Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 2, No.1, March 2019, pp. 52 – 60
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

A Comparison Framework of Classification Models based on Variable Extraction Model for Status Classify of Contraception Method in Fertile Age Couples in Indonesia

Laelatul Khikmah
Program Studi Statistika, Akademi Statistika (AIS) Muhammadiyah Semarang
Email: aisyah.salsabila17@gmail.com

Article Info

Article history:
Received Oct 14th, 2018
Revised Des 08th, 2018
Accepted Feb 11th, 2019

Keyword:
Contraception
Classification Models
Logistic Regression
Classifier Effectiveness

ABSTRACT

In terms of minimizing the risk of death in mothers the use of contraceptive methods really needs to be improved and the success of the use of contraceptive methods. This study aims to compare several popular classification models used to classify the status of the use of contraceptive methods in fertile age couples in Indonesia so that they can be used and the implementation of policies that are more impartial using the variable extraction integration method. The proposed model in this study is a comparative study of classification models include Logistic Regression (LR), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), C4.5, and CART. For the purpose of testing the model, Accuracy, AUC, F-measure, Sensitivity (SN), Specificity (SP), Positive Predictive Value (PPV), and Negative Predictive Value (NPV) are used to test frameworks comparative study of classification models. Based on the experimental results, RL shows superior and stable performance compared to other methods. It can be concluded, the RL method is the right choice method to classify the status of use of contraceptive methods in couples of childbearing ages in Indonesia.

Copyright © 2019 Puzzle Research Data Technology

Corresponding Author:
Laelatul Khikmah,
Program Studi Statistika
Akademi Statistika (AIS) Muhammadiyah Semarang
Jawa Tengah, Indonesia
Email: aisyah.salsabila17@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v2i1.7568

1. PENDAHULUAN

Program Keluarga Berencana (KB) merupakan program pemerintah yang bertujuan membangun keluarga kecil bahagia sejahtera. Program ini berupaya untuk menciptakan penduduk yang berkualitas yang akan mempercepat tercapainya pertumbuhan ekonomi dan tujuan pembangunan. Dasar penyelenggaraan pelayanan Keluarga Berencana (KB) adalah UU RI Nomor 36 Tahun 2009 tentang Kesehatan, pasal 78 tentang KB yang berbunyi: (1) Pelayanan kesehatan dalam keluarga berencana dimaksudkan untuk pengaturan kehamilan bagi pasangan kelompok umur usia subur untuk membentuk generasi penerus yang sehat dan cerdas; (2) Pemerintah bertanggung jawab dan menjamin ketersediaan tenaga, fasilitas pelayanan, alat dan obat dalam memberikan pelayanan KB yang aman, bermutu dan terjangkau oleh masyarakat; (3) Ketentuan mengenai pelayanan KB dilaksanakan sesuai dengan peraturan perundang-undangan.

Kebutuhan akan praktik metode kontrasepsi yang sehat dan penerapan manajemen peubah merupakan penentu keberhasilan dalam penggunaan metode kontrasepsi (dalam hal meminimalisasi resiko kematian pada ibu) perlu ditingkatkan.

Data mining merupakan metode yang banyak digunakan hampir di semua bidang bertujuan menemukan informasi yang bermanfaat dari sekumpulan data. *Data mining* adalah studi untuk mengumpulkan, membersihkan, mengolah, menganalisis, dan mendapatkan wawasan yang berguna dari data. Variasi pada data merupakan *domain* masalah untuk merepresentasi data yang ditemukan dalam aplikasi nyata [1]. Dalam penelitian ini *data mining* akan digunakan untuk menggambarkan berbagai aspek

pemrosesan data menggunakan model klasifikasi dalam mengetahui kelompok status penggunaan metode kontrasepsi pada pasangan usia subur di Indonesia.

