BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 6, Oktober 2023 | Hal 442-447
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i6.290

Naïve Bayes Classifier dengan Particle Optimize Weight Forward pada Dataset

Nuranisah*, Yanti Yusman

Fakultas Sains Dan Teknologi, Program Studi Komputer, Universitas Pembangunan Panca Budi, Medan, Indonesia
Email: 1,*nuranisahasriel123@gmail.com, 2yantiyusman@gmail.com
Email Penulis Korespondensi: nuranisahasriel123@gmail.com

Abstrak−Klasifikasi adalah proses mengidentifikasi dan mengelompokkan objek ke dalam kelas atau kategori berdasarkan karakteristiknya. Dalam data mining ada dua proses yaitu klasifikasi dan clustering yaitu yang digunakan untuk mengelompokkan objek berdasarkan kesamaan. Dalam proses klasifikasi, berbagai metode salah satunya seperti K-NN, SVM, dan Naïve Bayes yang sering digunakan dan dilakukan perkembangan dalam metodenya. Pengklasifikasi Naïve Bayes terbukti memiliki keunggulan, seperti penghitungan yang lebih cepat dan akurasi yang lebih baik. Namun, metode ini memiliki keterbatasan dalam proses pemilihan atribut. Untuk mengatasi keterbatasan ini, digunakan algoritma Particle Optimize Weights Forward untuk meningkatkan akurasi dengan memberikan bobot pada atribut dalam metode Naïve Bayes. Pendekatan ini meningkatkan efisiensi dan efektivitas pengklasifikasi Naïve Bayes dalam tugas klasifikasi data.

Kata Kunci: Kata Kunci**:** Naïve Bayes Classifier, Particle Optimize Weights Forward, Klasifikasi, Data Mining

Abstract−Classification is the process of identifying and grouping objects into classes or categories based on their characteristics. In data mining, there are two processes, namely classification and clustering, which are used to group objects based on similarities. In the classification process, various methods such as K-NN, SVM, and Naïve Bayes are often used and developments are made in the method. The Naïve Bayes classifier is proven to have advantages, such as faster calculation and better accuracy. However, this method has limitations in the attribute selection process. To overcome this limitation, the Particle Optimize Weights Forward algorithm is used to improve accuracy by assigning weights to attributes in the Naïve Bayes method. This approach improves the efficiency and effectiveness of the Naïve Bayes classifier in data classification tasks.

Keywords: Naïve Bayes Classifier; Particle Optimize Weights Forward; Classification; Data Mining

1. PENDAHULUAN

Data mining merupakan proses dalam menganalisa data dari perspektif yang berbeda dan menyimpulkan agar mengandung informasi – informasi penting yang dapat digunakan untuk meningkatkan perhitungan yang bermanfaat. [1] Data mining meliputi beberapa kegiatan diantaranya pemakaian,pengumpulan data historis untuk menemukan keteraturan, pola atau hubungan data berukuran besar lalu menjadikan data tersebut menjadi informasi – informasi yang nantinya dapat dimanfaatkan. [2] Didalam data mining terdapat beberapa metode dalam melakukan klasifikasi data. Klasifikasi data merupakan proses identifikasi objek pada sebuah kategori, kelompok atau kelas dengan melalui beberapa prosedur yang telah ditetapkan. [3].

Klasifikasi mempunyai tujuan sebagai penempatan objek yang bertugas ke salah satu kategori yang disebut kelas [4]. Selain proses klasifikasi. Adapun metode pengelompokan pada objek juga dapat dilakukan dengan metode clustering. Clustering merupakan pembentukan dari kesatuan yang membentuk suatu kelompok pada objek yang berdasarkan dari kemiripan antar objek. Perbedaan yang dapat dilihat dari kedua proses tersebut adalah pada pengelompokan objek dimana klasifikasi alur proses dilakukan dengan membagikan objek dasar kelompok/kategori yang telah ditentukan sebelumnya. Sedangkan, pada proses clustering dilakukan dengan mencari kemiripan antar objek, sehingga tidak dapat terindentifikasi dari definisi sebelumnya. Salah satu metode klasifikasi adalah mentode naïve bayes yang merupakan suatu model klasifikasian probabilistik sederhana dengan proses perhitungan pada kumpulan probabilitas dimana penjumlahann frekuensi dan kombinasi nilai pada dataset yang ada. [5]

