Kongres Internasional ke-8 tentang Informatika Terapan Lanjutan (IIAI-AAI) 2019

# Hutan Acak Baru dan Aplikasinya pada Klasifikasi Kualitas Udara

Hualing Yi
*Sekolah Data Besar & Perangkat Lunak Teknik Universitas Chongqing*
Chongqing, Cina
yihualing@cqu.edu.cn

Qingyu Xiong
*Sekolah Data Besar & Perangkat Lunak Teknik Universitas Chongqing*
Chongqing, Cina
xiong03@cqu.edu.cn

Qinghong Zou
*Sekolah Data Besar & Perangkat Lunak Teknik Universitas Chongqing*
Chongqing, Cina
cquzouqh@163.com

Rui Xu
*Sekolah Data Besar & Perangkat Lunak Teknik Universitas Chongqing*
Chongqing, Cina
cquxr@cqu.edu.cn

*Sekolah Otomasi* Kai Wang
*Universitas Chongqing*
Chongqing, Cina
akyle@163.com

Min Gao
*Sekolah Data Besar & Perangkat Lunak Teknik Universitas Chongqing*
Chongqing, Cina
gaomin@cqu.edu.cn

**Abstrak-Polusi *udara* memiliki dampak serius pada kehidupan sehari-hari. Penting untuk menginformasikan kualitas udara secara tepat waktu kepada masyarakat agar dapat mengambil tindakan sebelumnya. Metode pembelajaran mesin seperti random forest sangat baik dalam mengevaluasi nilai kualitas udara. Kami menemukan bahwa distribusi data udara tidak seimbang, yang menyebabkan efek negatif pada pengklasifikasi random forest. Kami mengusulkan sebuah metode hutan acak berdasarkan sampel yang dikelompokkan bootstrap untuk memecahkan masalah ini. Kemudian kami merancang tiga set eksperimen untuk mengevaluasi kinerja dari metode yang diusulkan. Hasil eksperimen menunjukkan bahwa metode yang diusulkan memberikan peningkatan terhadap random forest ketika keduanya diterapkan pada dataset yang seimbang. Peningkatannya sangat signifikan ketika diterapkan pada dataset yang tidak seimbang, di mana metode baru jauh lebih baik dalam mengklasifikasikan sampel minoritas.**

***Kata kunci-kualitas udara, set data ketidakseimbangan, koefisien ketidakseimbangan, hutan acak, bootstrap***

## I. PENDAHULUAN

Akhir-akhir ini, polusi udara menjadi semakin serius. Masalah seperti kabut asap sering terjadi di perkotaan. Kehidupan penduduk perkotaan sangat dipengaruhi oleh kualitas udara, terutama kesehatan. Oleh karena itu, untuk mengurangi dampak polusi udara terhadap kesehatan penduduk perkotaan, perlu dilakukan evaluasi kualitas udara secara ilmiah dan akurat. Metode yang didasarkan pada pembelajaran mesin bekerja dengan baik dalam tugas ini karena data udara yang sangat besar dalam sejarah.

Metode tradisional untuk mengevaluasi kualitas udara perkotaan umumnya menggunakan rumus tetap berdasarkan matematika fuzzy untuk membuat pemetaan antara data udara dan tingkat kualitas udara, yang terbukti memiliki kinerja yang buruk dengan toleransi dan efisiensi yang rendah [1]. Dengan akumulasi yang cepat dari sejumlah besar data udara, ada beberapa metode evaluasi berdasarkan pembelajaran mesin yang diusulkan oleh para ahli. Metode evaluasi jaringan syaraf tiruan yang diusulkan oleh Bai *dkk.* [2] memiliki kemampuan belajar mandiri dan adaptif, tetapi membutuhkan jumlah data yang besar dengan komputasi yang kompleks. Metode yang disebutkan dalam [3] didasarkan pada SVM

yang efisien dalam memproses masalah dua klasifikasi, tetapi berkinerja buruk dalam tugas multi-klasifikasi.

