Xplore (2023), Vol. 12, No. 2, 197–205
DOI: https://doi.org/10.29244/xplore.v12i2.335

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Penerapan Teknik Super Learner dalam Pemodelan Faktor yang Memengaruhi Rekomendasi Operator Seluler*

Aulia Fitriyani[1], Bagus Sartono[1‡],Septian Rahardiantoro[1]

[1,1]Department of Statistics, IPB University, Indonesia
[‡]corresponding author: bagusco@apps.ipb.ac.id

Abstract

Telecommunication refers to the exchange of information over long distances. Indonesia is one of the countries with the highest number of mobile network operators worldwide. This situation motivated Mobile Operator A to conduct a survey investigating store attendants' recommendations to customers regarding the use of Operator A's services. Classification methods can be applied to identify which operator a store attendant is likely to recommend based on several influencing factors. In this study, the super learner method is employed to integrate multiple base learners into a single optimized predictive model. The base learners used include random forest, bagging, and logistic regression. The resulting super learner model achieves an accuracy of 88.11% and an AUC of 0.9083. The most influential factor driving store attendants' recommendations is whether Operator A is the best-selling internet provider in the respective store. Beyond individual effects, several interactions between pairs of explanatory variables are also found to play a significant role.

Keywords: cross-validation, ensemble learning, classification, stacking, super learner.

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

1. Pendahuluan

Telekomunikasi merupakan proses pertukaran informasi pada jarak jauh dan menjadi kebutuhan yang semakin penting seiring perkembangan teknologi. Kemajuan tersebut mendorong meningkatnya kebutuhan layanan telekomunikasi di Indonesia, yang merupakan salah satu negara dengan jumlah operator seluler terbanyak di dunia. Menurut Syamni dan Martunis (2013), Indonesia memiliki sepuluh perusahaan operator seluler. Salah satu di antaranya adalah operator seluler A, sebuah perusahaan besar yang tetap menghadapi tantangan kompetitif, terutama terkait rendahnya tingkat rekomendasi pelanggan dibandingkan operator lain.

Untuk meningkatkan daya saing, perusahaan operator seluler A melakukan survei yang berfokus pada tiga pilar bisnis, yaitu availability, visibility, dan advocacy. Pilar advocacy menjadi perhatian khusus karena berkaitan dengan rekomendasi atau saran yang diberikan pelayan toko kepada konsumen. Oleh karena itu, penelitian ini diarahkan untuk memodelkan faktor-faktor yang memengaruhi rekomendasi pelayan toko terhadap penggunaan operator seluler A oleh pelanggan.

Proses klasifikasi digunakan untuk mengelompokkan kategori rekomendasi berdasarkan beberapa peubah penjelas (Warsono et al., 2016). Perkembangan metode klasifikasi telah menghasilkan berbagai pendekatan modern, salah satunya super learner, yaitu metode ensemble yang mengombinasikan sejumlah algoritme klasifikasi untuk memperoleh model dengan kinerja optimal (Polley dan Laan, 2010). Model super learner menentukan bobot masing-masing base learner berdasarkan kesalahan prediksi yang muncul pada tahap cross validation, sehingga algoritme dengan error lebih kecil mendapatkan bobot lebih besar. Pada penelitian ini digunakan tiga algoritme klasifikasi sebagai base learner, yaitu random forest (RF), bagging, dan regresi logistik, yang sebelumnya telah menunjukkan performa kuat pada berbagai studi.

Selain melakukan klasifikasi, penelitian ini juga mengidentifikasi peubah yang paling berpengaruh terhadap peubah respon menggunakan feature importance melalui pendekatan analisis degradasi akurasi. Pengaruh peubah penjelas dievaluasi lebih lanjut menggunakan partial dependence plot (PDP) yang memberikan gambaran visual mengenai hubungan antara peubah penjelas dan respon (Greenwell, 2017). Interaksi antarpeubah khususnya pada kelompok brand engagement juga dianalisis untuk memahami kombinasi faktor yang dapat memengaruhi keputusan rekomendasi.

