

Kajian Algoritme Super Learner sebagai Metode Ensemble dalam Kasus Klasifikasi*

Fransdana Nadeak¹, Bagus Sartono^{1‡}, Anwar Fitrianto¹

^{1,1}Department of Statistics, IPB University, Indonesia

‡corresponding author: bagusco@apps.ipb.ac.id

Copyright © 2023 Fransdana Nadeak, Bagus Sartono, Anwar Fitrianto. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Classification is a statistical approach used when the response variable is categorical. The classification process generally consists of two phases: model training and model testing. The Super Learner is an ensemble method that integrates multiple candidate algorithms into a single predictive model by using V-fold cross-validation to determine the optimal weighted combination of base learners. Although numerous studies in the Department of Statistics at IPB University have applied various classification techniques and achieved satisfactory average accuracy, misclassification remains an issue that could potentially be reduced through model optimization. This study investigates whether the Super Learner ensemble can improve classification accuracy relative to single-model approaches previously applied. In addition, the study examines the characteristics of the resulting Super Learner models and evaluates the conditions under which performance gains are most pronounced.

Keywords: accuracy, classification, ensemble method, super Learner.

*Received: Jan 2023; Reviewed: Mar 2023; Published: Jun 2023

1. Pendahuluan

Klasifikasi merupakan salah satu metode penting dalam statistika dan pembelajaran mesin untuk memetakan objek ke dalam kelas yang telah ditentukan ketika peubah respon bersifat kategorik. Proses klasifikasi terdiri atas tahap pelatihan dan pengujian, di mana model dibangun melalui data berlabel dan kemudian dievaluasi untuk mengukur kemampuan generalisasinya (James et al., 2013). Berbagai algoritme seperti pohon klasifikasi, Support Vector Machine (SVM), regresi logistik, analisis diskriminan, dan KNN telah diterapkan dalam banyak penelitian. Dalam konteks pengenalan pola, pendekatan berbasis neural network seperti FLVQ juga menunjukkan kinerja yang sangat baik dalam tugas identifikasi distribusi probabilitas, sebagaimana ditunjukkan oleh (Warsono et al., 2016). Namun, metode tunggal sering kali menghadapi keterbatasan seperti tingginya ragam prediksi, ketidakstabilan model, dan sensitivitas terhadap ketidakseimbangan kelas (Linoff and Berry, 2011; Hosmer and Lemeshow, 2000). Metode ensemble dikembangkan untuk mengatasi kelemahan tersebut dengan menggabungkan beberapa model dasar sehingga meningkatkan stabilitas dan akurasi prediksi. Pendekatan seperti bagging dan Random Forest (Breiman, 2001; Sartono and Syafitri, 2010) menunjukkan bahwa kombinasi model dapat mengurangi varians dan memperbaiki performa klasifikasi. Salah satu pendekatan ensemble yang lebih modern adalah Super Learner, yaitu algoritme yang mengombinasikan sekumpulan kandidat model menggunakan V-fold cross-validation untuk memperoleh bobot optimal berbasis fungsi kerugian tertentu (Naimi and Balzer, 2017). Secara teoritis, Super Learner bersifat asymptotically optimal karena mampu meniru kinerja model terbaik dalam library kandidat (Hastie et al., 2008). Penelitian klasifikasi yang dilakukan di Departemen Statistika IPB selama beberapa tahun terakhir telah menggunakan berbagai metode seperti CART, C5.0, regresi logistik, KNN, analisis diskriminan, dan CHAID. Walaupun tingkat akurasinya cukup baik, masih terdapat ruang untuk menurunkan kesalahan klasifikasi, terutama pada dataset dengan banyak variabel, struktur data kompleks, atau proporsi kelas yang tidak seimbang. Kondisi ini membuka pertanyaan apakah metode ensemble Super Learner dapat memberikan peningkatan evaluasi model yang konsisten dibandingkan metode tunggal yang digunakan sebelumnya. Penelitian ini bertujuan untuk mengkaji efektivitas Super Learner dalam meningkatkan akurasi, sensitivitas, dan spesifisitas pada berbagai dataset penelitian terdahulu. Selain mengevaluasi performanya secara komprehensif, studi ini juga menganalisis bagaimana karakteristik data—seperti jumlah variabel kategorik dan numerik, ukuran sampel, serta perbandingan kelas—mempengaruhi kinerja Super Learner. Penelitian ini juga mengidentifikasi kombinasi base learner terbaik untuk setiap dataset menggunakan pendekatan meta-learning berbasis regresi logistik, sehingga memberikan gambaran mengenai situasi empiris yang paling mendukung keunggulan metode ensemble ini.