Kemampuan generalisasi yang baik yang didapatkan dari model random forest telah memotivasi random forest menjadi salah satu algoritma yang paling banyak digunakan di area data mining [4]. Dalam makalah ini, kami mengusulkan sebuah metode yang didasarkan pada random forest dan menilai kinerjanya dengan merancang tiga set eksperimen. Hasil penelitian membuktikan bahwa metode ini efektif dalam evaluasi kualitas udara dan memiliki kinerja yang lebih baik dibandingkan dengan random forest.

Bagian selanjutnya dari makalah ini disusun sebagai berikut. Pada bagian II, kami menjelaskan data udara yang kami gunakan dalam makalah ini. Pada bagian III, kami memperkenalkan konsep dasar algoritma random forest. Pada bagian IV, kami menjelaskan algoritma yang diusulkan. Pada bagian V, kami menyajikan eksperimen yang kami rancang dan menganalisa hasilnya. Terakhir, bagian VI dikhususkan untuk kesimpulan.

## II. DATASET YANG DIPELAJARI

### A. *Deskripsi Data*

Data kualitas udara yang digunakan dalam makalah ini berasal dari situs web Stasiun Pemantauan Lingkungan China [5]. *Dataset1* adalah data udara dari Beijing. *Dataset2* dan *dataset3* adalah data udara yang berbeda dari Fangchenggang, provinsi Guangxi. *Dataset4* adalah data udara dari Beijing dan Fangchenggang. Ruang fitur dataset terdiri dari 6 polutan udara ($PM2.5$, $PM10$, $SO_2$, $NO_2$, $O_3$, $CO$). Nilai fitur adalah nilai konsentrasi polutan udara. Label set data adalah nilai ($1,2,3,4,5,6$) kualitas udara. Setiap nilai sesuai dengan tingkat kualitas udara, yang disajikan dalam TABEL I.

### B. *Analisis Data*

Kami mendefinisikan variabel $Ic$ yang disebut koefisien ketidakseimbangan untuk mengukur tingkat ketidakseimbangan suatu set data. Semakin besar $Ic$, semakin tidak seimbang set data tersebut. Kami menjelaskan proses untuk menghitung koefisien ketidakseimbangan suatu set data sebagai berikut:

Program Sains & Teknologi Utama Guangxi (Hibah No. GKAA17129002)
Program Penelitian Utama Komisi Sains & Teknologi Chongqing (Hibah No. CSTC2017jcyjBX0025)

978-1-7281-2627-2/19/$31.00 ©2019 IEEE
DOI 10.1109/IIAI-AAI.2019.00018

- (Langkah 1) Biarkan $C(C>=2)$ menjadi jumlah kategori dalam himpunan $S$. $i, j$ adalah bilangan bulat antara 1 dan $N(1<= i, j <= C)$.
- (Langkah 2) Untuk setiap kategori dalam $S$, hitung jumlah sampel dalam $S$ yang termasuk dalam kategori tersebut, dinotasikan sebagai $x_i$.
- (Langkah 3) $X$ adalah himpunan $x_i$. Untuk setiap $x_i$, $x_j$ dalam himpunan $X$ dan $0<x_i < x_j$, koefisien ketidakseimbangan antara kategori $i$ dan kategori $j$ didefinisikan sebagai:

$$Ic = \frac{x_j}{x_i} \qquad (1)$$

- (Langkah 4) Biarkan $I_{cx}$ menjadi himpunan $Ic$ yang dihasilkan pada langkah3, koefisien ketidakseimbangan $S$ didefinisikan sebagai:

$$Ic(s) = \frac{\sum Ic_x}{C} \qquad (2)$$

TABEL II menunjukkan tingkat ketidakseimbangan dari 4 dataset yang digunakan dalam makalah ini. *Dataset2* adalah yang paling tidak seimbang dengan koefisien ketidakseimbangan terbesar di antara 4 dataset. *Dataset3* adalah yang tidak seimbang kedua, diikuti oleh *dataset1*. Sedangkan *dataset4* adalah yang paling seimbang di antara 4 dataset.

## III. ALGORITMA HUTAN ACAK ASLI

### *A. Teori Dasar Hutan Acak*

Random Forest (RF) [6] adalah metode klasifikasi yang diawasi dengan baik berdasarkan kombinasi dari "bagging" Breiman
[7] dan pemilihan fitur secara acak [8], yang beroperasi dengan membangun beberapa pohon keputusan selama proses pelatihan. Prediksi akhir adalah agregasi dari keputusan yang dibuat oleh pohon-pohon di dalam hutan. Suara mayoritas digunakan dalam proses agregasi. Proses pelatihan random forest dibagi menjadi dua bagian: pengambilan sampel secara acak dan pemisahan secara penuh.