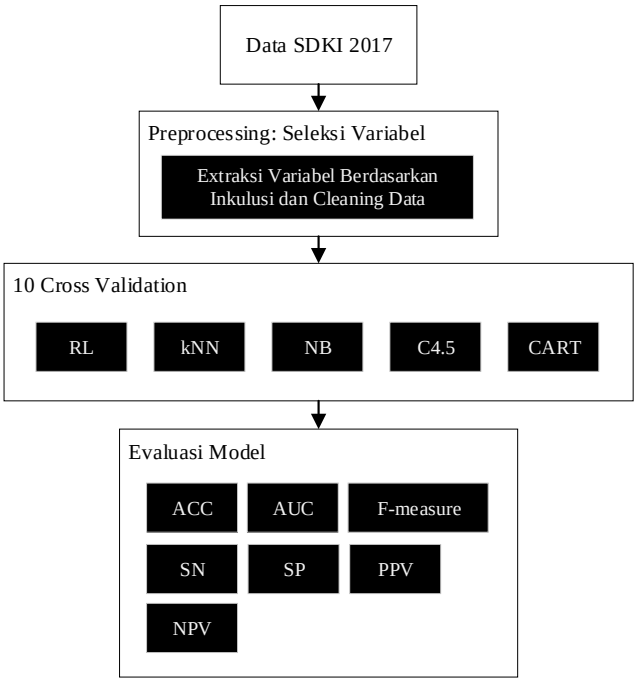
Model klasifikasi mempelajari struktur kumpulan data yang sudah dipartisi menjadi kelompok, yang disebut sebagai kategori atau kelas. Pembelajaran, dicapai dengan model. Input model klasifikasi adalah contoh kumpulan data yang telah dipartisi ke dalam kelas yang berbeda. Proses ini disebut *data training*, dan pengidentifikasi kelompok dari kelas-kelas ini disebut sebagai label kelas. Dalam banyak kasus, label kelas memiliki interpretasi semantik yang jelas dalam konteks aplikasi tertentu. Model yang dipelajari disebut sebagai model pelatihan. Poin data yang sebelumnya tidak terlihat yang perlu diklasifikasikan secara kolektif disebut sebagai *data testing*.

Beberapa penelitian pada kasus di atas telah banyak dilaporkan. Oleh [2], melaporkan hybrid metode Regresi Logistik dan CATPCA untuk menangani multikolinieritas pada metode kontrasepsi di Indonesia. [3] melaporkan Regresi Logistik untuk mengetahui faktor rendahnya keikutsertaan pengguna kontrasepsi jangka panjang pada pasangan usia subur. Selain itu oleh [4], juga telah mengusulkan Regresi logistik ganda untuk mengetahui faktor yang mempengaruhi pemilihan metode kontrasepsi.

Artikel ini disusun sebagai berikut. Di bagian 2, kerangka metode yang diusulkan dan kerangka teoritis dijelaskan. Pada bagian 3, hasil percobaan dan analisis disajikan. Akhirnya, karya kami dari artikel ini disimpulkan di bagian terakhir.

2. METODE PENELITIAN

Metode eksperimen digunakan dalam penelitian ini. Metode eksperimen yang kami usulkan adalah perbandingan kerangka model berdasarkan model ekstraksi variabel untuk klasifikasi status penggunaan metode kontrasepsi pada pasangan usia subur di Indonesia. Model ekstraksi variabel dan pembersihan data digunakan untuk memilih variabel yang relevan berdasarkan inklusi peneliti dan melakukan pembersihan data pada data yang missing pada data SDKI 2017 berdasarkan faktor yang berpengaruh terhadap penggunaan metode kontrasepsi. Model kerangka perbandingan yang diusulkan dievaluasi menggunakan data Survei Demografi dan Kependudukan Indonesia (SDKI) 2017.



Gambar 1. Blok diagram kerangka perbandingan model yang di usulkan

2.1. Pengumpulan Data

Dalam penelitian ini, data SDKI 2017 digunakan diperoleh dari *Demographic and Health Surveys* (DHS) dengan pilihan regional Indonesia, dapat diakses pada laman https://dhsprogram.com/data/available-datasets.cfm. Data asli pada SDKI 2017 memiliki jumlah data 49627 dan variabel 4959. Tujuan dari penelitian ini, untuk mengklasifikasi status penggunaan metode kontrasepsi pada pasangan usia subur di Indonesia sehingga pemilihan variabel hanya diambil berdasarkan kriteria inklusi oleh peneliti berdasarkan penelitian yang telah dilakukan dan dilaporkan oleh [2] dengan dataset yang sama. Banyaknya data missing pada masing-masing variabel juga menjadi perhatian khusus dalam pemilihan variabel bebas pada data

SDKI. Selain itu, data yang tersedia juga terdapat data bulanan dan data harian, sehingga peneliti hanya mengambil data secara keseluruhan dalam kurun waktu tahun diambilnya data tersebut. Deskripsi data variabel yang terpilih dari 4959 adalah 7 variabel terpilih di tampilkan pada Tabel 1.