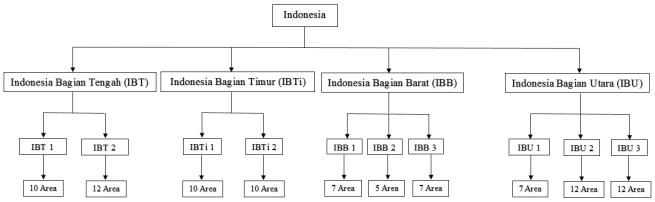
Berdasarkan konteks tersebut, penelitian ini bertujuan untuk membangun model klasifikasi rekomendasi operator seluler A menggunakan metode super learner dengan tiga algoritme dasar, yaitu random forest, bagging, dan regresi logistik. Model diharapkan mampu mengidentifikasi faktor-faktor yang paling berpengaruh terhadap rekomendasi pelayan toko serta memberikan pemahaman yang lebih mendalam mengenai pola keterkaitan antarpeubah dalam memengaruhi keputusan rekomendasi operator seluler.

2. Methodology

2.1. Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang didapatkan dari hasil survei salah satu *marketing research* mengenai operator seluler di Indonesia.

Survei dilakukan dengan menggunakan teknik pengambilan sampel *multi stage stratified random sampling* dengan 3 strata yang dapat dilihat pada Gambar 1.



Gambar 1: Strata-strata yang digunakan pada survei

Masing-masing area kemudian dipilih toko sebanyak 180-350 secara acak. Surveyor mengambil data dengan cara menjadi *mystery shopper* (pembeli misteri) serta melakukan *interview* dan observasi. Data terdiri dari 16 peubah penjelas (X), 1 peubah respon (Y) dan 12365 amatan. Peubah X terdiri dari 1 peubah mengenai *advocacy*, 7 peubah mengenai media promosi dan 8 peubah mengenai *brand engagement*. Peubah Y terdiri dari 2 kategori yaitu pelayan toko merekomendasikan operator seluler A dan pelayan toko tidak merekomendasikan operator seluler A. Sedangkan semua peubah X juga terdiri dari 2 kategori yaitu iya dan tidak.

2.2. Metode Penelitian

Tahapan analisis data dalam penelitian ini adalah sebagai berikut :

1. Mencari kejanggalan pada data.
   (a) Mencari data duplikat, jika ada kemudian duplikat dari data tersebut dihilangkan.
   (b) Mencari data hilang, jika ada kemudian data hilang tersebut diduga dengan modus.
2. Memasukkan data ke dalam *software* R 3.4.1.
3. Melakukan uji Chi Square
   Uji chi square merupakan salah satu pengujian untuk mengetahui hubungan atau kebebasan antar peubah (Tanty et al., 2013). Uji chi square dilakukan satu per satu pada masing-masing peubah penjelas terhadap peubah respon. Pada penelitian ini, uji chi square dilakukan pada *software* R yang menghasilkan *p-value*. *P-value* dapat digunakan untuk memutuskan apakah tolak atau tak tolak H0. Jika *p-value* lebih kecil dari α yang digunakan maka tolak H0 dan sebaliknya. Penelitian ini menggunakan α sebesar 0.05. Rumus yang digunakan adalah
   $$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
   Keterangan:
   O_{ij} = Frekuensi data baris ke-i kolom ke-j.
   E_{ij} = Nilai harapan dari data baris ke-i kolom ke-j.
   $$E_{ij} = \frac{O_{i.} * O_{.j}}{O_{..}}$$
4. Melakukan eksplorasi data.

(a) Membuat diagram lingkaran peubah respon untuk melihat proporsi dari kedua kategori.

(b) Membuat diagram batang X1 sampai X8 kemudian dijadikan dalam 1 grafik. Peubah X1 sampai X8 merupakan peubah-peubah mengenai media promosi yang digunakan oleh operator seluler A sehingga dapat digunakan untuk melihat perbandingan antara kedua kategori dari X1 sampai X8 dan melihat secara keseluruhan media promosi yang paling banyak digunakan oleh operator seluler A.