- Pengambilan sampel secara acak pada rekaman dan fitur

Pengacakan RF diinduksi oleh bootstrap dan subruang fitur acak. Bootstrap digunakan untuk memilih rekaman secara acak dari set data asli. Kumpulan sampel bootstrap dengan ukuran N diambil secara acak dengan penggantian dari kumpulan data asli, yang juga berukuran N [9]. Jadi, beberapa sampel dapat muncul lebih dari satu kali setelah bootstrap. Subruang fitur acak digunakan untuk memilih fitur secara acak dari ruang fitur asli tanpa penggantian. Jumlah fitur yang dipilih tidak lebih dari ukuran fitur asli dari dataset. Pengambilan sampel secara acak pada record dan fitur memungkinkan untuk melatih setiap pohon di dalam hutan dengan sampel yang sangat berbeda, sehingga random forest dapat menghindari overfitting dengan baik.

- Memisahkan Sepenuhnya:

Setiap pohon di hutan dibangun tanpa pemangkasan [6]. Dengan cara ini, sebuah pohon cenderung lebih berbeda dari yang lain. Keanekaragaman pohon adalah alasan penting lainnya untuk menghindari pemangkasan yang

Dalam dataset, bootstrap mungkin menghasilkan tiga jenis set sampel: dataset tanpa sampel minoritas, dataset dengan beberapa sampel minoritas tetapi juga koefisien ketidakseimbangan yang lebih besar daripada dataset asli, dataset dengan beberapa sampel minoritas dan koefisien ketidakseimbangan yang sama atau lebih kecil dibandingkan dengan dataset asli. Namun, jenis set sampel yang pertama dan kedua tidak dapat digunakan. Karena pohon keputusan yang dilatih berdasarkan dua jenis set sampel ini dapat mengganggu pemungutan suara akhir, yang pada akhirnya mengarah pada penurunan kinerja untuk

pengklasifikasi hutan acak.

TABEL I. HUBUNGAN ANTARA KUALITAS UDARA DAN KUALITAS UDARA

| Nilai (Label) | Deskripsi Kualitas Udara | Siapa yang perlu diperhatikan |
|---|---|---|
| *1* | Bagus. | Tidak perlu khawatir |
| *2* | Sedang | Beberapa orang yang mungkin tidak biasa sensitif terhadap ozon. |
| *3* | Tidak Sehat untuk Kelompok Sensitif | Kelompok sensitif meliputi: orang dengan penyakit paru-paru seperti asma, orang dewasa yang lebih tua, anak-anak dan remaja, dan orang-orang yang aktif di luar ruangan. |

kin
est
[10]. Kami berasumsi bahwa efek negatif tersebut dapat disebabkan oleh resampling dengan bootstrap dalam algoritma random forest. Bootstrap berarti pengambilan sampel dengan penggantian, jadi untuk setiap sampel dalam dataset, probabilitas untuk terpilih adalah sama [11]. Ketika terjadi ketidakseimbangan

TABEL II. TINGKAT KETIDAKSEIMBANGAN KUMPULAN DATA UDARA

| Dataset | Statistik kategori | | Koefisien ketidakseimbangan |
|---|---|---|---|
| *Dataset1* | *1* | 334 | 10.77 |
| | *2* | 733 | |
| | *3* | 412 | |
| | *4* | 251 | |
| | *5* | 133 | |
| | *6* | 46 | |
| *Dataset2* | *1* | 21443 | 1168.1 |
| | *2* | 12080 | |
| | *3* | 1266 | |
| | *4* | 135 | |
| | *5* | 39 | |
| | *6* | 6 | |
| *Dataset3* | *1* | 370 | 40.06 |
| | *2* | 204 | |
| | *3* | 30 | |
| | *4* | 10 | |
| | *5* | 5 | |
| | *6* | 2 | |
| *Dataset4* | *1* | 334 | 2.67 |
| | *2* | 334 | |
| | *3* | 320 | |
| | *4* | 318 | |
| | *5* | 318 | |
| | *6* | 285 | |