Tabel 1. Data SDKI 2017

	Jumlah Anak	Kelompok Umur	Pekerjaan Istri	Pendidikan Istri	Pendiidkan Suami	Pengetahuan	Metode KB
x	≤2, >2	30 tahun, 30-39 tahun, Diatas 39 tahun	Tidak bekerja Bekerja	SD, SMP, SMA, Diploma, Universitas	SD, SMP, SMA, Diploma, Universitas	Tahu, Tidak tahu	Menggunakan, Tidak Menggunakan
y	2	3	2	5	5	2	2
Jumlah Data		**29883**					

2.2. Model Klasifikasi

Kerangka perbandingan model yang diusulkan bertujuan untuk membandingkan kinerja berbagai model klasifikasi untuk klasifikasi status penggunaan metode kontrasepsi pada pasangan usia subur di Indonesia. Untuk tujuan penelitian ini, lima pengklasifikasi telah dipilih, setiap model dikelompokkan ke dalam kategori pengklasifikasi termasuk statistik tradisional (LR, dan NB), tetangga terdekat (k-NN dan K *), dan decision tree (C4.5, dan CART). Seleksi ini bertujuan untuk mencapai keseimbangan antara model klasifikasi yang ditetapkan dan digunakan dalam klasifikasi status penggunaan metode kontrasepsi pada kasus yang telah disebutkan sebelumnya.

2.2.1. Regresi Logistik

Regresi logistik (LR) adalah teknik pemodelan statistik di mana probabilitas variabel terikat yang berskala kategorik terkait dengan variabel bebas adalah yang berskala numerik atau kategorik. Misalkan variabel terikat memiliki M kategori [5]. Satu nilai (kejadian gagal) dari variabel terikat ditunjuk sebagai referensi kategori. Probabilitas keanggotaan dalam kategori dibandingkan dengan probabilitas keanggotaan dalam referensi kategori [6]. Secara umum probabilitas regresi logistik ditunjukkan pada persamaan 1:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \quad (1)$$

Di mana, $\pi(x)$ adalah probabilitas kategori sukses pada variabel terikat, dan β_0, β_1 adalah parameter pendugaan bagi variabel bebas.

2.2.2. K-NN

K-NN digunakan untuk melakukan klasifikasi terhadap objek berdasarkan pembelajaran yang jaraknya paling dekat dengan objek [7]. Pengukuran jarak terdekat pada penelitian ini menggunakan metode Euclidean distance [8]. Formula penghitungan euclidean distance dapat dilihat pada persamaan 2. Di mana $d(x, y)$ adalah jarak antara data *x* ke data *y*, x_i adalah data testing ke-*i*, dan y_1 adalah data training ke-*i*.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

2.2.3. Naïve Bayes (NB)

NB adalah salah satu model klasifikasi berdasarkan teorema Bayesian pada statistika [9]. Model NB dapat digunakan untuk memprediksi probabilitas keanggotaaan suatu kelas [10]. Teorema Bayesian menghitung nilai posterior probability $P(H|X)$, menggunakan probabilitas $P(H)$, $P(X)$ dan $P(X|H)$ [11]. Penghitungan model NB untuk tipe data nominal menggunakan persamaan 3. Apabila dataset bertipe numerik, maka digunakan penghitungan distribusi Gaussian [12]. Penghitungan distribusi Gaussian dapat dilihat dari persamaan 4, di mana dihitung terlebih dahulu nilai rata-rata μ sesuai persamaan 5, dan standard deviasi σ sesuai persamaan 6.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (3)$$

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi.\sigma}} exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (4)$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (5)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad (6)$$

2.2.4. C4.5

Model C4.5 menggunakan penghitungan $entropy$, $informationGain$, $splitInfo$ dan $gainRatio$ untuk pemilihan variabel menjadi *node*. Formula penghitungan $entropy$ dan $informationGain$ dapat dilihat pada persamaan 7 dan 8. Sedangkan untuk penghitungan $splitInfo$ dan $gainRatio$ dapat dilihat pada persamaan 9 dan 10. Pada $splitInfo$, nilai *D* adalah ruang data sampel yang digunakan untuk training, nilai D_j adalah jumlah sampel pada variabel *j*.