(c) Membuat diagram batang X9 sampai X16 kemudian dijadikan dalam 1 grafik. Peubah X9 sampai X16 merupakan peubah mengenai *brand engagement* sehingga dapat digunakan untuk melihat perbandingan antara kedua kategori dari X9 sampai X16 dan melihat secara keseluruhan mengenai *brand engagement* dari operator seluler A.

5. Membagi data menjadi data training dan data testing dengan perbandingan 80:20 (Level 0).

6. Menetapkan 9 kombinasi *hyperparameter*
Hyperparameter merupakan suatu peubah yang mempengaruhi *output* model. Perubahan yang dilakukan pada *hyperparameter* bisa menyebabkan model menjadi lebih baik ataupun lebih buruk. Perubahan yang diberikan kepada *hyperparameter* RF adalah diperkecil 1/2 kali dan diperbesar 2 kali lipat dari *default* sedangkan untuk *hyperparameter bagging* diperkecil 1/5 kali dan diperbesar 5 kali lipat sehingga didapatkan kombinasi seperti pada Tabel I.

Tabel 1: Kombinasi hyperparameter super learner

Kombinasi	Banyak Pohon RF	Banyak Peubah Penjelas RF	Ulangan *Bootstrap Bagging*
1	250	2	5
2	250	2	125
3	250	8	5
4	250	8	125
5	500*	4*	25*
6	1000	2	5
7	1000	2	125
8	1000	8	5
9	1000	8	125

7. Melakukan teknik *super learner* pada data *training*.

(a) Menerapkan masing-masing algoritme klasifikasi yang digunakan dengan tahapan *cross validation*
Cross validation merupakan salah satu metode untuk mengevaluasi dan membandingkan kinerja suatu algoritme dengan cara mebagi data menjadi dua yaitu data *training* dan *testing* (Browne, 2000). *Fold* yang digunakan pada penelitian ini adalah 10 karena menurut Han et al. (2012), 10 *fold* merupakan penggunaan *fold* terbaik yang relatif mampu memberikan bias

dan ragam yang kecil. Algoritme klasifikasi yang digunakan pada penelitian ini adalah *bagging*, *random forest* dan regresi logistik. Pada tahapan ini dihasilkan $\hat{\Psi}$ dari masing-masing algoritme klasifikasi.

(b) Mengubah $\hat{\Psi}$ menjadi X *super learner* (X_{SL}) (Level 1).
$\hat{\Psi}_{Bagging} \rightarrow X_{1SL}$, $\hat{\Psi}_{RF} \rightarrow X_{2SL}$, $\hat{\Psi}_{Reglog} \rightarrow X_{3SL}$

(c) Memprediksi koefisien dengan menggunakan regresi logistik.
Peubah X: X *Super Learner* (SL).
Peubah Y: Y data aktual.

(d) Didapatkan model.

(e) Agar lebih mudah dipahami ubah XSL menjadi bentuk awalnya sehingga terbentuk model *super learner* seperti dibawah
$\hat{\Psi}_{SL} = \alpha_{RF}\hat{\Psi}_{RF} + \alpha_{bag}\hat{\Psi}_{bag} + \alpha_{reglog}\hat{\Psi}_{reglog}$

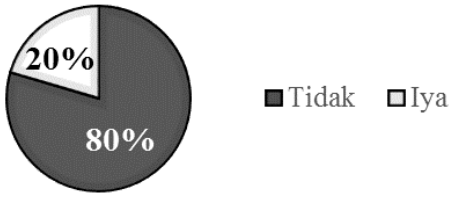
8. Melakukan klasifikasi pada data *testing* dengan model yang didapatkan pada nomor 7 bagian d.
9. Mengevaluasi hasil data *testing*.
10. Lakukan tahap 7 sampai 9 sebanyak 20 kali untuk masing-masing kombinasi
11. Buat *boxplot* dari 20 akurasi dan 20 nilai AUC yang didapatkan dari masing-masing kombinasi sehingga didapatkan 9 *boxplot* akurasi dan AUC.
12. Memilih 1 kombinasi terbaik
    Cara untuk memilih kombinasi terbaik adalah menghitung nilai rata-rata dan simpangan baku dari masing-masing kombinasi yang diulang sebanyak 20 kali. Kombinasi yang terpilih adalah kombinasi yang memiliki nilai rata-rata terbesar dan simpangan baku terkecil. Jika tidak ada yang memenuhi, pilih kombinasi yang stabil yaitu nilai rata-rata yang cukup besar dan simpangan baku yang cukup kecil.
13. Menerapkan *super learner* dengan kombinasi *hyperparameter* terpilih pada data lengkap.
14. Mengevaluasi hasil yang didapatkan dari tahap 13
15. Menginterpretasi peubah penjelas
    Peubah penjelas terpenting, pengaruh peubah penjelas terhadap peubah respon dan interaksi antar 2 peubah penjelas.

3. Hasil dan Pembahasan

Sebelum melakukan analisis, dilakukan pencarian terhadap kejanggalan data untuk melihat apakah terdapat data hilang dan data duplikat pada data yang akan digunakan. Setelah dicari ternyata tidak ada data hilang dan duplikat pada data yang akan digunakan untuk penelitian, sehingga dapat dilanjutkan untuk tahap selanjutnya. Tahap selanjutnya adalah melakukan identifikasi untuk mengetahui peubah penjelas mana yang berpengaruh terhadap peubah respon menggunakan uji chi square. Hasil uji chi square yang dilakukan pada *software* R 3.4.1 menunjukkan bahwa semua peubah penjelas yang digunakan berpengaruh terhadap peubah responnya.

3.1. Karakteristik Rekomendasi Serta Faktor yang Mempengaruhi Rekomendasi

Menurut Syamni dan Martunis (2013) Indonesia memiliki 10 perusahaan operator seluler sehingga jika ada pertanyaan mengenai rekomendasi penggunaan operator seluler tentunya akan memberikan banyak jawaban. Survey ini dilakukan oleh sebuah perusahaan operator seluler (operator seluler A). Peubah respon pada data yang digunakan adalah rekomendasi untuk menggunakan operator seluler A oleh pelayan toko kepada pelanggan dengan persentase yang didapatkan disajikan pada Gambar 2.



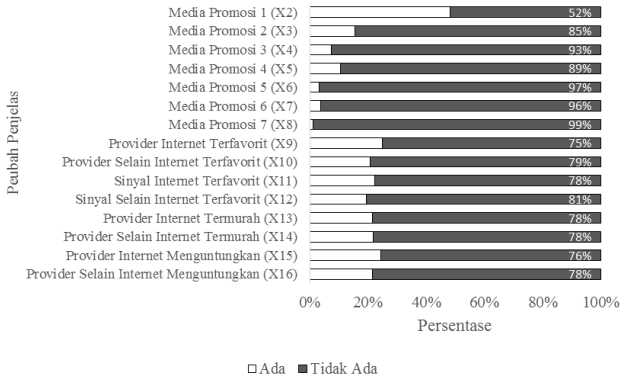
Gambar 2: Persentase rekomendasi penggunaan operator seluler A oleh pelayan toko untuk konsumen

Gambar 2 menunjukkan bahwa lebih banyak pelayan toko yang tidak merekomendasikan operator seluler A kepada konsumen dengan persentase sebesar 80% (9837 toko). Penelitian ini menggunakan 16 peubah penjelas (faktor yang mempengaruhi) Peubah pertama yang akan dibahas adalah peubah X1 yang merupakan peubah mengenai keberadaan kartu perdana operator seluler A pada toko yang disurvey.

Gambar 3: Persentase keberadaan kartu perdana operator seluler A (X1) pada toko yang disurvey

Berdasarkan Gambar 3 terlihat bahwa 64% (7961 toko) yang disurvey memiliki kartu perdana operator seluler A dan hanya 36% saja (4404 toko) yang tidak memiliki kartu perdana operator seluler A. Selanjutnya yang akan dibahas adalah peubah-peubah mengenai media promosi yang dilakukan oleh operator seluler A serta mengenai peubah *brand engagement* yang disajikan pada Gambar 4.

Berdasarkan Gambar 4 terlihat bahwa media promosi yang dilakukan operator seluler A pada toko lebih banyak dilakukan dengan menggunakan media promosi 1 yaitu ada sebanyak 48% atau ada sebanyak 6397 dari 12365 toko. Media promosi terbanyak yang dilakukan oleh operator seluler A selanjutnya adalah media promosi 2, media promosi 4 dan media promosi 3 dengan persentase berturut-turut yaitu sebesar 15%, 11% dan 7%. Selain itu terdapat 3 media promosi lain yang dilakukan



Gambar 4: Gambaran umum media promosi dan *brand engagement* operator seluler A

oleh operator seluler A namun, persentase dari keberadaan promosi tersebut tidak mencapai 5%. Ketiga media promosi tersebut adalah media promosi 6, media promosi 5, dan media promosi 7 dengan persentase masing-masing sebesar 4%, 3% dan 1%.

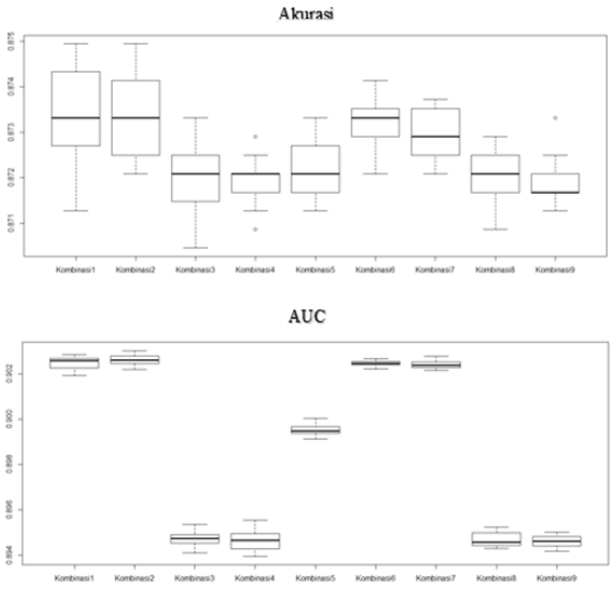
Menurut Gambar 4, persentase tertinggi dari *brand engagement* dimiliki oleh *provider* internet terfavorit dengan persentase sebesar 25% sedangkan untuk *provider* selain internet terfavorit, operator seluler A mendapatkan persentase sebesar 21%. Sama halnya mengenai kedua peubah yang telah disebutkan, keempat peubah ini juga tidak memiliki perbedaan persentase yang sangat jauh, Peubah-peubah tersebut adalah sinyal internet terbaik & sinyal selain internet terbaik (22% dan 19%) serta *provider* internet menguntungkan & *provider* selain internet menguntungkan (24% dan 22%). Terdapat 2 layanan lagi yang belum disebutkan, yaitu *provider* internet termurah serta *provider* selain internet termurah, operator seluler A memiliki persentase yang sama pada kedua peubah ini yaitu sebesar 22%.

3.2. *Super Learner*

Setelah menerapkan metode *super learner* dengan 9 kombinasi *hyperparameter*, masing-masing kombinasi diulang sebanyak 20 kali sehingga didapatkan pula 20 akurasi dan AUC. Nilai-nilai akurasi dan AUC yang didapatkan dari masing-masing kombinasi kemudian dibuat *boxplot* dan hasilnya disajikan pada Gambar 5.

Jika *boxplot* digunakan untuk memilih kombinasi yang akan diterapkan pada data lengkap, tentunya sulit untuk memilih kombinasi terbaik karena *boxplot* yang dihasilkan antar 2 kombinasi hampir sama. Hal ini dapat diatasi dengan melihat nilai rata-rata dan simpangan baku yang dihasilkan dari masing-masing kombinasi. Kombinasi ke-6 merupakan kombinasi yang cukup stabil karena mendapatkan nilai rata-rata terbesar ke-3 dengan simpangan baku terkecil kedua baik dari akurasi maupun AUC. Model yang didapatkan dari penerapan metode *super learner* menggunakan kombinasi ke-6 adalah $\hat{\Psi}_{SL} = 0.8021\hat{\Psi}_{RF} + 0.1980\hat{\Psi}_{Reglog} + 0.0000\hat{\Psi}_{Bagging}$. Nilai akurasi, sensitivitas serta spesifisitas yang didapatkan berturut-turut adalah sebesar 88.11%, 93.42% dan 67.44%. Model yang dihasilkan memperlihatkan bahwa untuk data operator seluler, *base learner* terbaik adalah RF kemudian diikuti regresi logistik karena nilai koefisien yang didapatkan oleh RF lebih besar. Koefisien yang dimiliki oleh *bagging* adalah 0 yang artinya kurang cocok untuk digunakan pada data ini sehingga tidak perlu untuk

