Penerapan Teknik *Random Undersampling* untuk Mengatasi *Imbalance Class* dalam Prediksi Kebakaran Hutan Menggunakan Algoritma *Decision Tree*

Ilham Kurniawan[1], Duwi Cahya Putri Buani[2], Abdussomad[3], Widya Apriliah[4], Eka Fitriani[5]
[1,3,4] Universitas Bina Sarana Informatika, Karawang, Indonesia
[2] Universitas Nusa Mandiri, Jakarta, Indonesia
[5] Universitas Bina Sarana Informatika, Jakarta, Indonesia
Email: [1]ilham.imk@bsi.ac.id, [2]duwi.dcp@nusamandiri.ac.id, [3]abdussomad.bdu@bsi.ac.id, [4]widya.wyr@bsi.ac.id, [5]eka.ean@bsi.ac.id

Abstrak - Kebakaran hutan ialah bencana yang memicu kerusakan ekonomi dan ekologi juga meneror kehidupan manusia. Oleh karena itu, memprediksi problem lingkungan semacam kebakaran hutan benar-benar penting untuk meminimalisir ancaman kejadian bencana alam seperti kebakaran hutan. Dalam penelitian ini kami mengusulkan algoritma klasifikasi decision tree untuk memprediksi kebakaran hutan. Prediksi kebakaran hutan dilandaskan pada data meteorologi yang sesuai dengan elemen cuaca yang mempengaruhi terjadinya kebakaran hutan, yaitu suhu, kelembaban relatif dan kecepatan angin. Kami telah mendapati akurasi sekitar 94,52% mengenai prediksi kebakaran hutan dengan algoritma klasifikasi decision tree yang diusulkan. Nilai akurasi tersebut diperkuat dengan nilai ROC sebesar 0,950 yang melambangkan representasi dari algoritma klasifikasi yang dibangun untuk memprediksi kebakaran hutan, semakin mendekati angka 1 maka semakin baik juga algoritma klasifikasi yang dibangun..

Kata Kunci - ***Kebakaran hutan, algoritma klasifikasi, decision tree*.**

Abstract - Forest fires are disasters that cause economic and ecological damage and threaten human life. Therefore, predicting environmental problems such as forest fires is very important to reduce the threat of natural disasters such as forest fires. In this study we propose a decision tree classification algorithm for forest fires. Forest predictions are based on meteorological data that are in accordance with weather elements that affect the occurrence of forest fires, namely temperature, relative humidity and wind speed. We have obtained about 94.52% accuracy regarding forest fires with the proposed decision tree classification algorithm. a value of that magnitude with an ROC value of .950 which is a representation of the classification built for forest fires, the closer to number 1 the better the algorithm built.

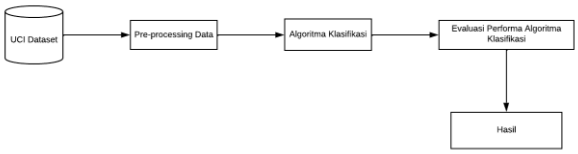
Keywords - forest fires, clasification algorithm, decision tree.

I. PENDAHULUAN

Kebakaran hutan ialah bencana lingkungan yang meneror keselamatan manusia, infrastruktur, dan heterogenitas hayati[1]. Perputaran iklim global menyebabkan penurunan curah hujan yang lumayan besar dan peningkatan suhu, yang selanjutnya mempengaruhi terjadinya kebakaran hutan [2]. Kebakaran hutan dapat mempengaruhi sebagian besar penduduk, mengakibatkan persoalan ekonomi dan hambatan terhadap bisnis dan persoalan kesehatan jangka pendek dan panjang [3][4][5]. Asap dan kabut yang dihasilkan oleh kebakaran juga mampu memicu gangguan dan defisit ekonomi bagi negeri-negeri jiran dalam ratusan juta [6]. Oleh karena itu, perlu untuk mampu memprediksi kebakaran hutan dan memilih langkah-langkah mitigasi untuk meminimalisir dampak buruknya [7]. Dalam *data mining* dan *machine learning*, tidak mudah untuk melatih model *machine learning* yang berhasil jika pembagian kelas pada data set tidak seimbang. Ini dikenal sebagai masalah *imbalance class* [8][9][10]. Satu kelas boleh jadi diwakili dengan sejumlah besar contoh, sebaliknya contoh yang lain diwakili dengan hanya beberapa contoh [11][12]. Random undersampling (RUS) secara acak menghapus instance dari kelas mayoritas hingga keseimbangan yang diinginkan tercapai [13].Tanpa memperhitungkan masalah *imbalance class*, algoritma atau model yang dibangun bisa dikuasai dengan kelas mayoritas dan mampu melupakan kelas minoritas. Sebagai contoh, misalnya sebuah dataset mempunyai 2 kelas dan rasio ketidakseimbangan datanya 99%, kelas mayoritas ialah 99% dari data set dan kelas minoritas hanya berisi 1% dari data [14][15]. Untuk mengurangi tingkat kekeliruan, algoritma yang dibangun mengklasifikasikan semua contoh ke dalam kelas mayoritas, yang menjadikan tingkat kesalahan 1%. Dalam hal ini, semua contoh milik kelas minoritas adalah yang terpenting dan harus diidentifikasi sebagai klasifikasi yang salah [16][17]. Dalam penelitian ini kami mengusulkan sebuah metode untuk memprediksi kebakaran hutan menggunakan integrasi teknik pendekatan level data yaitu Random Under-Sampling (RUS).