$$entropy = \sum_{i=1}^{n} -P_i^* log_2 p_i \quad (7)$$

$$Gain(S,A) = entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times entropy(S_i) \quad (8)$$

$$splitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{D_j}{D}) \quad (9)$$

$$gainRatio = \frac{Gain(S,A)}{splitInfo_A(D)} \quad (10)$$

2.2.5. CART

Model *classification and regression tree* (CART) menggunakan penghitungan $IndexGini$ untuk pembentukan cabang. Sedangkan untuk pembentukan node, pada algoritme CART digunakan penghitungan $GiniGain$. Formula penghitungan $IndexGini$ dan $GiniGain$ dapat dilihat pada persamaan 11 dan 12.

$$IndexGini = 1 - \sum_{i=1}^{k} P_i^2 \quad (11)$$

$$GiniGain = Gini(A,S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Gini(S_i) \quad (12)$$

2.3. Validasi Model

Kami menggunakan 10 cross validation untuk pembelajaran dan pengujian data secara otomatis. Hal ini dilakukan karena laporan dari [13]–[14], dan pada beberapa tes juga menunjukkan bahwa penggunaan stratifikasi validasi model sedikit meningkatkan hasil. Ini berarti bahwa kami membagi data pelatihan menjadi 10 bagian yang sama dan kemudian melakukan proses pembelajaran 10 kali. Seperti yang ditunjukkan pada Tabel 2, setiap kali, kami memilih bagian dataset lain untuk pengujian dan menggunakan sembilan bagian yang tersisa untuk belajar. Setelah itu, kami menghitung nilai rata-rata dan nilai deviasi dari sepuluh hasil pengujian yang berbeda. Kami menggunakan validasi silang bertingkat 10 kali lipat, karena metode ini telah menjadi metode validasi standar dan canggih dalam istilah praktis [15].

Tabel 2. 10 Cross Validation

n-Validasi	Dataset partisi									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

2.4. Evaluasi Model

Pada penelitian ini, model yang diusulkan dievaluasi menggunakan *classifier effectiveness*. Hasil dari proses ini akan menghasilkan *confusion matrix* (matriks kebingungan) yang berisi nilai *true positive* (TP), *true negative* (TN), *false positive* (FP) dan *false negative* (FN). TP berarti ketika label yang diprediksi adalah "menggunakan" dan label sebenarnya adalah "menggunakan" juga. Ketika label yang diprediksi adalah "menggunakan" tetapi label sebenarnya adalah "tidak-menggunakan", itu disebut FP. TN sama dengan TP tetapi dalam hal label tidak-menggunakan, sedangkan FN adalah ketika label yang diprediksi adalah "tidak-menggunakan" tetapi sebenarnya labelnya adalah "menggunakan". Evaluasi di atas dihitung berdasarkan matriks kebingungan yang dihasilkan dari model. Berdasarkan matriks kebingungan, perhitungan pengukuran adalah sebagai berikut:

(i) SN : mengukur proporsi instance pola positif yang diakui dengan benar sebagai positif

$$SN = TP / (TP + FN) \quad (13)$$

(ii) SP : mengukur proporsi instance pola negatif yang dikenali dengan benar sebagai negative

$$SP = TN / (TN + FP) \quad (14)$$

(iii) PPV : mengukur probabilitas bahwa instance pola yang diprediksi secara positif dilabeli sebagai positif

$$PPV = TP / (TP + FP) \quad (15)$$

(iv) NPV : mengukur probabilitas bahwa instance pola prediksi negatif dilabeli sebagai negative

$$NPV = TN / (TN + FN) \quad (16)$$

(v) ACC : mengukur persentase sampel yang diklasifikasikan dengan benar

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (17)$$

(vi) F-measure : ukuran yang menggabungkan nilai prediksi positif dan sensitivitas

$$F\text{-measure} = (2 * PPV * SN) / (PPV + SN)$$

Penelitian ini juga menerapkan *area under curve* (AUC) sebagai akurasi indikator dalam percobaan untuk mengevaluasi kinerja pengklasifikasi. AUC adalah area under curve ROC. Lessmann et al. [16] menganjurkan penggunaan AUC untuk meningkatkan komparabilitas studi silang. AUC memiliki potensi untuk secara signifikan meningkatkan konvergensi lintas eksperimen empiris dalam klasifikasi status penggunaan metode kontrasepsi pada usia subur di Indonesia, karena memisahkan kinerja klasifikasi dari kondisi operasi, dan mewakili ukuran umum pengklasifikasi. Selanjutnya, AUC memiliki interpretasi statistik yang jelas. Ini mengukur probabilitas bahwa classifier membuat peringkat status pengguna rawan kesalahan yang dipilih secara acak lebih tinggi daripada status pengguna yang tidak rawan dipilih secara acak. Akibatnya, setiap pengklasifikasi yang mencapai AUC jauh di atas 0,6 terbukti efektif untuk mengidentifikasi status pengguna metode kontrasepsi rawan kesalahan dan memberikan saran yang berharga tentang status penggu mana yang harus mendapat perhatian khusus pada klasifikasi status pengguna metode kontrasepsi.

Panduan kasar untuk mengklasifikasikan keakuratan tes diagnostik menggunakan AUC adalah sistem tradisional, yang disajikan oleh Gorunescu [17]. Dalam kerangka yang diusulkan, kami menambahkan simbol untuk interpretasi dan pemahaman AUC yang lebih mudah (Tabel 3).

Tabel 3. Nilai AUC dan deskripsinya

AUC	Deskripsi
0.90 – 1.00	klasifikasi yang sangat baik
0.80 – 0.90	klasifikasi yang baik
0.70 – 0.80	klasifikasi yang adil
0.60 – 0.70	klasifikasi buruk
< 0.60	klasifikasi gagal

3. HASIL DAN ANALISIS

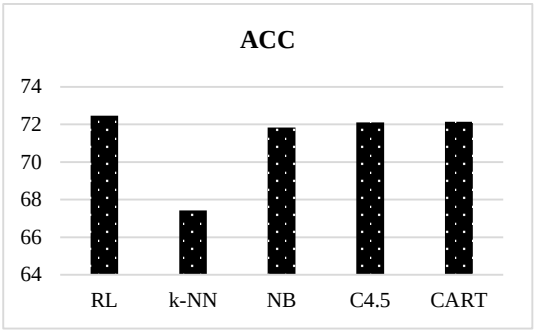
Percobaan dilakukan dengan menggunakan platform komputasi berbasis Intel Celeron 2.16 GHz CPU, 8 GB RAM, dan sistem operasi Microsoft Windows 10 Pro 64-bit. Lingkungan pengembangan model menggunakan Rapid Miner 6 dan SPSS versi 23.

Kami telah melakukan percobaan pada dataset SDKI 2017 dengan menggunakan 5 algoritma klasifikasi. Dalam penelitian ini classification effectiveness digunakan untuk mengevaluasi model. Percobaan pertama kami menguji semua model dengan data SDKI 2017 tanpa pre-processing. Tabel 4 menunjukkan laporan percobaan pertama. Hasil percobaan pertama, berdasarkan hasil *accuracy* (ACC) dan *area under curve* (AUC), dilaporkan bahwa model Regresi Logistik (RL) mendapatkan performa yang baik

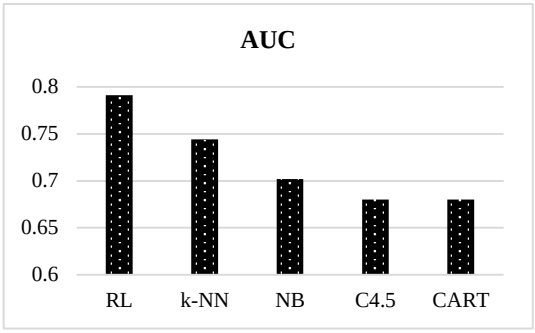
dibandingkan model pembanding lainnya. Model terbaik kedua oleh k-NN (48.75 % & 0.756), posisi ketiga oleh NB (53.64% & 0.684), selanjutnya oleh CART (54.04% & 0.662) dan C4.5 (54.04% & 0.662). Diagram perbandingan model berdasarkan ACC dan AUC dilaporkan pada Gambar 2, dan 3.

Tabel 4. Hasil evaluasi percobaan pertama tanpa pre-processing pada semua model hanya menggunakan *accuracy* (ACC), dan *area under curve* (AUC).

Model	ACC (%)	AUC
RL	**57.47**	**0.776**
k-NN	48.37	0.756
NB	53.64	0.684
C4.5	53.90	0.662
CART	54.04	0.662



Gambar 2. Diagram perbandingan model klasifikasi pada percobaan pertama tanpa pre-processing berdasarkan *acccuracy* (ACC).

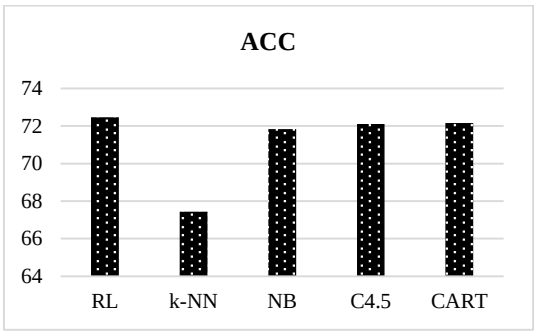


Gambar 3. Diagram perbandingan model klasifikasi pada percobaan pertama tanpa pre-processing berdasarkan *area under curve* (AUC) mean.

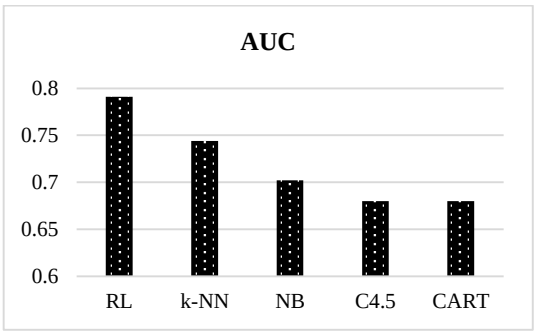
Percobaan kedua, kami menguji semua model menggunakan dataset yang sama dan menggunakan teknik ekstraksi variabel dan pembersihan data yang missing untuk menemukan variabel yang berpengaruh untuk status penggunaan metode kontrasepsi pada pasangan usia subur. Tabel 5 menunjukkan laporan percobaan kedua. Hasil percobaan kedua berdasarkan hasil *accuracy* (ACC) dan AUC dilaporkan bahwa model RL unggul di kedua kalinya dibandingkan model pembanding lainnya. Model terbaik kedua oleh k-NN (48.75 % & 0.756), posisi ketiga oleh NB (53.64% & 0.684), selanjutnya oleh CART (54.04% & 0.662) dan C4.5 (54.04% & 0.662). Bila dicermati hasil menunjukkan dari percobaan pertama dan percobaan kedua terdapat hasil yang konsisten yaitu RL selalu unggul dibandingkan model lainnya baik berdasarkan ACC maupun AUC.

Tabel 5. Hasil evaluasi percobaan kedua menggunakan teknik pre-processing pada semua model hanya menggunakan *accuracy* (ACC), dan *area under curve* (AUC).

Model	ACC (%)	AUC
RL	**72.47**	**0.791**
k-NN	67.43	0.744
NB	71.84	0.702
C4.5	72.11	0.680
CART	72.15	0.680



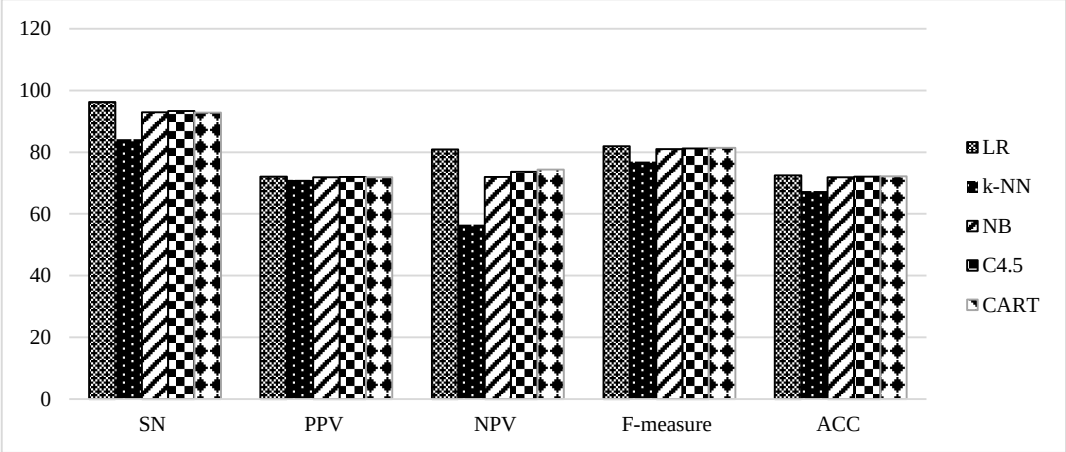
Gambar 4. Diagram perbandingan model klasifikasi pada percobaan kedua menggunakan pre-processing berdasarkan *acccuracy* (ACC).



Gambar 5. Diagram perbandingan model klasifikasi pada percobaan kedua menggunakan pre-processing berdasarkan *area under curve* (AUC) mean.

Seperti yang ditunjukkan pada Tabel 5, Gambar 4, dan Gambar 5 dilaporkan terjadi peningkatan performa yang signifikan setelah menambahkan metode ekstraksi variabel dan penghapusan data. Hal ini sesuai dengan hasil laporan penelitian sebelumnya [18], [19], [14] bahwa metode seleksi variabel dapat meningkatkan performa model-model klasifikasi tradisional.

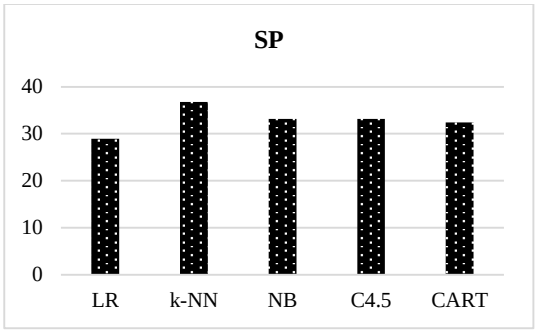
Percobaan terakhir, kami membandingkan model yang diusulkan dan model pembanding disajikan dengan menggunakan tujuh evaluasi model termasuk *sensitivity* (SN), *specificity* (SP), *positive predictive value* (PPV), *negative predictive value* (NPV), F-measure, *accuracy* (ACC), and *area under curve* (AUC).



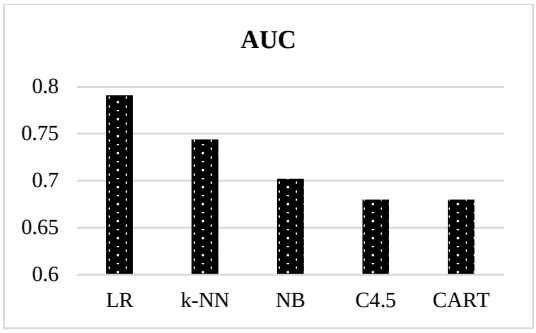
Gambar 6. Diagram perbandingan model klasifikasi pada percobaan ketiga menggunakan pre-processing berdasarkan SN, PPV, NPV, F-measure, dan ACC.

Tabel 6. Hasil evaluasi percobaan menggunakan classification effectiveness pada lima model klasifikasi.

Model	SN (%)	SP (%)	PPV (%)	NPV (%)	F-measure (%)	ACC (%)	AUC (%)	R
LR	**96.18**	**28.96**	**72.11**	**80.88**	**81.89**	**72.47**	**0.791**	**61.90**
k-NN	84.16	36.75	71.02	56.57	76.91	67.43	0.744	56.23
NB	92.91	33.18	71.87	71.96	81.03	71.84	0.702	60.50
C4.5	93.33	33.18	71.97	73.59	81.24	72.11	0.680	60.87
CART	92.82	32.40	71.88	74.37	81.36	72.15	0.680	60.81



Gambar 7. Diagram perbandingan model klasifikasi pada percobaan ketiga menggunakan pre-processing berdasarkan *specificity* (SP).



Gambar 8. Diagram perbandingan model klasifikasi pada percobaan ketiga menggunakan pre-processing berdasarkan *area under curve* (AUC).

Seperti yang ditujukkan pada pada Tabel 6, LR mengungguli model pembanding lainnya di semua evaluasi. Untuk SN, PPV, NPV, F-measure, ACC, dan AUC nilai tertinggi adalah mengidentifikasi performa model yang baik (ditampilkan pada Tabel 6, Gambar 7, dan Gambar 8), berbeda pada evaluasi SP di mana prediksi kelas negatif nilai terendah berarti memiliki nilai terbaik.

4. KESIMPULAN

Hasil percobaan menunjukkan dari kerangka model yang diusulkan menunjukkan performa yang variative. Penggunaan ekstraksi variabel berdasarkan inklusi peneliti dan melakukan pembersihan data pada data yang missing pada data SDKI 2017 berdasarkan faktor yang berpengaruh terhadap penggunaan metode kontrasepsi terbukti meningkatkan kinerja Regresi Logistik (RL). Oleh karena itu, dapat disimpulkan bahwa ekstraksi variabel dan pembersihan data mampu meningkatkan kinerja RL untuk klasifikasi status penggunaan metode kontrasepsi pada pasangan usia subur di Indonesia.

