 E-ISSN 2548-7779

Performa Klasifikasi K-NN dan *Cross-validation* pada Data Pasien Pengidap Penyakit Jantung

Huzain Azis[a,1,*], Purnawansyah [a,2], Farniwati Fattah[a,3] dan Inggrianti Pratiwi Putri [a,4]

[a] *Universitas Muslim Indonesia, Urip Sumoharjo km.5, Makassar 90231, Indonesia*
[1] *huzain.azis@umi.ac.id;* [2]*Purnawansyah@umi.ac.id;* [3] *farniwati.fattah@umi.ac.id;* [4]*pratiwiinggrianti@gmail.com*
**corresponding author*

INFORMASI ARTIKEL

Dikirim : 27 Juli 2020
Diulas : 01 Agustus 2020
Direvisi : 15 Agustus 2020
Diterbitkan : 27 Agustus 2020

Kata Kunci:
K-Nearest Neighbor
Cross-validation
Analisis performa
Penyakit *cardiovascular*

ABSTRAK

Secara global, penyebab kematian nomor satu setiap tahunnya adalah penyakit *cardiovascular*. Penyakit *cardiovascular* adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah, seperti Penyakit Jantung Koroner, Penyakit Gagal jantung atau Payah Jantung, Hipertensi dan Stroke. Tujuan dari penelitian ini adalah mengukur performa akurasi, presisi, *recall* dan *f-measure* metode K-NN dan Crossvalidation pada dataset pasien pengidap *cardiovascular*. dataset yang digunakan sebanyak 1000 *record* terdiri dari 11 atribut (*age, gender, height*, dll.) data pasien *cardiovascular* dan non *cardiovascular*, dataset tersebut diperoleh dari UCI *Machine Learning Repository* yang dikelola oleh *Hungarian Institute of Cardiology Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland*. Tahapan yang dilakukan yaitu: membagi rasio simulasi dataset 20:80, 50:50 dan 80:20, penerapan *crossvalidation* (k-fold=10) dan klasifikasi menggunakan metode K-NN (k=2 hingga K=900). Hasil penelitian dari simulasi *rasio dataset* 50:50 memperoleh nilai akurasi 82%, presisi 82%, *recall* 82% dan *f-measure* 80% pada nilai K=13, kemudian hasil penelitian dari simulasi *rasio dataset* 20:80 memperoleh nilai akurasi 87%, presisi 87%, recall 97% dan *f-measure* 92% pada nilai K=3, serta hasil penelitian dari simulasi rasio dataset 80:20 memperoleh nilai akurasi 91%, presisi 92%, recall 60% dan f-measure 72% pada nilai K=5.

Keywords:
K-Nearest Neighbor
Crossvalidation
Perform analysis
Cardiovascular

ABSTRACT

Globally, the number one cause of death each year is cardiovascular disease. Cardiovascular disease is a disease caused by impaired function of the heart and blood vessels, such as coronary heart disease, heart failure or heart failure, hypertension and stroke. The purpose of this study was to measure the performance of accuracy, precision, recall and f-measure of the K-NN and Crossvalidation methods on a dataset of cardiovascular patients. The dataset used was 1000 records consisting of 11 attributes (age, gender, height, etc.) cardiovascular and non cardiovascular patient data, the dataset was obtained from the UCI Machine Learning Repository managed by the Hungarian Institute of Cardiology Budapest: Andras Janosi, MD, University Hospital, Zurich, Switzerland. The steps taken are: dividing the simulation ratio of the dataset to 20:80, 50:50 and 80:20, applying crossvalidation (k-fold = 10) and classification using the K-NN method (k = 2 to K = 900). The research results from the simulation of the dataset ratio 50:50 obtained an accuracy value of 82%, 82% precision, 82% recall and 80% f-measure at a value of K = 13, then the research results from the simulation of the dataset ratio 20:80 obtained an accuracy value of 87%, 87% precision, 97% recall and 92% f-measure at the value of K = 3, and the results of research from the simulation of the dataset ratio 80:20 obtained an accuracy value of 91%, 92% precision, 60% recall and 72% f-measure at the value K = 5.

I. Pendahuluan

Cardiovascular merupakan organ manusia yang berperan dalam sistem peredaran darah. Penyakit *cardiovascular* adalah sebuah kondisi jantung tidak dapat melaksanakan tugasnya dengan baik. Secara global, penyebab kematian nomor satu setiap tahunnya adalah penyakit *cardiovascular*. Penyakit *cardiovascular* adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah, seperti penyakit jantung koroner,

penyakit gagal jantung atau payah jantung, hipertensi dan stroke. Berdasarkan data dari kemenkes RI Pada tahun 2013 diperkirakan sebanyak 12,3 juta kematian disebabkan oleh penyakit cardiovascular. Lebih dari 3 juta kematian tersebut terjadi sebelum usia 60 tahun dan seharusnya dapat dicegah. Masyarakat berusia 20 hingga 40 menderita penyakit cardiovascular sebanyak 37% sedangkan yang berusia 40 hingga 60 sebanyak 71%.

Penentuan kemiripan suatu data dapat menggunakan metode pengukuran jarak, salah satunya yaitu K-Nearest Neighbor (K-NN). K-NN merupakan metode pengukuran yang paling sering digunakan untuk menentukan kesamaan dua vektor. Berdasarkan penelitian sebelumnya[1], data yang digunakan 110 records pasien. 100 *records* digunakan sebagai data latih (*training data*) dan 10 records digunakan sebagai data uji (testing data). Untuk menentukan apakah seorang pasien terkena penyakit jantung digunakan 9 data terdekat atau K= 9. Perhitungan kedekatan data training dengan kasus pada data testing 3 data diprediksikan masuk kedalam kelas "1" tetapi ternyata termasuk kedalam kelas "2". Dari 100 data training dan 10 data testing dan menggunakan metode K-NN dengan nilai K = 9 diperoleh tingkat akurasi sebesar 70%.

Lanjutan penelitian tersebut[2], dengan menggunakan metode yang sama yaitu K-NN dengan jumlah data yang di tingkatkan sebesar 1000 data pasien serta melakukan simulasi nilai K=3 hingga K=9. Hasil yang diperoleh dari penelitiannya yaitu nilai K=6 memiliki nilai akurasi paling baik yaitu sebesar 85%, dan nilai presisinya 78%, recall 93%, dan F-measure 85%. Diakhir penelitiannya Hasran menyebutkan bahwa nilai akurasi 85% belum cukup baik maka di perlukan beberapa tindakan diantaranya menerapkan *crossvalidation*[3]–[5] untuk simulasi data pada saat pengujian. Penelitian ini berupaya melanjutkan penelitian sebelumnya yaitu dengan menganalisis performa metode K-NN serta menerapkan *Crossvalidation* pada data pasien *Cardiovascular*, harapan pada penelitian ini yaitu diperolehnya nilai performa yang baru dan dapat menjadi pembanding dengan penelitian-penelitian sebelumnya.

II. Metode

A. *Datamining*

Datamining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi. Proses dalam tahap data mining terdiri dari tiga langkah utama, yaitu : a. Data Preparation pada langkah ini, data dipilih, dibersihkan, dan dilakukan preprocessed mengikuti pedoman dan *knowledge* dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh. b. Algoritma data mining penggunaan algoritma data mining dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai[5]–[9].

B. *K-Nearest Neighbor*

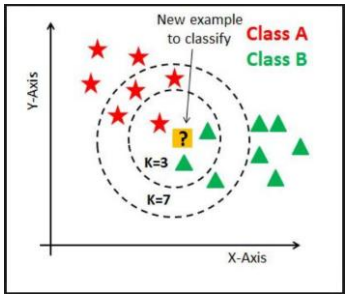
K-Nearest Neighbor (K-NN) termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik lazy learning. K-NN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Untuk mencari solusi dari pasien baru tersebut digunakan kedekatan dengan kasus pasien lama, solusi dari kasus lama yang memiliki kedekatan dengan kasus baru digunakan sebagai solusinya.

Terdapat pasien baru dan 4 pasien lama, yaitu P, Q, R, dan S. Ketika ada pasien baru maka yang diambil solusi adalah solusi dari kasus pasien lama yang memikili kedekatan terbesar. Misal D1 adalah jarak antara pasien baru dengan pasien P, D2 adalah jarak antara pasien baru dengan pasien Q, D3 adalah jarak antara pasien baru dengan pasien R, D4 adalah jarak antara pasien baru dengan pasien S. Dari ilustrasi gambar terlihat bahwa D2 yang paling terdekat dengan kasus baru. Dengan demikian maka solusi dari kasus pasien Q yang akan digunakan sebagai solusi dari pasien baru tersebut.

Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (data training), diantaranya *euclidean distance* dan *manhattan distance* (*city block distance*), yang paling sering digunakan adalah *euclidean distance*. Persamaan *euclidean* ditunjukkan pada persamaan 1 [2], [10], [11].

$$euc = \sqrt{((a\_1 - b\_1)\text{^}2 + (a\_2 - b\_2)\text{^}2 + \cdots + (a\_n - b\_n)\text{^}2)} \quad (1)$$

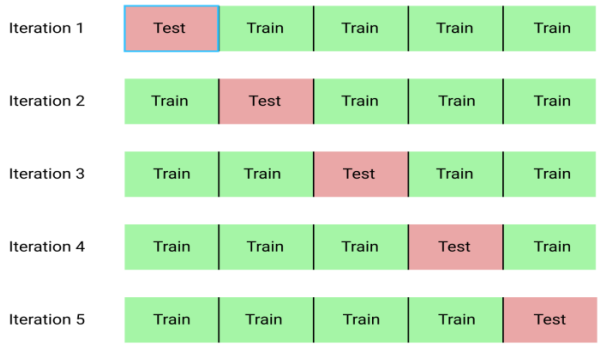
Alur kerja K-NN diilustrasikan pada Gambar 1.



Gambar 1. Contoh Alur Kerja K-NN

C. Crossvalidation

Crossvalidation atau dapat disebut estimasi rotasi adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen[12][13]. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Salah satu teknik dari validasi silang adalah *k-fold cross-validation*, yang mana memecah data menjadi K bagian set data dengan ukuran yang sama. Penggunaan *k-fold*=5 *cross-validation* untuk menghilangkan bias pada data. Pelatihan dan pengujian dilakukan sebanyak K. Alur kerja *cross-validation* diilustrasikan pada Gambar 2.



Gambar 2. Contoh simulasi *crossvalidation*.

D. Analsis Performa

Tahap terakhir setelah penerapan metode klasifikasi adalah menghitung performa, adapun persamaan performa menggunakan *confusion matrix* yang mana untuk membantu menghitung nilai akurasi, presisi, *recall* dan *f-measure*[14]–[16]. Persamaan yang digunakan untuk menghitung akurasi ditunjukkan pada persamaan 2, penerapan perhitungan akurasi yang digunakan pada penelitian ini adalah *balanced-accuracy* (ba) dimana berfungsi untuk menangani *multiclass* klasifikasi dengan data yang tidak seimbang [17], [18].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

$$b.a = 1/2(\frac{TP}{TP + FN} + \frac{TN}{TN + FP})$$

Proses klasifikasi *multiclass*, perhitungan performa presisi, *recall*, dan *f-measure* dapat diterapkan pada setiap label secara *independen*[19], [20]. presisi menggunakan persamaan 3, persamaan 4 menunjukan perhitungan performa *recall* dan persamaan 5 untuk *f-measure* [21], [22].

$$precision = \frac{TP}{TP + FP} \qquad (3)$$

$$recall = \frac{TP}{TP + FN} \qquad (4)$$

$$f - measure = \frac{precision * recall}{precison + recall} \qquad (5)$$

III. Hasil dan Pembahasan

Implementasi metode K-NN pada tahap ini merupakan contoh perhitungan manual mulai dari penetapan data latih dan data uji, implementasi metode K-NN hingga perhitungan performa. Berikut ini 12 *sample* data yang akan di terapkan untuk perhitungan manual di tunjukkan pada Tabel 1 menunjukan 12 sampel data.