II. METODE PENELITIAN

Metode yang diusulkan disajikan dalam gambar 1 di bawah ini dengan bentuk diagram. Gambar 1 menampilkan alur penelitian yang dilakukan dalam membuat model.



Gambar 1. Metode penelitian

Dari konteks model penelitian yang sudah diusulkan dari Gambar 1. Diagram model yang diusulkan.

A. UCI Dataset

UCI *machine learning repository* ialah berkas basis data, teori domain, dan generator data yang digunakan komunitas *machine learning* untuk studi empiris algoritma *machine learning*. Dibuat seperti arsip ftp pada tahun 1987 oleh David Aha dan mahasiswa pascasarjana di UC Irvine [18]. UCI Dataset adalah dataset yang digunakan pada penelitian ini, dataset dari UCI *repository* yakni dataset *Algerian Forest Fires*, dengan jumlah data sebanyak 244 data, 14 atribut dan 1 kelas.

B. Pre-processing Data

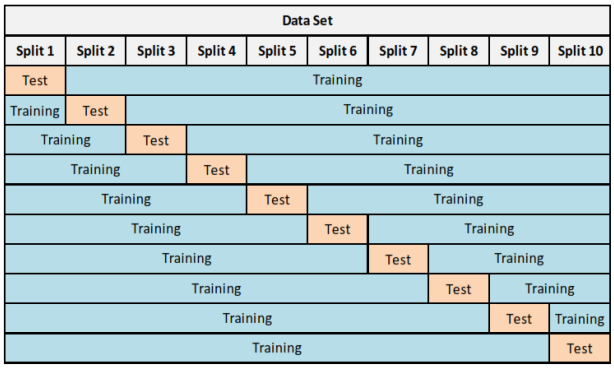
Pre-processing data merupakan langkah penting untuk menerapkan algoritma klasifikasi. Dalam pelatihan model *supervised learning* seperti *decision tree*, input data *training* sangat memengaruhi kinerja model, sehingga memiliki data yang diproses sebelumnya dan dianotasi dengan baik merupakan langkah penting dalam mencapai kinerja moedel algoritma klasifikasi yang baik [19].

C. Algoritma Klaifikasi

Algoritma klasifikasi yang digunakan pada penelitian ini adalah algoritma klasifikasi *Decision Tree*. Decision tree merupakan salah satu pendekatan sangat kuat yang digunakan pada pemodelan prediktif dalam *machine learning* dan *machine learning*. Peran algoritma decision tree dilandaskan pada rangkaian data variabel input dan membentuk model untuk prediksi nilai variabel dependen. Pembangunan *decision tree* direalisasikan dengan partisi satu set fitur ke dalam hierarki serta mengekspresikan kaitan antara fitur dan variabel dependen [20].

D. Evaluasi Performa Algoritma Klasifikasi

Percobaan dilakukan menerapkan 10 *fold cross validation*. *Cross-validation* (CV) yaitu proses statistik yang mampu digunakan untuk menilai kinerja model atau algoritma dimana data dipisahkan sebagai dua subset yakni data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih bagi subset pembelajaran dan dibuktikan oleh subset validasi. Selanjutnya penentuan jenis CV dapat dilandaskan pada ukuran dataset. Lazimnya CV K-fold digunakan karena mampu mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Pengukuran *Accuracy, F-Measure, Recall, Precision* dan ROC (*Receiver Operating Curve*), digunakan untuk klasifikasi penelitian ini.