2. Metode Penelitian

2.1. Data

Gugus data yang digunakan adalah data penelitian-penelitian yang sudah dilakukan sebelumnya oleh mahasiswa Statistika IPB. Ada 16 penelitian yang dibandingkan

hasil klasifikasinya. Tabel 1 menunjukkan daftar penelitian yang dibandingkan dalam penelitian ini beserta tahun penelitian dan metode klasifikasi yang digunakan oleh peneliti sebelumnya. Semua data diusahakan untuk semirip mungkin dengan data dalam penelitian sebelumnya.

Tabel 1: Daftar penelitian yang dibandingkan

Nama data	Nama Peneliti	Tahun penelitian
A1	Resty Indah Sari	2011
A2	Rakhmawati	2011
A3	I Dewa G Amory	2012
A4	Rindy Pertiwi (2)	2013
A5	Rossi A Barro	2013
A6	Meita A Rubiati	2014
A7	Alfin Khairi	2014
A8	Gitania N Rahisti	2015
A9	Dimas Adiangga	2015
A10	Nur Fitriani	2015
A11	Nita Nurgenita	2015
A12	Adi Nugraha	2015
A13	Shafa R Surbakti	2015
A14	Devi Adrian	2018
A15	Dairul Fuhron	2018
A16	Dairul Fuhron(2)	2018

Tabel 2: Daftar akurasi, sensitivitas, dan spesifisitas

Nama data	Akurasi	Sensitivitas	Spesifisitas
A1	72.20	81.43	59.33
A2	74.40	43.33	72.67
A3	97.80	97.50	100.00
A4	74.60	92.70	34.20
A5	98.30	98.98	33.33
A6	70.60	76.90	62.60
A7	75.00	30.00	92.30
A8	96.67	99.78	21.05
A9	87.80	95.00	64.40
A10	68.70	12.60	100.00
A11	70.80	83.38	50.40
A12	70.20	38.46	86.67
A13	92.76	71.27	97.10
A14	88.42	3.06	98.94
A15	66.79	42.32	84.87
A16	72.73	100.00	1.77

2.2. Metode Penelitian

Tahapan analisis data dalam penelitian ini adalah sebagai berikut :

1. Memasukkan data ke dalam *software Python*
2. memilih peubah yang digunakan sesuai peubah yang terdapat di dalam model penelitian sebelumnya
3. mengubah peubah kategorik menjadi peubah boneka.
4. membagi gugus data menjadi data latih dan data uji dengan proporsi 70% dan 30%
5. menentukan *Base Learner* dan *Meta Learner* yang akan digunakan dalam algoritma *Super Learner*. *Meta Learner* yang digunakan adalah Regresi Logistik
6. melakukan pengklasifikasian menggunakan metode *Super Learner*
7. melakukan prediksi kelas/kategori respon terhadap data uji
8. hasil prediksi dari *Super Learner* yang berupa peluang ditransformasi menjadi 2 kelas dengan nilai *cut off* sebesar 0.5
9. menghitung nilai akurasi, sensitivitas, dan spesifisitas dari metode *Super Learner*
10. menghitung rata-rata Geometrik dari setiap data baik model sebelumnya maupun model *Super Learner*.
11. Mengulangi tahap 1 sampai tahap 9 untuk data-data penelitian yang dibandingkan sebanyak 20 kali, lalu
12. membandingkan hasil klasifikasi *Super Learner* dengan metode yang digunakan pada penelitian sebelumnya berdasarkan nilai akurasi klasifikasi, sensitivitas, spesifisitas, dan rata-rata geometrik
13. membuat grafik radar untuk perbandingan model *Super Learner* dan model sebelumnya
14. melakukan pemodelan regresi logistik untuk melihat faktor apa yang berpengaruh terhadap kebaikan model *Super Learner*
15. melakukan analisis korelasi dari ukuran evaluasi model terhadap karakteristik data untuk mengetahui pola data dalam pemodelan *Super Learner*.