TABEL III. KONFIGURASI PARAMETER HIPER

| Parameter hiper | Nilai-nilai | Deskripsi |
|---|---|---|
| *kriteria* | "gini" | Ketidakmurnian Gini untuk set data |
| *n_estimator* | 178 | Jumlah pohon di hutan. |
| *max_features* | Tidak ada | Jumlah fitur yang perlu dipertimbangkan ketika mencari split terbaik. |
| *max_depth* | Tidak ada | Kedalaman maksimum pohon. Tidak ada berarti simpul diperluas sampai semua sampai semua daun mengandung kurang dari *min_sampel_split* sampel |
| *bootstrap* | Benar (RF) | Apakah dataset bootstrap digunakan saat membangun pohon. |
| *sbootstrap* | Benar (SGB-RF) | Apakah sampel dikelompokkan secara bootstrap dataset digunakan saat membangun pohon. |

| | | |
|---|---|---|
| *min_samples_split* | 5 | Jumlah sampel minimum yang diperlukan untuk membagi node internal. |
| *min_sampel_daun* | 2 | Jumlah sampel minimum harus berada di simpul daun |

## IV. ALGORITMA YANG DIUSULKAN SGB-RF

Pada bagian ini, kami akan memperkenalkan modifikasi dari algoritma RF asli, yang dapat beradaptasi dengan baik pada set data yang tidak seimbang. Algoritma RF yang telah diperbaiki dinamakan Samples Grouped Bootstrap based Random Forest, disingkat SGB-RF.

Algoritma SGB-RF menggunakan metode bootstrap sampel-terkelompok alih-alih metode bootstrap dalam fase pengambilan sampel acak. Metode bootstrap sampel-terkelompok didefinisikan sebagai berikut:

- (Langkah 1) Biarkan $C$ menjadi jumlah kategori dalam set sampel asal $S$. Kelompokkan dataset asli $S$ berdasarkan label kelas, sampel dengan label kelas yang sama berada dalam kelompok yang sama. Ada $C$ set sampel baru yang dihasilkan, yang berbeda satu sama lain dengan label kelas yang berbeda.
- (Langkah2) Gunakan metode bootstrap untuk mengambil sampel ulang pada setiap set sampel yang dihasilkan pada langkah1 untuk menghasilkan set sampel bootstrap $C$ lainnya.
- (Langkah 3) Gabungkan set sampel $C$ yang dihasilkan pada langkah 2 ke set sampel baru yang dilambangkan sebagai $S_g$, yang ukurannya sama dengan S.

Metode bootstrap berkelompok sampel menjamin keacakan sambil mempertahankan tingkat ketidakseimbangan set sampel baru $S_g$. Di satu sisi, samples-grouped bootstrap mewarisi skema bootstrap ketika mengambil sampel dalam subsampel dengan label yang sama. Di sisi lain, koefisien ketidakseimbangan set data baru $S_g$ sama dengan koefisien ketidakseimbangan set data asli $S$.

## V. EKSPERIMEN DAN ANALISIS

Untuk menilai kinerja metode yang diusulkan, kami merancang tiga set eksperimen berdasarkan set data dalam TABEL
II. Dalam percobaan, pohon keputusan dibangun dengan menggunakan algoritma CART yang menggunakan indeks Gini untuk mengukur kemurnian node dan menggunakan indeks Gini berbasis jarak minimum untuk memilih atribut pemisah [12]. Pada percobaan berikut, beberapa parameter yang dapat dikonfigurasi dikonfigurasikan sebagai TABEL III, yang kami pilih yang terbaik dengan bereksperimen dengan strategi pencarian acak.