Gambar 2. Pembagian dataset untuk 10-*fold cross validation*

III. HASIL DAN PEMBAHASAN

Penelitian ini memanfaatkan python machine learning melalui hardware Laptop processor AMD Ryzen 3 dengan RAM 4 GB dan hardisk 320GB. Penelitian ini menggunakan data dari UCI machine learning repository yaitu *Algerian Fire Forest* [21], dataset ini memiliki jumlah atribut 14 atribut dan 1 kelas dengan jumlah data sebanyak 244 data yang akan disajikan pada Tabel 1, dengan uraian seperti pada Tabel 2.

Tabel 1. Deskripsi dataset

Dataset	*Jumlah Atribut*	*Jumlah Data*
Algerian Fire Forest	14	244

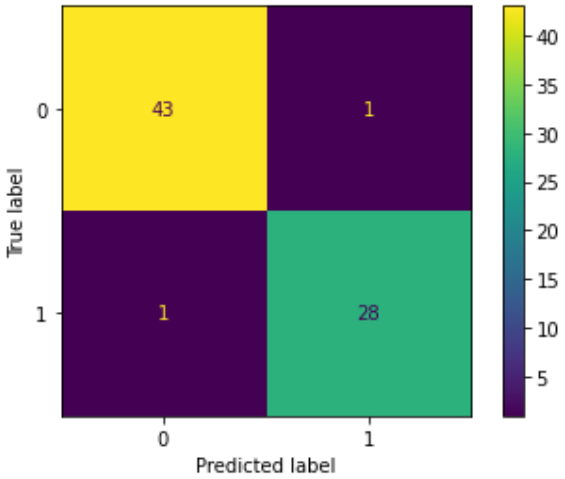
Tabel 2. Sampel dataset

Day	*Month*	*Year*	*Temperature*	*RH*	*Ws*	*Rain*	*FFMC*	*DM*
1	6	2012	29	57	18	0.0	65.7	1
2	6	2012	29	61	13	1.3	64.4	2
3	6	2012	26	82	22	13.1	47.1	3
4	6	2012	25	89	13	2.5	28.6	4
5	6	2012	27	77	16	0.0	64.8	5
6	6	2012	31	67	14	0.0	82.6	6
7	6	2012	33	54	13	0.0	88.2	7
8	6	2012	30	73	15	0.0	86.6	8
9	6	2012	25	88	13	0.2	52.9	9
10	6	2012	28	79	12	0.0	73.2	10

11	6	2012	31	65	14	0.0	84.5	11
12	6	2012	26	81	19	0.0	84.0	12
13	6	2012	27	84	21	1.2	50.0	13
14	6	2012	30	78	20	0.5	59.0	14
15	6	2012	28	80	17	3.1	49.4	15
16	6	2012	29	89	13	0.7	36.1	16
17	6	2012	30	89	16	0.6	37.3	17
18	6	2012	31	78	14	0.3	56.9	18
19	6	2012	31	55	16	0.1	79.9	19
20	6	2012	30	80	16	0.4	59.8	20
21	6	2012	30	78	14	0.0	81.0	6.3
22	6	2012	31	67	17	0.1	79.1	7.0
23	6	2012	32	62	18	0.1	81.4	8.2
24	6	2012	32	66	17	0.0	85.9	11.2
25	6	2012	31	64	15	0.0	86.7	14.2
26	6	2012	31	64	18	0.0	86.8	17.8
27	6	2012	34	53	18	0.0	89.0	21.6
28	6	2012	32	55	14	0.0	89.1	25.5
29	6	2012	32	47	13	0.3	79.9	18.4
30	6	2012	33	50	14	0.0	88.7	22.9

Metode *pre-processing data* yang digunakan pada penlitian ini yakni metode Resample dengan teknik *Random Under-Sampling*, gunanya untuk menciptakan subsampel acak dari kumpulan data melalui pengambilan sampel dengan penggantian atau tanpa penggantian, sehingga rasio perbedaan kelas mayoritas dengan kelas minoritas tidak terlalu signifikan.