3. Hasil dan Pembahasan

3.1. Karakteristik Data

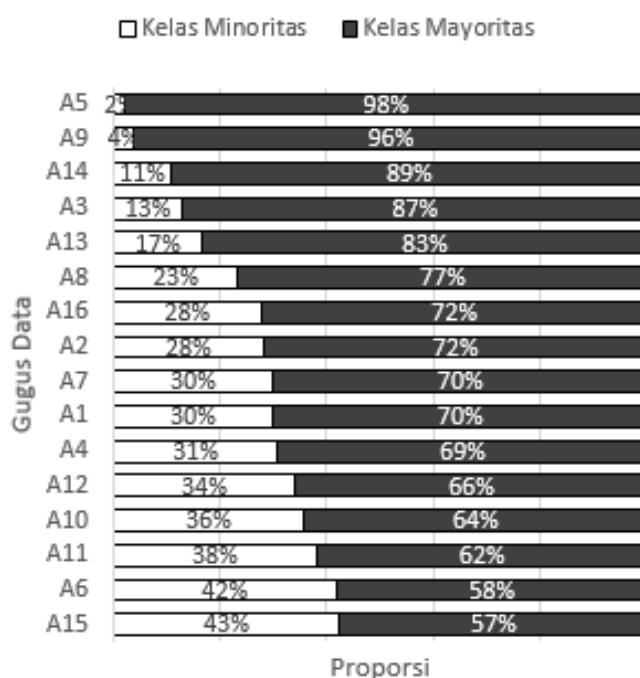
Data yang digunakan dalam penelitian ini adalah gugus data hasil penelitian sebelumnya sebanyak 16 gugus data dari tahun 2011 sampai 2018. Data tersebut merupakan data dengan peubah respon biner dengan berbagai proporsi kelas. Berikut merupakan karakteristik dari ke-16 data tersebut yang disajikan dalam bentuk tabel Peubah penjelas dari gugus data tersebut berupa variabel kategorik dan variabel numerik. Dari tabel 4 terlihat bahwa banyaknya variabel tiap data berbeda – beda. Data yang memiliki variabel kategorik terbanyak adalah data A5 dengan variabel kategorik sebanyak 294 variabel, sedangkan data yang memiliki variabel numerik terbanyak adalah data A3 yang memiliki variabel numerik sebanyak 18 variabel. Dari tabel tersebut juga dapat dilihat bahwa rentang banyaknya observasi dari keseluruhan data ada pada rentang 46 sampai 78609 observasi.

Tabel 3: Karakteristik data

Nama data	Banyak Observasi	Kategorik	Numerik
A1	1000	13	7
A2	1000	15	3
A3	46	0	18
A4	7024	8	1
A5	1002	294	0
A6	1651	15	1
A7	117	7	4
A9	2042	10	2
A8	311	5	3
A10	243	10	2
A11	3242	6	1
A12	114	8	3
A13	78609	14	1
A14	4951	7	2
A15	3371	3	2
A16	3371	3	2

3.2. Keseimbangan Data

Imbalance adalah suatu keadaan di mana salah satu kelas dari peubah responnya memiliki jumlah yang jauh lebih banyak daripada kelas yang lain sehingga proporsi kelasnya menjadi tidak seimbang. Keadaan *Imbalance* ini akan sangat mempengaruhi hasil prediksi dari pemodelan yang dilakukan.