### *A. Percobaan (1)*

Kami melatih satu set pengklasifikasi RF berdasarkan *dataset1* dan melatih satu set pengklasifikasi RF lainnya berdasarkan *dataset4.* Keduanya membuat prediksi pada *dataset3.* Laporan klasifikasi prediksi dari dua set pengklasifikasi RF disajikan pada TABEL IV dan TABEL V. Menurut nilai presisi, recall, dan F1 dari laporan klasifikasi, pengklasifikasi RF berdasarkan *dataset4* berkinerja lebih baik daripada pengklasifikasi RF berdasarkan *dataset1* sebesar 2 poin persentase. Dari TABEL II, koefisien ketidakseimbangan *dataset1* adalah 10.77, sedangkan koefisien ketidakseimbangan *dataset4* adalah 2.67. Hasil percobaan ini membuktikan bahwa kinerja pengklasifikasi RF berhubungan dengan tingkat ketidakseimbangan dataset. Pengklasifikasi RF yang dilatih pada dataset yang tidak seimbang, biasanya memiliki kinerja yang tidak memuaskan.

TABEL IV. LAPORAN KLASIFIKASI PENGKLASIFIKASIAN RF BERDASARKAN DATASET1

| Label | Laporan Klasifikasi | | | |
|---|---|---|---|---|
| | *Presisi* | *Ingat* | *F1 - skor* | *Dukungan* |
| *1* | 1.00 | 0.89 | 0.94 | 370 |
| *2* | 0.83 | 0.92 | 0.87 | 204 |
| *3* | 0.64 | 0.93 | 0.76 | 30 |
| *4* | 0.82 | 1.00 | 0.90 | 9 |
| *5* | 1.00 | 1.00 | 1.00 | 5 |
| *6* | 0.00 | 0.00 | 0.00 | 2 |
| rata-rata / | 0.90 | 0.89 | 0.89 | 620 |

TABEL V. LAPORAN KLASIFIKASI PENGKLASIFIKASI RF BERDASARKAN DATASET4

| Label | Laporan Klasifikasi | | | |
|---|---|---|---|---|
| | *Presisi* | *Ingat* | *F1 - skor* | *Dukungan* |
| *1* | 0.99 | 0.91 | 0.95 | 370 |
| *2* | 0.84 | 0.90 | 0.87 | 204 |
| *3* | 0.60 | 1.00 | 0.75 | 30 |
| *4* | 1.00 | 0.89 | 0.94 | 9 |
| *5* | 1.00 | 1.00 | 1.00 | 5 |
| *6* | 1.00 | 1.00 | 1.00 | 2 |
| rata-rata / | 0.93 | 0.91 | 0.91 | 620 |

*B. Percobaan (2)*

Eksperimen ini dirancang untuk memverifikasi bahwa algoritma SGB-RF dapat meningkatkan performa klasifikasi pada dataset yang tidak seimbang dibandingkan dengan RF. Sampel pelatihan berasal dari *dataset2*, sedangkan sampel pengujian berasal dari *dataset1*. Pengklasifikasi RF dan pengklasifikasi SGB-RF dilatih dengan konfigurasi hyper-parameter yang ditunjukkan pada TABEL III.

Kami menggunakan confusion matrix untuk mengevaluasi kinerja klasifikasi. Confusion matrix adalah cara yang sangat efektif untuk menilai akurasi dataset yang tidak seimbang, di mana akurasi setiap kategori digambarkan dengan jelas bersama dengan kesalahan inklusi dan kesalahan eksklusi yang ada dalam klasifikasi [13]. Gbr. 1, Gbr. 2 adalah matriks kebingungan dari pengklasifikasi RF dan pengklasifikasi SGB-RF yang diuji pada *dataset1*. Untuk set sampel minoritas, yang berukuran 46 dan label kelas *"6"*, semua sampel salah diklasifikasikan pada pengklasifikasi RF sementara 16 dari 46 sampel diklasifikasikan dengan benar oleh pengklasifikasi SGB-RF. Untuk sampel yang labelnya *"3"* dan *"5"*, akurasi pengklasifikasi SGB-RF lebih tinggi daripada akurasi pengklasifikasi RF. Untuk sampel lainnya yang berlabel *"1", "2"* dan *"4"*, pengklasifikasi RF dan pengklasifikasi SGB-RF mendapatkan kinerja yang sama. Hasil percobaan ini membuktikan bahwa SGB-RF memiliki kinerja yang lebih baik daripada RF, terutama dalam klasifikasi pada sampel minoritas.