dimasukkan ke dalam model.



Gambar 5: *Boxplot* akurasi dan AUC dari masing-masing kombinasi

Peubah penjelas terpenting pada data ini adalah peubah X9 (*provider* internet terfavorit) karena ketika X9 tidak diikutsertakan, nilai akurasi yang didapatkan berkurang cukup besar dibandingkan dengan peubah penjelas lainnya. Semakin penting suatu peubah maka semakin besar pula peubah penjelas tersebut dalam memberi pengaruh terhadap peubah responnya. Karena peubah terpenting pada data ini adalah X9 maka dapat diartikan jika operator seluler A menjadi *provider* internet terfavorit di toko tersebut akan semakin besar pula kecenderungan pelayan toko merekomendasikan operator seluler A kepada konsumen. Tidak hanya pengaruh dari masing-masing peubah penjelas yang mempengaruhi tetapi ada pula pengaruh dari perpaduan 2 peubah (interaksi) terutama dari peubah mengenai *brand engagement* yang mempengaruhi rekomendasi pelayan toko kepada pelanggan untuk menggunakan operator seluler A.

4. Simpulan

Penelitian ini menggunakan 3 *base learner* yaitu *random forest*, *bagging* dan regresi logistik dengan kombinasi *hyperparameter* banyak pohon pada *random forest*, banyak peubah penjelas pada *random forest* serta ulangan *bootsrap* pada *bagging* dengan nilai masing-masing sebesar 1000, 2 dan 5. Model yang dihasilkan adalah $\hat{\Psi}_{SL} = 0.8021\hat{\Psi}_{RF} + 0.1980\hat{\Psi}_{Reglog} + 0.0000\hat{\Psi}_{Bagging}$, yang berarti dari ketiga *base learner* yang digunakan hanya 2 yang sesuai digunakan pada data ini yaitu *random forest* dan regresi logistik. Pada data ini, metode *super learner* mampu memberikan nilai akurasi sebesar 88.11% dengan nilai sensitivitas dan spesifisitas sebesar 93.42% dan 67.44%. Peubah operator seluler A yang merupakan *provider* internet terlaris pada toko yang disurvei merupakan peubah terpenting pada data ini. Secara keseluruhan hampir semua peubah mengenai *brand engagement* memiliki interaksi dengan peubah lain baik dengan sesame peubah *brand engagement* ataupun dengan peubah media

promosi, yang berarti terdapat pengaruh dari perpaduan peubah *brand engagement* dengan peubah lain yang mempengaruhi rekomendasi pelayan toko.

Pustaka

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, 44(1), 108–132.

Greenwell, B. M. (2017). Pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), 421–436.

Han, J., Kamber, M., dan Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc, Waltham (US).

Polley, E. C., dan Laan, M. J. V. D. (2010). Super Learner in Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 266.

Syamni, G., dan Martunis (2013). Pengaruh OPM, ROE dan ROA terhadap perubahan laba pada perusahaan telekomunikasi di Bursa Efek Indonesia. *Jurnal Kebangsaan*, 2(4), 19.

Tanty, H., Bekti, R. D., dan Rahayu, A. (2013). Metode Nonparametrik untuk Analis Hubungan Perilaku dan Pengetahuan Masyarakat Tentang Kode Plastik. *Jurnal Mat Stat*, 13(2), 97–104.

Warsono, Dito, G.A., Kurniasari, D., dan 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.