Model algoritma klasifikasi yang diusulkan dievaluasi pada dataset kebakaran hutan yaitu menggunakan algoritma klasifikasi *dicision tree*. Percobaan dilakukan memanfaatkan metode *10-fold cross-validation. accuracy, f-measure, recall, precision* and ROC (*Receiver Operating Curve*) *measures* digunakan untuk klasifikasi penelitian ini, sehingga menghasilkan confussion matrix yang akan disajikan pada Gambar 3. Tabel.3 mendeskripsikan ukuran akurasi di bawah ini:



Gambar 3. *Confussion Matrix*

Tabel 3. Pengukuran akurasi

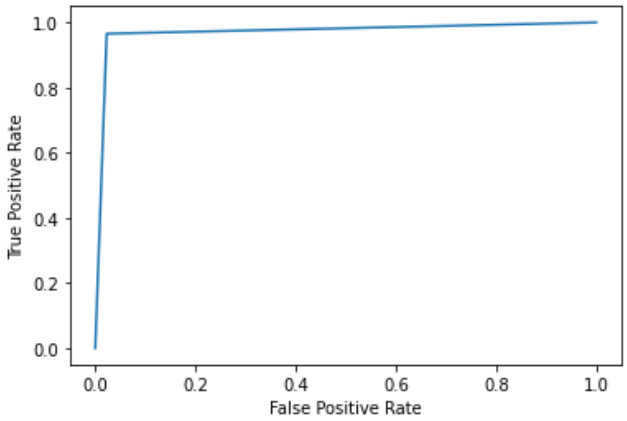
Pengukuran	*Definisi*	*Formula*
Accuracy (A)	Akurasi mendefinisikan keakuratan algoritma memprediksi instance	A=(TP+TN) / (Jumlah total sampel)
Prescision (P)	Classifier, correctness/accuracy diukur dengan Precision	P = TP / (TP + FP)
Recall (R)	Untuk menguji pengklasifikasi completeness atau sensitivity, memanfaatkan Recall	R = TP / (TP + FN)
F-Measure (F)	F-Measure ialah rata-rata dari precision dan recall.	F = 2*(P * R) / (P + R)
ROC	ROC (Receiver Operating Curve) digunakan untuk menimbang-nimbang kegunaan pengujian	

Tabel 4 menampilkan nilai kinerja dari algoritma klasifikasi yang dihitung dalam berbagai ukuran. Dari Tabel 4 dapat dianalisis algoritma klasifikasi *decision tree* menonjolkan nilai akurasi dan nilai akurasi sebesar 94,52% dan nilai ROC sebesar 0,950. ROC ialah gambaran dari algoritma klasifikasi yang dibangun untuk memprediksi kebakaran hutan, semakin mendekati angka 1 maka semakin bagus algoritma klasifikasi yang dibangun, metode *random under-sampling* dapat mengatasi

permasalahan *imbalance class* dan algoritma klasifikasi *decision tree* mampu memprediksi probabilitas kebakaran hutan dengan lebih akurat yang disajikan pada Gambar 4.

Tabel 4. Kinerja algoritma klasifikasi

Algoritma	*Precision*	*Recall*	*F-measure*	*Accuracy*	*ROC*
Decision Tree	0.950	0.950	0.950	0.954	0.950



Gambar 4. Hasil ROC measures

IV. KESIMPULAN DAN SARAN

Kebakaran hutan menjadi salah satu bencana paling umum yang tercatat menyebabkan rusaknya hutan berhektar-hektar. Kebakaran hutan menimbulkan ancaman tidak hanya untuk sumber daya hutan tetapi untuk flora dan fauna, yang secara serius mengganggu keanekaragaman hayati, ekosistem, dan lingkungan suatu daerah. Karena metode, teknik, dan tools data mining menjadi lebih menjanjikan untuk memprediksi kebakaran hutan dan pada akhirnya mengurangi resiko terjadinya kebakaran hutan. Kontribusi utama penelitian ini adalah untuk mengetahui algoritma klasifikasi terbaik untuk prediksi terjadinya kebakaran hutan. Percobaan dilakukan pada dataset Algerian Fire Direst yang diambil dari UCI repository. Hasil percobaan menentukan kecukupan sistem yang dirancang dengan akurasi yang dicapai sebesar 95,52%. Kedepannya, sistem yang dirancang dengan algoritma klasifikasi machine learning dapat digunakan untuk memprediksi atau mendiagnosis permsalahan yang lainnya. Kedepannya, sistem yang dirancang dengan algoritma klasifikasi machine learning dapat digunakan untuk memprediksi permasalahan lainnya. Penelitian dapat diperpanjang dan ditingkatkan untuk prediksi kemungkinan kebakaran hutan dengan algoritma machine learning lainnya atau bahkan dengan menggunakan sistem optimasi seperti particle swarm optimization.

DAFTAR PUSTAKA

[1] D. Tien Bui, Q. T. Bui, Q. P. Nguyen, B. Pradhan, H. Nampak, dan P. T. Trinh, "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area," *Agric. For. Meteorol.*, vol. 233, hal. 32–44, 2017.

[2] K. Tshering, P. Thinley, M. Shafapour Tehrany, U. Thinley, dan F. Shabani, "A Comparison of the Qualitative Analytic Hierarchy Process and the Quantitative Frequency Ratio Techniques in Predicting Forest Fire-Prone Areas in Bhutan Using GIS," *Forecasting*, vol. 2, no. 2, hal. 36–58, 2020.

[3] Z. Jaafar dan T. L. Loh, "Linking land, air and sea: Potential impacts of biomass burning and the resultant haze on marine ecosystems of Southeast Asia," *Glob. Chang. Biol.*, vol. 20, no. 9, hal. 2701–2707, 2014.

[4] R. A. Chisholm, L. S. Wijedasa, dan T. Swinfield, "The need for long-term remedies for Indonesia's forest fires," *Conserv. Biol.*, vol. 30, no. 1, hal. 5–6, 2016.

[5] V. Huijnen, M. J. Wooster, J. W. Kaiser, D. L. A. Gaveau, J. Flemming, dan M. Parrington, "Fire carbon emissions over maritime southeast Asia in 2015 largest since 1997," *Nat. Publ. Gr.*, no. February, hal. 1–8, 2016.

[6] S. H. Doerr dan C. Santı, "Global trends in wildfire and its impacts : perceptions versus realities in a changing world," 2016.

[7] S. Yang, M. Lupascu, dan K. S. Meel, "Predicting Forest Fire Using Remote Sensing Data And Machine Learning," *35th AAAI Conf. Artif. Intell. AAAI 2021*, vol. 17A, hal. 14983–14990, 2021.

[8] G. Douzas dan F. Bacao, "Effective Data Generation for Imbalanced Learning sing Conditional Generative Adversarial Networks," *Expert Syst. Appl.*, 2017.

[9] A. Pinto, D. Ferreira, C. Neto, A. Abelha, dan J. Machado, "Data mining to predict early stage chronic kidney disease," *Procedia Comput. Sci.*, vol. 177, no. 2018, hal. 562–567, 2020.

[10] L. In, R. Using, dan L. Based, "TACKLING IMBALANCED CLASS IN SOFTWARE DEFECT PREDICTION USING TWO-STEP CLUSTER BASED RANDOM UNDERSAMPLING AND STACKING TECHNIQUE," *J. Teknol.*, vol. 2, hal. 45–50, 2017.

[11] V. J. L. Gan, I. M. C. Lo, J. Ma, K. T. Tse, J. C. P. Cheng, dan C. M. Chan, "Enhanced automatic twin support vector machine for imbalanced data classification," *J. Pre-proof*, hal. undefined-undefined, 2020.

[12] I. Kurniawan, Abdussomad, M. F. Akbar, D. F. Saepudin, M. S. Azis, dan M. Tabrani, "Improving the Effectiveness of Classification Using the Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention

Prediction," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020.

[13] A. R. Hassan dan M. I. H. Bhuiyan, "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Comput. Methods Programs Biomed.*, vol. 140, hal. 201–210, 2017.

[14] X. Zhu, Y. Liu, Z. Qin, dan J. Li, "Emotion Classification with Data Augmentation Using Generative Adversarial Networks."

[15] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, dan S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Comput. Electr. Eng.*, vol. 0, hal. 1–12, 2018.

[16] M. F. Akbar, I. Kurniawan, dan A. Fauzi, "Mengatasi Imbalanced Class Pada Software Defect Prediction Menggunakan Two-Step Clustering-Based Undersampling dan Bagging Tehcnique," *J. Inform.*, vol. 6, no. 1, hal. 107–113, 2019.

[17] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, dan Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, hal. 1623–1637, 2015.

[18] R. Hagan, C. J. Gillan, dan F. Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease," *Informatics Med. Unlocked*, vol. 24, hal. 100606, 2021.

[19] J. Bernal *et al.*, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artif. Intell. Med.*, vol. 95, no. April, hal. 64–81, 2019.

[20] G. Koulinas, P. Paraschos, dan D. Koulouriotis, "A decision trees-based knowledge mining approach for controlling a complex production system," *Procedia Manuf.*, vol. 51, no. 2019, hal. 1439–1445, 2020.

[21] F. Abid dan N. Izeboudjen, *Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm*, vol. 1105 AISC. 2020.