*C. Percobaan (3)*

Untuk memverifikasi lebih lanjut keefektifan SGB-RF, kami juga mengevaluasi metode yang diusulkan pada 4 dataset dari repositori UCI [14]. TABEL VI menunjukkan deskripsi dataset yang digunakan dalam percobaan ini, termasuk koefisien ketidakseimbangan untuk mengukur tingkat ketidakseimbangan dataset. Berdasarkan koefisien ketidakseimbangan, *Glass* adalah yang paling tidak seimbang, diikuti oleh *Wine>Breast>Iris*.

Breiman telah memberikan bukti empiris dalam [15] untuk menunjukkan bahwa estimasi out-of-bag sama akuratnya dengan menggunakan set pengujian dengan ukuran yang sama dengan set pelatihan. Jadi dalam percobaan ini kami mengevaluasi kinerja pengklasifikasi SGB-RF dan pengklasifikasi RF dengan skor out-of-bag. Semakin tinggi skornya, semakin baik sebuah model. Gbr.3 menunjukkan hasil yang diperoleh dari percobaan ini. Terlihat bahwa SGB-RF memberikan peningkatan (sekitar 1 poin persentase)

dari RF ketika keduanya diterapkan pada dataset *Iris* yang seimbang. Peningkatannya signifikan ketika keduanya diterapkan pada dataset ketidakseimbangan (hingga 6 poin persentase pada *Glass*).

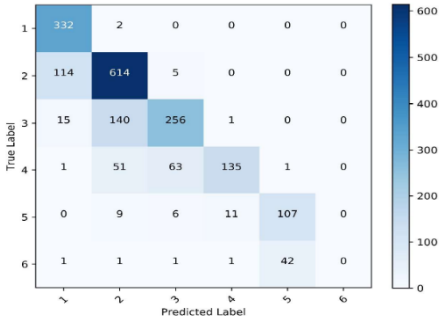
## VI. KESIMPULAN

Polusi udara adalah masalah serius bagi kesehatan penduduk, jadi penting untuk membuat evaluasi yang akurat tentang tingkat kualitas udara sesegera mungkin. Tingkat kualitas udara membantu penduduk mengambil tindakan sebelumnya untuk mengurangi efek polusi udara terhadap kesehatan. Dalam makalah ini, kami menemukan bahwa distribusi data udara tidak seimbang, yang menyebabkan penurunan kinerja dari pengklasifikasi hutan acak. Untuk beradaptasi dengan set data yang tidak seimbang, kami mengusulkan algoritma random forest berbasis bootstrap yang dikelompokkan sampel (SGB-RF). Melalui tiga set percobaan, kami membuat pengamatan berikut: kinerja pengklasifikasi random forest tidak memuaskan pada dataset yang tidak seimbang. SGB-RF menyajikan peningkatan dari RF ketika keduanya diterapkan pada dataset yang seimbang. Peningkatannya signifikan ketika keduanya diterapkan pada dataset yang tidak seimbang, di mana SGB-RF jauh lebih baik daripada RF dalam hal kemampuan yang kuat untuk mengklasifikasikan sampel minoritas dengan benar.

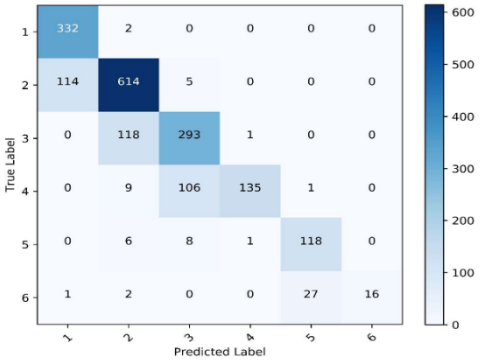
Untuk penelitian selanjutnya, kami akan membandingkan algoritma yang diusulkan dengan metode ensemble berdasarkan boosting, dan algoritma yang diusulkan akan dieksplorasi lebih lanjut untuk aplikasi yang lebih luas pada dataset lainnya.


## UCAPAN TERIMA KASIH

Pekerjaan ini didukung oleh Program Sains & Teknologi Utama Guangxi (Hibah No. GKAA17129002) dan Program Penelitian Utama Komisi Sains & Teknologi Chongqing (Hibah No. CSTC2017jcyjBX0025).