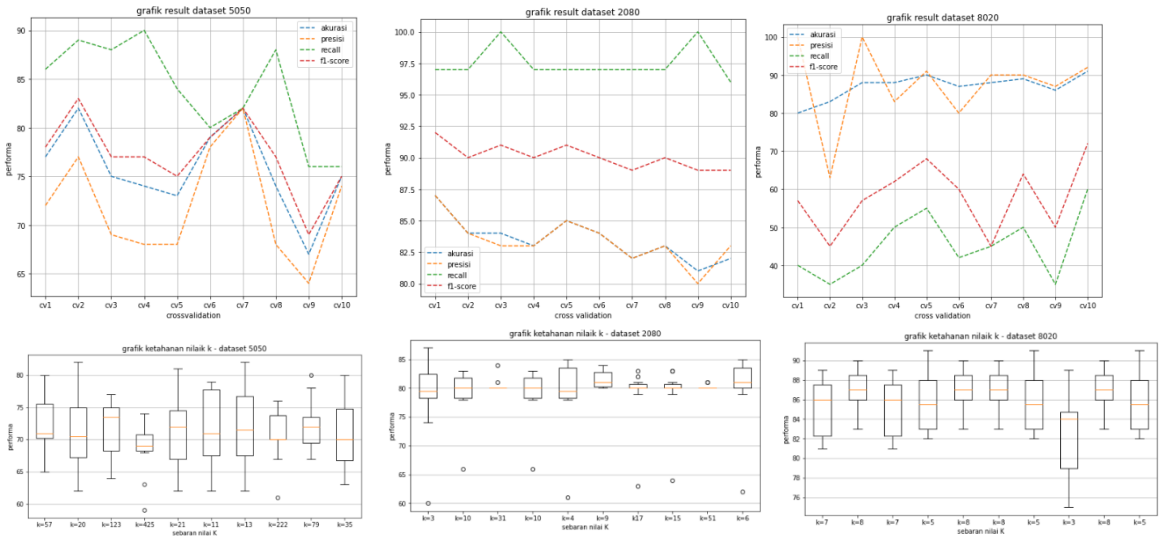
Tabel 1 *DataSet Sample*

age	gender	Height	weight	ap_hi	ap_lo	chol	gluc	smoke	alco	acti ve	cardio
16003	1	162	74.0	140	100	2	1	0	0	1	0
16039	2	180	90.0	140	90	2	2	0	0	0	0
19029	1	115	64.0	120	70	1	1	0	0	1	0
22821	2	168	80.0	160	100	1	1	0	0	1	0
20395	1	164	70.0	120	70	1	1	0	0	1	0

age	gender	Height	weight	ap_hi	ap_lo	chol	gluc	smoke	alco	acti ve	cardio
21473	2	174	90.0	140	80	1	1	1	1	0	1
21758	1	165	65.0	140	90	1	1	0	0	1	1
21166	2	162	59.0	140	90	2	1	1	0	1	1
23444	1	175	76.0	120	80	1	1	0	1	1	1
18958	1	149	62.0	180	110	2	1	0	0	1	1

Hal yang dilakukan setelah mempersiapkan data adalah melakukan pembagian data latih dan data uji sesuai perencanaan yaitu menggunakan konsep *cross-validation*. Adapun jumlah *k-fold* yang digunakan adalah *k-fold*=10. Gambar 3 menunjukkan implementasi *croos-validation* pada penelitian ini

Gambar 3. Implementasi *cross-validation k-fold*=10



Gambar 4. Hasil implementasi metode KNN dan *cross-validation*

Setelah proses pembagian data telah dilakukan, maka tahap selanjutnya adalah penerapan metode K-NN, implementasi metode K-NN pada penelitian ini menggunakan *library machine learning sklearn*, penerapannya dilakukan pada pembagian data *cross-validation* yang telah dilakukan sebelumnya dan pada rasio data 20:80, 80:20 serta 50:50. Nilai ketetanggapan yang diuji coba adalah nilai K=2 hingga k=jumlah data uji atau sebesar k=900.

Gambar 4 menunjukkan tahap akhir dari penelitian ini yaitu merepresentasikan hasil pengujian performa yaitu akurasi, presisi, *recall* dan *f-measure*, keseluruhan output dari penelitian ini, dibagi menjadi dua bagian yaitu *line graph* dan *boxplot.* Pada *line graph* terbagi menjadi tiga bagian yang mana mewakili setiap rasio yaitu rasio 20:80, 80:20 dan 50:50. Setiap *line graph* merepresentasikan nilai akurasi, presisi dan *recall* untuk 10 *k-fold crossvalidation*. Sedangkan representasi *boxplot* pada Gambar 4 terbagi menjadi tiga bagian yang mewakili setiap rasio dimana nilai yang dibentuk menjadi *boxplot* tersebut adalah 10 nilai akurasi tertinggi pada setiap *crossvalidation*. Adanya *boxplot* maka dapat terlihat nilai-nilai performa yang menjadi nilai outlier atau nilai yang tidak dapat dijadikan acuan sebagai model.

Berdasarkan Gambar 4 dapat dilihat bahwa terdapat tiga nilai performa terbaik pada setiap rasionya yaitu 82%, 87% dan 91% untuk akurasinya, namun performa terbaik dengan mempertimbangkan nilai akurasi, presisi, *recall*, *f-measure* serta representasi *boxplot* jatuh pada rasio 80:20 dengan nilai akurasi 91%, dikarenakan nilai tersebut memiliki keseimbangan antara presisi dan recall yang paling baik serta pada rasio tersebut nilai outlier pada *boxplot* tidak ditemukan. Tabel 2 merepresentasikan keseluruhan hasil pengujian disertai dengan perbandingan hasil pengujian penelitian ini dengan penelitian sebelumnya.

Tabel 2. Perbandingan hasil penelitian

No	Peneliti	Metode	Jumlah / Rasio dataset	Jumlah Data Training	Jumlah Data Testing	Akurasi	Presisi	Recall	f-measure
1.	Mei (2014) [1]	K-NN K= 6	110	100	10	85%	78%	93%	85%
2.	Hasran (2019) [2]	K-NN K=6	1000 / 50:50	900	100	85%	79%	93%	85%
		K-NN K=7	1000 / 80:20	900	100	82%	46%	47%	41%
3.	Huzain (2020)	K-NN K=13	1000 / 50:50	900	100	82%	82%	82%	82%
		K-NN K=3	1000 / 20:80	900	100	87%	87%	97%	92%
		K-NN K=5	1000 / 80:20	900	100	91%	92%	60%	72%

IV. Kesimpulan

Berdasarkan hasil penelitian ini maka penulis dapat menarik kesimpulan yaitu dataset1 (dataset 50:50) di peroleh nilai performa paling baik pada nilai akurasi sebesar 82%, presisi 82%, *recall* 82% dan *f-measure* 82%, pada K=13. Dataset2 (dataset 20:80) di peroleh nilai performa paling baik pada nilai akurasi sebesar 87%, presisi 87%, *recall* 97%, dan *f-measure* 92%, pada K=3. Dataset3 (dataset 80:20) di peroleh nilai performa paling baik pada nilai akurasi sebesar 91%, presisi 92%, *recall* 60% dan *f-measure* 72%, pada K=5. Performa terdapat pada rasio 80:20 dengan akurasi 91% dengan pertimbangan bahwa baiknya keseimbangan nilai presisi dan recall serta tidak adanya nilai *outlier* pada *boxplot.* Berdasarkan seluruh nilai yang telah didapatkan serta hasil perbandingan dengan penelitian sebelumnya, nilai performa yang diperoleh pada penelitian ini lebih baik di banding dengan nilai performa penelitian sebelumnya, serta menunjukkan bahwa nilai performa yang diperoleh dapat lebih baik karena dilakukan berbagai simulasi rasio data, penerapan *cross-validation* serta pengujian keseluruh nilai K pada K-NN

Daftar Pustaka

[1] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) Untuk Mendeteksi Penyakit Jantung," *Fakt. Exacta*, vol. 7, no. September 2010, pp. 366–371, 2014.

[2] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.

[3] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.

[4] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[5] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[6] Y. Lukito and A. R. Chrismanto, "Perbandingan Metode-Metode Klasifikasi untuk Indoor Positioning System," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, no. 2, pp. 123–131, 2015, doi: 10.28932/jutisi.v1i2.373.

[7] N. Fadhillah, H. Azis, and D. Lantara, "Validasi Pencarian Kata Kunci Menggunakan Algoritma Levenshtein Distance Berdasarkan Metode Approximate String Matching," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 3–7, 2018.

[8] A. A. Karim, H. Azis, and Y. Salim, "Kinerja Metode C4.5 dalam Penyaluran Bantuan Dana Bencana 1," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 84–87, 2018.

[9] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[10] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.

[11] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020.

[12] M. J. Hartmann and G. Carleo, "Neural-Network Approach to Dissipative Quantum Many-Body Dynamics," *Phys. Rev. Lett.*, vol. 122, no. 25, p. 250502, Jun. 2019, doi: 10.1103/PhysRevLett.122.250502.

[13] K. Crammer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res. - JMLR*, vol. 2, no. 2, pp. 265–292, 2002.

[14] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, p. 145, 2016, doi: 10.1504/ijapr.2016.079050.

[15] P. A. Flach and M. Kull, "Precision-Recall-Gain curves: PR analysis done right," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 838–846, 2015.

[16] L. Nurhayati and H. Azis, "Perancangan Sistem Pendukung Keputusan Untuk Proses Kenaikan Jabatan Struktural Pada Biro Kepegawaian," *Semin. Nas. Teknol. Inf. dan Multimed.*, pp. 6–7, 2016.

[17] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, 2018, doi: 10.1016/j.aci.2018.08.003.

[18] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3121–3124, 2010, doi: 10.1109/ICPR.2010.764.

[19] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Stat. Interface*, vol. 2, no. 3, pp. 349–360, 2009, doi: 10.4310/sii.2009.v2.n3.a8.

[20] R. Puri and K. Khamrui, "Application of Quantitative Descriptive Analysis (QDA), Principal Component Analysis (PCA) and Response Surface Methodology (RSM) in standardization of cham-cham making.," 2015.

[21] S. Paembonan *et al.*, "Combination of K-Means and Profile Matching for Drag Substitution," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Nov. 2018, pp. 180–183, doi: 10.1109/EIConCIT.2018.8878539.

[22] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.