Penelitian di masa depan akan berkaitan dengan pemilihan variabel untuk meningkatkan performa akurasi, karena set data SDKI 2017 memiliki banyak variabel yang tidak semuanya relevan. Menggunakan teknik meta-learning memiliki peluang untuk meningkatkan kinerja RL.

UCAPAN TERIMA KASIH

Penulis berterima kasih kepada Ahmad Ilham dari Universitas Muhammadiyah Semarang yang telah memberi meteri dan bersedia diskusi tentang metodologi penelitian dan teknik penulisan artikel ilmiah, serta kepada teman-teman yang tergabung dalam Data Science Indonesia Regional Jawa Tengah atas diskusi hangat yang membangun terkait topik penelitian di atas.

REFERENCES

[1] C. C. Aggarwal, *Data Mining: The Textbook*. new york: Springer Berlin Heidelberg, 2015.

[2] L. Khikmah, H. Wijayanto, and U. D. Syafitri, "Modeling Governance KB with CATPCA to Overcome Multicollinearity in the Logistic Regression Modeling," 2017.

[3] P. Hariyani, C. Dewi, B. Notobroto, and D. Biostatistika, "Rendahnya Keikutsertaan Pengguna Metode Kontrasepsi Jangka Panjang Pada Pasangan Usia Subur," *Biometrika dan Kependud.*, vol. 3, no. 1, pp. 66–72, 2015.

[4] R. Septalia and N. Puspitasari, "Faktor yang Memengaruhi Pemilihan Metode Kontrasepsi," pp. 91–98.

[5] J. A. Vallejos and S. D. McKinnon, "Logistic regression and neural network classification of seismic records," *Int. J. Rock Mech. Min. Sci.*, vol. 62, pp. 86–95, 2013.

[6] A. Agresti, *Categorical Data Analysis - 2nd Ed.*, vol. 13. 2002.

[7] B. Lantz, *Machine Learning with R*. Birmingham: Packt Publishing Ltd, 2013.

[8] A. A. Aburomman and M. Bin Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Appl. Soft Comput. J.*, vol. 38, pp. 360–372, 2016.

[9] J. Suntoro, F. Wahyu, and H. Indriyawati, "Software Defect Prediction Using AWEIG + ADACOST Bayesian Algorithm for Handling High Dimensional Data and Class Imbalanced Problem," *Int. J. Inf. Technol. Bus.*, vol. 1, no. 1, pp. 36–41, 2018.

[10] Han and Kamber, *Data Mining Concepts and Techniques Third Edition*, vol. 1. 2012.

[11] M. Kantardzic, *Data Mining : Concepts, Models, Methods, and Algorithms*, Second Edi. Canada: A John Willey & Sons, Inc., Publication, 2011.

[12] D. Ryu and J. Baik, "Effective multi-objective naïve Bayes learning for cross-project defect prediction," *Appl. Soft Comput. J.*, vol. 49, pp. 1062–1077, 2016.

[13] R. Kosfeld and J. Lauridsen, "Factor analysis regression," *Stat. Pap.*, vol. 49, no. 4, pp. 653–667, 2008.

[14] A. Ilham, L. Khikmah, A. Qahslim, I. B. A. Indra Iswara, F. E. Laumal, and R. Rahim, "A systematic literature review on attribute independent assumption of Naive Bayes: research trend, datasets, methods and frameworks," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, p. 012086, Oct. 2018.

[15] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. MK Morgan Kaufman, 2011.

[16] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings," *IEEE Trans. Softw. Eng.*, vol. 34, no. 4, pp. 485–496, Jul. 2008.

[17] F. Gorunescu, *Data Mining: Concepts,Models and Techniques*. Springer Berlin Heidelberg, 2011.

[18] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Four. Morgan Kauvmann, 2016.

[19] R. S. Wahono and N. S. Herman, "Genetic Feature Selection for Software Defect Prediction," *Adv. Sci. Lett.*, vol. 20, no. 1, pp. 239–244, Jan. 2014.

BIBLIOGRAFI PENULIS

Laelatul Khikmah. Memperoleh gelar S.Si dari Prodi Studi Statistika Fakultas MIPA Universitas Islam Indonesia, dan gelar M.Si dari Program Pascasarja Statistika IPB. Dia saat ini sebagai dosen program D3 Statika di Akademi Statistika Muhammadiyah Semarang, Indonesia. Selain itu dia tergabung di Data Science Indonesia Regional Jawa Tengah. Minat Penelitian saat ini di bidang *data mining*, Regresi Logistik dan Data Kategorik.