Gambar 1: Persentase komposisi kelas peubah respon

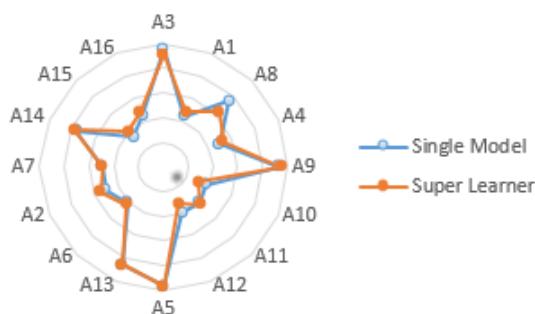
Gambar 1 merupakan gambaran dari persentase komposisi kelas peubah respon

dari setiap data yang sudah diurutkan dari yang perbandingan persentase komposisi kelas yang terbesar sampai yang terkecil.

Data A5 merupakan data yang paling tidak seimbang dikarenakan perbandingan antara kelas minoritas dan mayoritas yang sangat jauh. Sedangkan data yang perbandingan komposisi kelas yang paling kecil terdapat pada data A15 dengan proporsi sebesar 43% untuk kelas minoritas dan 57% untuk kelas mayoritas

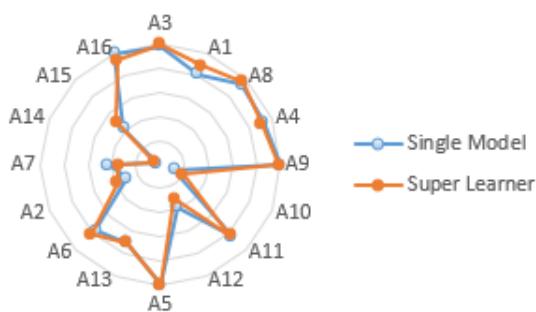
3.3. Hasil Evaluasi Model Super Learner

Perbandingan nilai evaluasi tiap data dari kedua model, baik dari *single model* maupun *Super Learner* akan digambarkan dengan grafik radar yang terlihat pada grafik di bawah ini.



Gambar 2: Perbandingan akurasi Single Model dan *Super Learner*

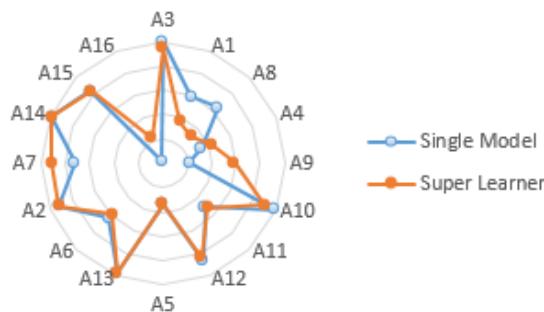
Secara keseluruhan terlihat bahwa mayoritas data memiliki akurasi yang lebih tinggi saat menggunakan *Super Learner*, namun ada beberapa data juga yang memiliki akurasi lebih rendah walapun nilainya tak terlalu jauh yaitu gugus data A3, A8, A10 dan A12. Data A8 juga memiliki perbedaan akurasi yang paling tinggi antar model ketimbang ketiga data lain.



Gambar 3: Perbandingan sensitivitas Single Model dan *Super Learner*

Dari grafik radar pada gambar 3, terlihat bahwa mayoritas gugus data memiliki sensitivitas yang lebih tinggi saat menggunakan *Super Learner*, namun ada beberapa data juga yang memiliki sensitivitas lebih rendah. Dari grafik radar di atas, terlihat bahwa data A7, A12, A13, dan A16 merupakan data dengan sensitivitas yang lebih rendah ketika menggunakan *Super Learner*. Terlihat bahwa gugus data A14 dan A10

merupakan data yang memiliki sensitivitas sangat rendah bila dibandingkan dengan data lain baik menggunakan *Super Learner* maupun *single model*.

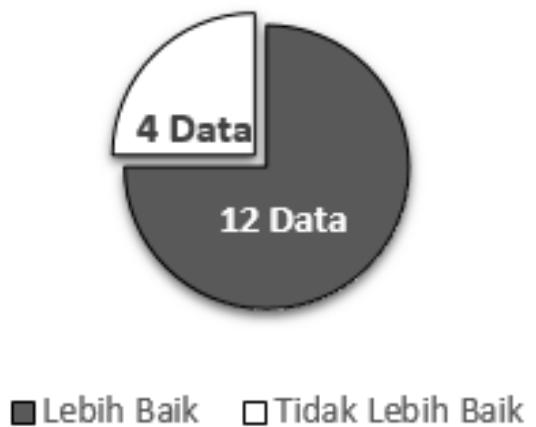


Gambar 4: Perbandingan spesifisitas *Single Model* dan *Super Learner*

Untuk perbandingan spesifisitas antar model, dapat diketahui dari gambar 4 bahwa data A1, A3, A8 dan A10 merupakan gugus data dengan sensitivitas yang lebih rendah ketika menggunakan *Super Learner*. Nilai spesifisitas dari keempat data tersebut pun terlihat memiliki perbedaan yang jelas. Terlihat juga bahwa data A1 dan A8 merupakan gugus data yang memiliki perbedaan spesifisitas yang jauh lebih tinggi bila dibandingkan model *Super Learner*-nya.

3.4. Perbandingan rata-rata Geometrik Antar Model

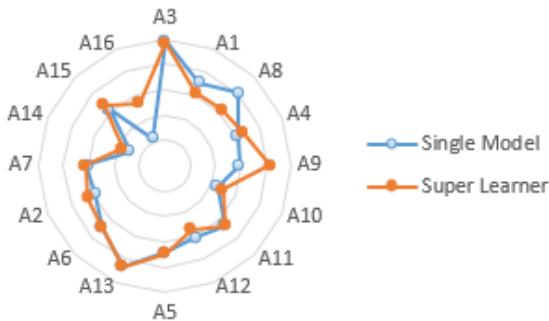
Dari 16 data yang dimiliki, dihitung rata - rata geometrik untuk model *Super Learner* dan *single model* sehingga didapatkan 16 rata-rata geometrik untuk masing- masing model.



Gambar 5: Diagram kue perbandingan rata-rata geometrik

Berdasarkan Gambar 5, dari 16 data, ada 12 data yang rata-rata geometrik *Super Learner* nya lebih tinggi daripada *single model*, sedangkan 4 data sisanya memiliki rata-rata geometrik *Super Learner* yang lebih rendah dari *single model*.

Berdasarkan gambar 6, terlihat ada 3 data yang rata-rata geometrik dari model *Super Learner*nya jauh lebih rendah dari *single model*, yaitu data A2, A3, dan A8. Sedangkan mayoritas data lainnya terlihat jauh lebih baik menggunakan model *Super Learner* dibandingkan *single model*. Oleh karena itu dapat dikatakan bahwa model *Super*



Gambar 6: Perbandingan rata-rata geometrik *single model* dan *super learner*

Learner lebih bagus untuk digunakan karena mayoritas data memiliki evaluasi yang lebih baik dibandingkan model sebelumnya.

3.5. Regresi Logistik

Salah satu cara untuk melihat faktor mana yang berpengaruh terhadap kinerja *Super Learner* adalah dengan memodelkannya. Model yang dipakai adalah regresi logistik. Peubah Y yang digunakan adalah keputusan lebih baik atau tidak yang di dapat dari perbandingan rata-rata geometrik kedua model.

Tabel 4: Tabel Uji G Regresi Logistik

Faktor	Estimate	Std.Error	Z Value	P Value
(Intercept)	-17.716	5.46	-3.24	0.00
Banyak variabel	0.695	0.23	3.04	0.00
Kelas Mayoritas	0.071	0.08	0.87	0.38
Perbandingan Kelas	5.292	1.91	2.77	0.01
Banyak Observasi	0.023	0.01	3.12	0.00
Banyak Variabel Numerik	-2.032	0.60	-3.39	0.00
Ukuran data	-0.002	0.00	-3.14	0.00

Berdasarkan tabel 6, terlihat bahwa semua peubah memiliki pengaruh yang signifikan terhadap peubah y kecuali variabel mayoritas. Variabel mayoritas merupakan besarnya proporsi untuk kelas mayoritas dari peubah responnya. Semakin besar proporsi mayoritas, maka semakin tidak seimbang data tersebut.

3.6. Feature Importance

Untuk melihat variabel mana yang paling berpengaruh atau yang paling penting, digunakan *feature importance*.

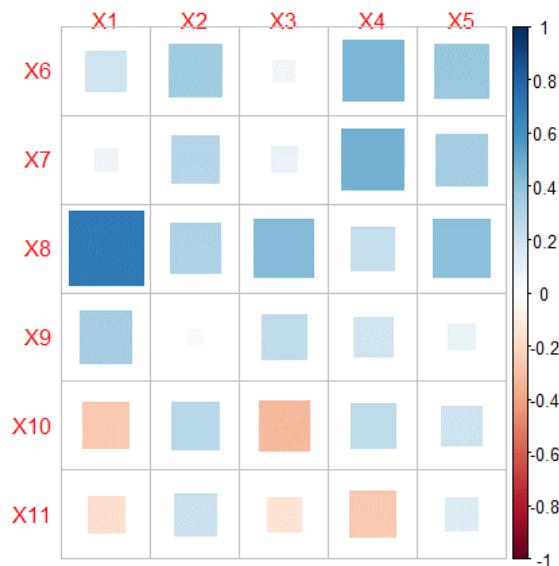


Gambar 7: Grafik korelasi hubungan antara ukuran evaluasi dengan karakteristik data

Berdasarkan gambar 7, terlihat bahwa banyaknya variabel, banyaknya observasi, dan ukuran data merupakan 3 faktor terpenting yang mempengaruhi kebaikan kinerja model *Super Learner*. Faktor yang paling penting diantara ketiga itu adalah banyaknya variabel, diikuti dengan banyak observasi, sedangkan variabel mayoritas dan perbandingan kelas tidak terlalu berpengaruh terhadap kebaikan model *Super Learner* terhadap *single model*.

3.7. Kecenderungan Hubungan Antara Ukuran Evaluasi dengan Karakteristik Data

Korelasi digunakan untuk melihat apakah ada hubungan yang linear antara variabel-variabel yang akan dilihat. Hubungan yang akan dilihat adalah Ukuran evaluasi terhadap karakteristik data.



Gambar 8: Grafik korelasi hubungan antara ukuran evaluasi dengan karakteristik data

Keterangan peubah pada grafik korelasi :

X1 = Perbandingan kelas

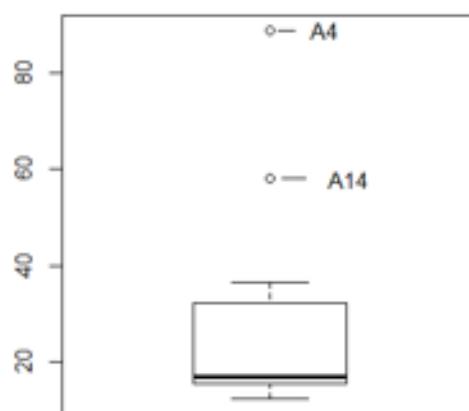
X2 = Banyak observasi

- X3 = Banyaknya variabel kategorik
- X4 = Banyaknya variabel numerik
- X5 = Ukuran data
- X6 = rata-rata geometrik *Super Learner*
- X7 = rata-rata geometrik *Single Model*
- X8 = Akurasi
- X9 = Sensitivitas
- X10 = Spesifisitas
- X11 = Waktu komputasi

Berdasarkan grafik korelasi, dapat diketahui bahwa perbandingan kelas dan banyak observasi memiliki korelasi yang positif terhadap rata-rata geometrik baik *Super Learner* maupun *single model*. Sehingga, dapat dikatakan bahwa semakin besar perbandingan kelas dan banyak observasi, maka akan semakin meningkatkan rata-rata geometrik kedua model. Namun, terlihat jelas bahwa besar korelasi perbandingan kelas terhadap rata-rata geometrik *Super Learner* lebih besar daripada *single model*. Sehingga dapat dikatakan bahwa model *Super Learner* menghasilkan evaluasi yang lebih baik saat terjadi masalah ketidakseimbangan data ketimbang *single model*.

3.8. Waktu Komputasi Pemodelan *Super Learner*

Waktu komputasi yang dihitung adalah seberapa lama model *Super Learner* yang telah kita buat melakukan pemodelan terhadap data latih.



Gambar 9: Boxplot Waktu komputasi

Rata-rata waktu komputasi keseluruhan adalah selama 27.39 detik. Sedangkan, persebaran waktu komputasinya kebanyakan berkisar di antara 12 detik sampai 31 detik. Dari gambar 9, dapat dilihat bahwa ada 2 data yang merupakan penculan atas untuk lama waktu komputasi. Kedua data tersebut adalah data A4 dan A14 yang memiliki lama waktu komputasi masing - masing selama 88.86 detik dan 58.05 detik.

4. Simpulan

Dari hasil dan pembahasan yang telah dilakukan dalam pemodelan Super Learner setelah dibandingkan dengan pemodelan oleh penelitian sebelumnya, didapat bahwa sebanyak 12 dari 16 data memiliki rata-rata geometrik yang lebih baik dibandingkan model sebelumnya. Empat data yang memiliki rata-rata geometrik lebih rendah memiliki banyak observasi yang ternyata bisa dibilang lebih sedikit. Untuk rata-rata ukuran akurasi, sensitivitas, dan spesifisitas dari keseluruhan gugus data mengalami kenaikan walaupun tidak terlalu besar saat menggunakan model *Super Learner*.

Faktor banyaknya variabel, ukuran data, perbandingan kelas, dan banyaknya observasi merupakan faktor yang memiliki pengaruh signifikan terhadap kebaikan model *Super Learner*. Dari faktor-faktor tersebut, faktor banyaknya variabel merupakan faktor yang paling penting diikuti dengan faktor banyaknya observasi, dan ukuran data dalam mempengaruhi kebaikan model *Super Learner*. Faktor yang perbandingan kelas tidak terlalu penting dalam mempengaruhi kebaikan model *Super Learner*.

Pustaka

- Breiman, L. (2001). Random forest. *Machine Learning*, 45, 5–32.
- Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher, Burlington (UK).
- Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2008). *The Elements of Statistical Learning: Data-Mining, Inference and Prediction*. Springer-Verlag, New York.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression Second Edition*. John Wiley and Sons, New Jersey (US).
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York (US).
- Linoff, G. S., and Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. John Wiley and Sons, New York (US).
- Naimi, A. I., and Balzer, L. (2017). An Introduction to Super Learning. *Stacked Generalization*, 35, 3–9.
- Sartono, B., and Syafitri, U. D. (2010). Metode Pohon Gabungan: Solusi Pilihan untuk Mengatasi Kelemahan Pohon Regresi dan Klasifikasi Tunggal. *Forum Statistika dan Komputasi*, 15(1), 1–7.
- Warsono, Dito, G.A., Kurniasari, D., and Usman, M. (2016). Neural Network Fuzzy Learning Vector Quantization (FLVQ) to Identify Probability Distributions. *IJCSNS International Journal of Computer Science and Network Security*, 16(10), 16–25.