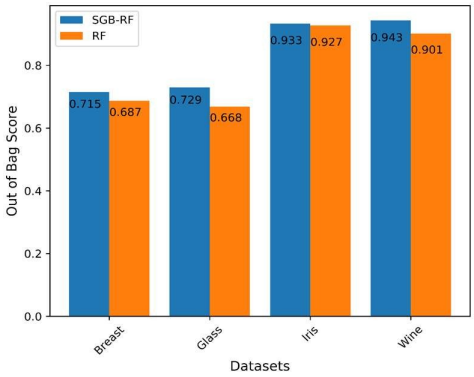

Gbr. 1. Matriks kebingungan dari pengklasifikasi RF




Gbr. 2. Matriks kebingungan dari pengklasifikasi SGB-RF

TABEL VI. DESKRIPSI DATASET UCI YANG DIGUNAKAN DALAM EKSPERIMEN

| Dataset | Jumlah catatan | Jumlah fitur | Jumlah kategori | Koefisien ketidakseimbangan |
|---|---|---|---|---|
| *Payudara* | 116 | 9 | 2 | 1.39 |
| *Kaca* | 214 | 9 | 6 | 8.99 |
| *Iris* | 150 | 4 | 3 | 1 |
| *Anggur* | 178 | 13 | 3 | 3.91 |




Gbr. 3. Skor di luar kantong RF dan SGB-RF pada dataset UCI

## REFERENSI


[1] Sarkheil, H. & Rahbari, S. Environ Earth Sci (2016) 75: 1319. https://doi.org/10.1007/s12665-016-6131-2

[2] Yun Bai, Yong Li, Xiaoxue Wang, Jingjing Xie, Chuan Li, Peramalan konsentrasi polutan udara menggunakan jaringan syaraf tiruan perambatan balik berdasarkan dekomposisi wavelet dengan kondisi meteorologi, Penelitian Pencemaran Atmosfer, Volume 7, Edisi 3, 2016, Halaman 557-566

[3] A. Shawabkeh, F. Al-Beqain, A. Redan dan M. Salem, "Model Pemantauan Polusi Udara Benzena menggunakan ANN dan SVM," Tren Teknologi Informasi HCT Kelima (ITT) 2018, Dubai, Uni Emirat Arab, 2018, hlm. 197-204

[4] Sagi, O, Rokach, L. Pembelajaran ansambel: Sebuah survei. WIREs Data Mining Knowl Discov. 2018; 8:e1249. https://doi.org/10.1002/widm.1249

[5] [Online] Tersedia: https://www.cnemc.cn/

[6] Breiman, L. (1996). Mengantongi prediktor. Pembelajaran Mesin, 24(2), 123-
140. doi:10.1023/A:1018054314350

[7] Breiman, L. (2001). Hutan acak. Machine Learning, 45(1), 5-32.

[8] Carmen Lai, Marcel J.T. Reinders, Lodewyk Wessels, Metode subruang acak untuk pemilihan fitur multivariat, Pattern Recognition Letters, Volume 27, Edisi 10, 2006, Halaman 1067-1076

[9] Echeverri, A.C., von Harling, B. & Serone, M. J. High Energ. Phys. (2016) 2016: 97. https://doi.org/10.1007/JHEP09(2016)097

[10] Salvador García, Zhong-Liang Zhang, Abdulrahman Altalhi, Saleh Alshomrani, Francisco Herrera, Pemilihan ansambel dinamis untuk dataset multi-kelas yang tidak seimbang, Ilmu Pengetahuan Informasi, Volume 445-446, 2018,
Halaman 22-37

[11] José A. Sáez, Bartosz Krawczyk, Michał Woźniak, Menganalisis pengambilan sampel berlebih dari berbagai kelas dan jenis contoh dalam dataset tidak seimbang multi-kelas, Pengenalan Pola, Volume 57, 2016, Halaman 164-178. https://doi.org/10.1016/j.patcog.2016.03.012

[12] Breiman L, Friedman J, Stone C. Klasifikasi dan Pohon Regresi. Wasworth, 1984

[13] Caelen, O. Ann Math Artif Intell (2017) 81: 429. https://doi.org/10.1007/s10472-017-9564-8

[14] [Online] Tersedia: http://archive.ics.uci.edu/ml/

[15] Breiman, L. (1996) Di luar Tas (Out-of-Bag) Estimasi.
https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf