera.

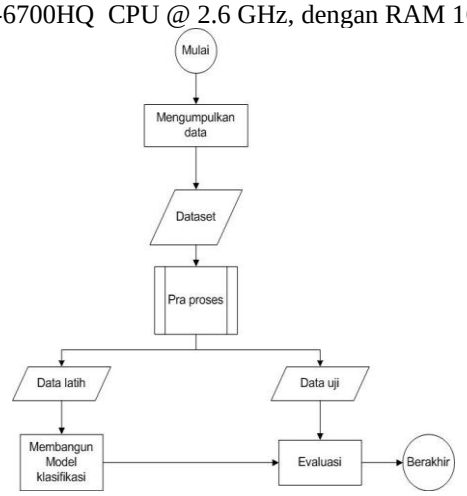
Prediksi kebakaran hutan merupakan sebuah tindakan pencegahan atau minimalisasi risiko dari dampak kebakaran hutan di masa depan [2]. Faktanya ada banyak jenis titik panas berdasarkan data satelit, bisa berasal dari volkano aktif, kebakaran lahan, kebakaran kota dsb. Tujuan utama adalah mengklasifikasikan untuk type titik api berdasarkan variable lokasi dan data satelit di area Sumatera. Manfaatnya adalah menentukan titik api yang prioritas untuk dipadamkan. Kebakaran lahan vegetasi menjadi titik api yang menjadi utama sedangkan titik panas di laut lepas dan vulkano menjadi tidak prioritas.

Sebuah penelitian menganalisis beberapa algoritma untuk mengklasifikasi titik api dengan menggunakan pendekatan penambangan data [3], dalam penelitiannya membandingkan hasil klassifkasi dari beberapa algoritma. Hasil penelitiannya mengemukakan bahwa algorithm Random Forest adalah algoritma terbaik untuk mengklasifikasikan untuk kasus tersebut. Informasi tentang algoritma terbaik lalu diimplementasi dalam suatu studi kasus untuk memprediksi kebakaran, seperti memprediksi titik api berdasarkan informasi klimatologuui di Borneo menggunakan algoritma Random Forest [2].

Dalam penelitian ini, upaya mengklasifikasikan titik api yang bersumber dari Global Forest Watch (GFW) dilakukan dengan mengimplementsikan algoritma XGBoost. Sebelum itu pra proses data dilakukan dengan normalisasi dan pemilihan variable dengan metode *feature importance*.

2. Metode Penelitian

Secara umum, proses dalam penelitian ini meliputi empat proses, yakni mengumpulkan data, pra-proses data, membangun model klasifikasi dan mengevaluasi model (gambar 1). Keseluruhan proses akan dijalankan oleh *processor* Intel® Core™ i7-6700HQ CPU @ 2.6 GHz, dengan RAM 16 GB.



Gambar 1 Alur proses penelitian

A. Dataset

Dalam penelitian ini, data kebakaran hutan dikumpulkan dari Global Forest Watch (GFW). GFW adalah platform daring yang menyediakan data dan alat untuk memantau hutan. Data yang berhasil dikumpulkan sebanyak 300 titik panas dengan 12 variabel (tabel 1) dan empat tipe titik panas. Contoh dataset dapat dilihat pada tabel (2).

Tabel 1 Deskripsi variabel

Variable	Deskripsi
Lat	*Latitude*
Long	*Longitude*
Bright_ti4	Temperatur kecerahan I-4 dalam kelvin
Scan	Ukuran scan dalam pixel
Track	Ukuran *Track* dalam pixel
Acq_Date	Tanggal akuisisi dari VIIRS
Acq_Time	Waktu akuisisi (dalam UTC).
Satellite (S)	N= *Suomi National Polar-orbiting Partnership* (Suomi NPP), 1=NOAA-20 (*designated* JPSS-1 *prior to launch*)
Confidence (C)	Nilai ini berdasarkan pada proses deteksi.
Version (V)	Versi
Bright_ti5	Temperatur kecerahan I-5
FRP	Kekuatan radiative titik panas
Type	Type titik panas 0 = kebakaran vegetasi 1 = Vulkano aktif 2 = Sumber lahan lainnya 3 = titik panas di lepas pantai

Tabel 2 Contoh dataset

latitude	longitude	bright_ti4	scan	track	acq_date	acq_time	S	C	V	bright_ti5	frp	type
-2.7868	120.3739	325.5	0.57	0.52	1/1/2018	455	N	n	1	290.7	4.7	0
-0.8244	127.8687	332.6	0.39	0.36	1/1/2018	455	N	n	1	292.5	3.7	3
-1.4326	132.3212	333.3	0.57	0.43	1/1/2018	455	N	n	1	290.1	5.5	0
0.90354	127.6283	337.7	0.39	0.36	1/1/2018	456	N	n	1	296.4	8.2	0
-0.8648	122.2072	329.5	0.4	0.44	1/1/2018	456	N	n	1	287.4	2.2	0
-0.3357	128.0134	331.2	0.39	0.36	1/1/2018	456	N	n	1	295.2	2	0
0.55568	121.8954	334.5	0.4	0.44	1/1/2018	456	N	n	1	286.8	3.8	0
0.52534	121.4961	340.5	0.42	0.46	1/1/2018	456	N	n	1	291.9	6.9	0
0.52593	121.4999	341.7	0.42	0.46	1/1/2018	456	N	n	1	292.3	11.4	0
3.08473	117.4119	340.7	0.37	0.58	1/1/2018	457	N	n	1	274.5	6.7	0

B. Pra-proses

Delapan puluh persen proses penambangan data adalah proses mengumpulkan data dan pra-proses[4]. Artinya sebagian besar energi penelitian dihabiskan untuk mengumpulkan dan mempersiapkan data sebelum diolah. Pada tahapan ini akan dilakukan pra proses data. Ada beberapa kondisi yang tidak ideal pada dataset yang diperoleh, pertama jumlah variable yang terlalu banyak sehingga sulit menetukan fokus dalam membangun model. Permasalahan yang kedua adalah skala dari masing masing nilai variabel yang yang berbeda jauh, kondisi tersebut dapat menyebabkan menurunkan akurasi model. Untuk mengatasi permasalahan yang pertama, peneliti melakukan penguarangan variabel dengan menggunakan teknik *feature importance* sedangkan untuk mengatasi permasalahan yang kedua peneliti melakukan normalisasi data.

Feature importance menghitung skor penting untuk setiap fitur dalam data set secara eksplisit. Hal ini memungkinkan fitur untuk diberi peringkat dan dibandingkan satu sama lain. Semakin tinggi skor maka data fitur atau atribut tersebut semakin relevan. Untuk memilih fitur yang akan dipakai dibutuhkan sebuah nilai batas (threshold). Fitur yang memiliki skor < nilai threshold akan dihapus sedangkan fitur yang memiliki skor > nilai threshold akan digunakan.

Normalisasi data dilakukan untuk menstandarkan data dari setiap variabel. Sehingga setiap nilai akan memiliki skala yang sama. Proses normalisasi data pada penelitian ini menggunakan metode z-score [5] metode normalisasi yang digunakan adalah z-score dengan formula pada persamaan (1) [5]. Notasi Z merepresentasikan nilai normaliasi, x adalah data , μ adalah nilai rata rata dari data dan σ adalah standar deviasi.

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

C. Metode Klasifikasi

Setelah proses normalisasi, hasilnya berupa dataset yang sudah terstandar. Proses selanjutnya adalah membagi data menjadi dua bagian, yakni data latih dan data test dengan menggunakan dengan prosentase 80 % data digunakan sebagai data latih dan 20 % data tes. Data latih digunakan untuk membangun model klasifikasi sedangkan data tes digunakan untuk menguji model. Proses pemodelan dilakukan dengan metode Extreme Gradient Boosting (XGBoost).

XGBoost adalah algoritma yang ditingkatkan berdasarkan gradient boosting decision tree dan dapat membangun boosted trees secara efisien dan beroperasi secara parallel[6]. XGBoost merupakan salah satu teknik pembelajaran mesin untuk mengatasi permasalahan regresi dan klasifikasi berdasarkan Gradient Boosting Decision Tree (GBDT)[7]. XGBoost pada dasarnya adalah metode ensemble yang didasarkan pada gradient boosting tree[6]. Didalam pohon regresi, nodes bagian dalam mewakili nilainilai untuk tes atribut dan leaf nodes dengan skor mewakili keputusan. Hasil prediksi adalah jumlah skor yang diprediksi oleh pohon K, seperti ditunjukkan pada persamaan (2):

$$\hat{y}_l = \sum_{k}^{K} f_k(x_i), f_k \ \in F \tag{2}$$

Metode penelitian berisikan tentang bagaimana penelitian dikerjakan yang di jelaskan secara detail. Pada setiap paragraph bisa terdiri dari beberapa subparagraph yang ditunjukan pada persamaan (3).

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_l) + \sum_{k}^{K} \Omega\,(f_k) \tag{3}$$

Dimana$\sum_{i=1}^{n} l(y_i, \hat{y}_l)$ adalah differentiable *loss function* untuk mengukur apakah model tersebut cocok untuk set data pelatihan dan $\sum_{k}^{K} \Omega\,(f_k)$ adalah item yang menentukkan kompleksitas model. Ketika kompleksitas model meningkat skor yang sesuai dikurangi nilainya. Sebelum membangun model prediksi, dilakukan optimasi parameter tuning XGBoost dengan beberapa parameter yang digunakan seperti pada Tabel 3.

Tabel 3 Parameter XGBoost

No	Parameter	Rentang Nilai
1	*Max_depth*	5
2	*seed*	7
3	*Test_size*	0.35
4	*Feature*	1-9
5	*Learning rate*	0.05

D. Validasi Model

Proses akhir dalam penelitian ini adalah validasi model. Beberapa penelitian tidak cukup menggnakan akurasi sebagai parameter validasi model klasifikasi[7,8]. Sehingga proses validasi perlu menggunakan parameter lainnya Sensitivity (SE), Specificity (SP), akurasi (Q) dan Matthews Correlation Coefficient (MCC). Setiap proses validasi diakumulasi kedalam confusion matrix (tabel 4). TP adalah kelas yang didata benar dan hasil model juga benar. FN adalah di data benar,sedangkan model memprediksinya salah. FP adalah kelas di data salah sedangkan hasil model memprediksi benar. TN adalah kelas yang salah dan model memprediksinya salah juga.

Tabel 4 Confusion Matrix

Kelas	Kelas Kelas 1 (prediksi)	Kelas Kelas 2 (prediksi)
Kelas 1 (Aktual)	*TP (True Positive)*	*FN (False Negative)*
Kelas 2 (Aktual)	*FP (False Positive)*	*TN (True Negative)*

Sensitivity (SE) merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. SE menjawab pertanyaan "Berapa persen titik panas yang diprediksi bukan titik panas dibandingkan keseluruhan mahasiswa yang sebenarnya titik panas. Formula untuk menghitung SE menggunakan persamaan (4). *Specifity* (SP) merupakan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif. Specificity menjawab pertanyaan "Berapa persen titik panas yang benar diprediksi tidak dibandingkan dengan keseluruhan titik panas sebenarnya tidak ". Formula untuk menghitung ini menggunakan persamaan (5). Akurasi merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan "Berapa persen titik api yang benar diprediksi benar dan tidak titik api dari kesuluruhan titik api". Formula untuk menghitung akurasi dengan persamaan (6). MCC merupakan ukuran kualitas klasifikasi biner yang mewakili korelasi antara klasifikasi biner yang diamati dan diprediksi. MCC akan mengembalikan nilai kedalam -1 dan 1, dengan nilai koefisien korelasi 1 mewakili prediksi benar dan nilai koefisien -1 sebagai prediksi salah[7]. Jadi, MCC bisa digunakan untuk setiap klasifikasi biner. Prameter tersebut dievaluasi dengan menggunakan persamaan (7) :

$$SE = \frac{TP}{TP + FN} \tag{4}$$

$$SP = \frac{TN}{TN + FP} \tag{5}$$

$$Q = \frac{TP + TN}{TN + TN + FN + FP} \tag{6}$$

$$MCC = \frac{(TP\ x\ TN)x\ (FP\ xFN)}{\sqrt{(TP + FP)(FN + TN)(TP + FN)(FP + TN)}} \tag{7}$$

3. Hasil dan Pembahasan

Terdapat 12 variabel awal pada dataset kebakaran hutan. Sebagian acuan awal, peneliti mengklasifikasi data tersebut dengan melibatkan semua variabel dan dan parameter di tabel 3 menggunakan metode XGBoost. Berdasarkan hasil klasifikasi diperoleh akurasi sebesar 82.51 % dengan *running time* 10 menit 35 detik. Kami percaya bahwa hasil tersebut tidaklah efektif, mengingat jumlah data yang diolah tidak besar dengan akurasi yang tidak menyentuh angka 85% terlebih lagi memakan waktu yang lama. Selanjutnya akan dipilih beberapa variabel terbaik dengan menggunakan *feature importance*.

A. Feature Importance

Proses pemilihan variabel dilakukan sebagai proses mengurangi variabel. Dari 12 variabel yang dianalisis, terdapat 9 variabel yang nilainya diatas *threshold* (tabel 5). Variabel *acq_time* merupakan variabel dengan nilai tertinggi yang terpilih dan variabel *confidence* merupakan variabel dengan nilai *feature importance* terendah yang terpilih. Sembilan variabel ini yang digunakan sebagai variabel dalam memodelkan.

Tabel 5 Nilai *feature importance*

Variabel	Nilai *important*
Latitude	73.2526
Longitude	51.3742
Bright_ti4	36.2183
Scan	13.3127
Track	11.1997
Acq_time	87.5616
confidence	0.6567
Bright_ti5	33.6024
frp	31.1704

B. XGBoost

Dibagian awal diklasifikasikan data kebakaran hutan dengan melibatkan seluruh variabel dan diperoleh hasil yang kurang baik. Pada bagian ini merupakan analisis hasil pemilihan variabel terbaik terhadap akurasi model. Berdasarkan hasil klasifikasi dengan menggunakan metode XGBoost diperoleh akurasi tertinggi ketika menggunakan enam atau tujuh variabel (tabel 6) yakni *acq_time, latitude, longitude, bright_ti4, bright_ti5, frp* dan atau *scan.*

Tabel 6 Pengaruh jumlah variabel terhadap akurasi

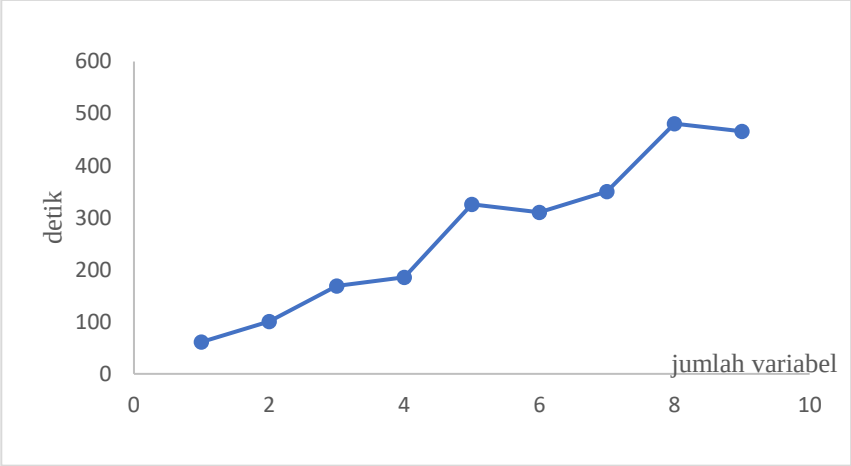
Jumlah Variabel	*Threshold*	Akurasi (%)
1	0.304	69.52
2	0.174	88.57
3	0.139	87.62
4	0.112	88.57
5	0.089	88.57
6	0.066	89.52
7	0.063	89.52
8	0.052	88.57
9	0	88.57

Informasi tersebut dilanjutkan dengan mengukur kinerja model dengan nilai SE dan SP. Nilai dari masing masing metode tertera pada tabel 7. Kelas 0 dapat dideteksi dengan baik oleh hasil model XGBoost, kelas 0 ini lah yang menjaadi fokus dari pemadaman titik panas, karena kelas 0 mengindentifkasi titik api di lahan vegetasi. Secara keseluruhan nilai dari SE, SP dan MCC dari metode XGBoost dengan enam variabel adalah 91.32%, 93.16 % dan 92.75 %.

Tabel 7 SE dan SP model XGBoost

Metode validasi	Kelas	Nilai (%)
SE	Type 0	94.36
	Type 1	47.81
	Type 2	14.52
	Type 3	50
SP	Type 0	87.76
	Type 1	3.64
	Type 2	24.62
	Type 3	0.03

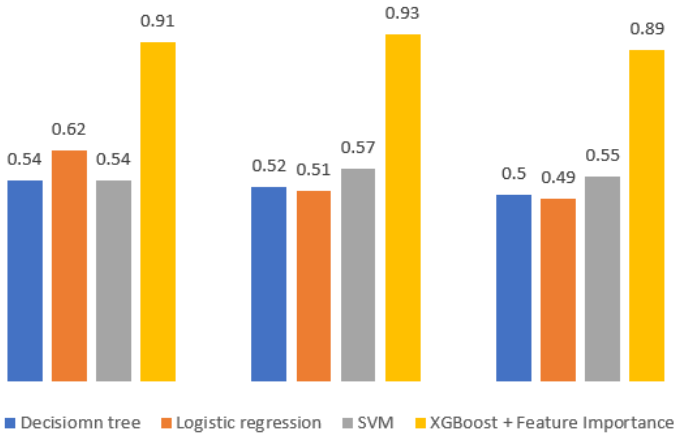
Sebagai tambahan, penulis juga mendata *running time* dari kerja algoritma XGBoost terhadap perubahan jumlah variabel (gambar 2). Secara umum penigkatan jumlah variabel yang digunakan akan meingkatkan waktu komputasi dari algoritma XGBoost. Namun yang perlu dicermati adalah perubahan waktu ketika menggunakan lima variabel dengan enam variabel, terjadi penurunan *running time* yang sangat signifikat dibandingkan dengan kondisi lainnya, misalnya perbandingan dari 3 variabel dengan empat variabel atau delapan variabel dengan Sembilan variabel. Sehingga temuan tersebut memperkuat hasil bahwa enam variabel terbaik hasil *feature importance* menghasilkan model klasifikasi terbaik dengan metode XGBoost.



Gambar 2 *running time*

C. Perbandingan dengan Metode lain

Pada bagian ini peneliti juga membandingkan kinerja dari beberpa algoritma klasifikasi dari penelitian sebelumnya. Hasil perbandingan disajikan pada gambar 3. Berdasarkan parameter validasi apapun, metode XGBoost dan *feature importance* memiliki hasil yang lebih baik dengan algoritma klasifikasi lainnya. Dengan kata lain *feature importance* memiliki pengaruh yang besar dalam meningkatakan kualitas model klasifikasi.



Gambar 3 Perbandingan dengan metode lain [9]

4. Kesimpulan

Feature importance membantu dalam pemilihan fitur terbaik dalam proses klasifikasi kebakaran hutan. Dari 12 feature, terdapat enam dan atau tujuh feature yang memiliki pengaruh terbesar dalam menghasilkan model klasifikasi. Dengan kata lain, enam atau tujuh variabel itulah yang sangat berpengaruh dalam menentukan hasil prediksi type titik panas. Adapun metode XGBoost dan *feature importance* menghasilkan akurasi sebesar 89.52%. SE, SP dan MCC secara berturut turut 91.32 %, 93.16 % dan 92.75 %. Metode ini juga lebih baik dibandingkan dengan hasil penelitian sebelumnya [9]. Sebagai catatan dalam penelitian ini terdapat informasi lokasi penangannya masih menggunakan cara klasik. Kedepannya akan ditangani secara khusus atribut lokasi dalam proses klasifikasi.

5. Daftar Rujukan

[1] Globalforestwatch.com (2020, 10 November). Kehilangan Hutan Primer Di Indonesia https://www.globalforestwatch.org/

[2] Latifah, A. L., Shabrina, A., Wahyuni, I. N., & Sadikin, R. (2019, October). Evaluation of Random Forest model for forest fire prediction based on climatology over Borneo. In *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* (pp. 4-8). IEEE.

[3] Molovtsev, M. D., & Sineva, I. S. (2019, September). Classification Algorithms Analysis in the Forest Fire Detection Problem. In *2019 International Conference" Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS)* (pp. 548-553). IEEE.

[4] Karo, I. M. K., & Huda, A. F. (2016, October). Spatial clustering for determining rescue shelter of flood disaster in South Bandung using CLARANS Algorithm with Polygon Dissimilarity Function. In *2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA)* (pp. 70-75). IEEE.

[5] Karo, I. M. K. (2017). *Spatial Clustering Based on Dissimilarity Region Using CLARANS with Polygon Dissimilarity Function* (Doctoral dissertation, TELKOM UNIVERSITY).

[6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[7] Cherif, I. L., & Kortebi, A. (2019, April). On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification. In *2019 Wireless Days (WD)* (pp. 1-6). IEEE.

[8] Karo, I. M. K., Ramdhani, R., Ramadhelza, A. W., & Aufa, B. Z. (2020, October). A Hybrid Classification Based on Machine Learning Classifiers to Predict Smart Indonesia Program. In *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)* (pp. 1-5). IEEE.

[9] Dimitrakopoulos, G. N., Vrahatis, A. G., Plagianakos, V., & Sgarbas, K. (2018, July). Pathway analysis using XGBoost classification in Biomedical Data. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (pp. 1-6).