Secara keseluruhannya sudah banyak penelitian yang diangkat dengan metode naïve bayes menggunakan berbagai perkembangan metode dan dataset yang berbeda- beda. Beberapa penelitian yang menggunakan metode naïve bayes menjelaskan bahwa penggunaan metode naïve bayes yang di optimasikan dengan particle swarm optimize menghasilkan data akurasi sebesar 98,76% dengan menggunakan dataset penyakit tuberculios [6]. Lalu pada penelitian sebelumnya dengan menggunakan data iris yang nantinya akan digunakan sebagai dataset pengujian metode naïve bayes dalam penelitian dengan proses optimize yang berbeda, hasil dari penelitian tersebut menjelaskan bahwa metode naïve bayes dengan menggunakan particle swarm optimize dengan dataset iris menghasilkan fitness tertinggi dalam penentuan bobot atribut optimumnya. [7] Proses peningkatan akurasi banyak dilakukan oleh penelitian yang sebagian besar menggunakan PSO dalam peningkatan akurasi karena naïve bayes memiliki kelemahan pada seleksi atribut sehingga sering kali mempengaruhi nilai akurasi.

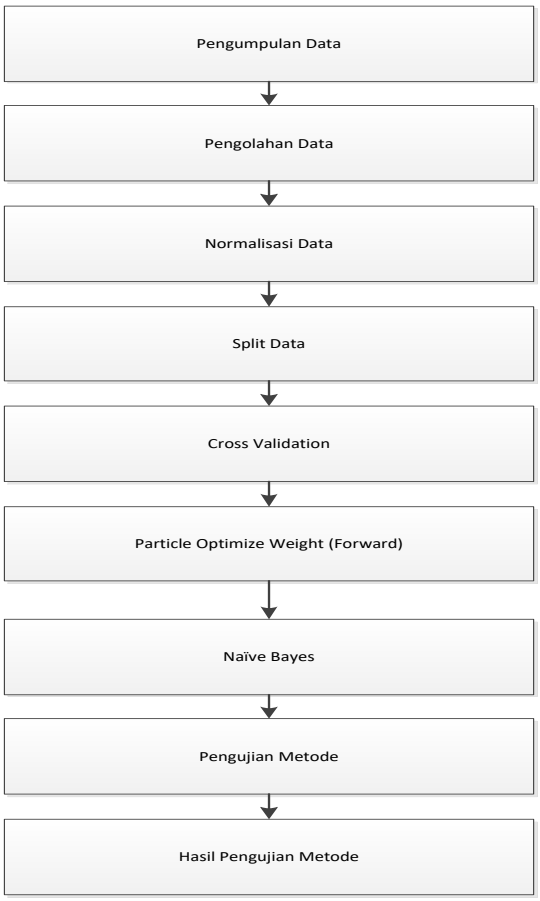
Penguji cobaan dengan algoritma Particle Optimize Weight Forward diharapkan mampu memberikan dampak hasil akurasi yang baik dengan melakukan pembobotan atribut. Data yang akan digunakan pada penelitian ini adalah data dari UCI Machine Learning yang terdiri dari 150 dataset terdapat 3 kelas dan 4 atribut yaitu sepal length, petal length, sepal width dan petal width.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 6, Oktober 2023 | Hal 442-447
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i6.290

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Proses dari tahapan – tahapan penelitian memerlukan metode yang dapat digunakan agar menjelaskan alur klasifikasi yang akan dilakukan sehingga mengetahui informasi dari pengolalan hasil klasifikasi dari dataset yang digunakan. Proses optimasi Naïve Bayes Classifier menggunakan algoritma Particle Optimize Weight Forward untuk melakukan optimasi bobot dari suatu dataset yang telah ditentukan yaitu pada data iris ditampilkan pada Gambar 1 berikut :



Gambar 1 Metode Penelitian

Berdasarkan gambar 1 bagan dari pada tahapan penelitian diatas , uraian dari setiap Langkah-langkah tersebut adalah proses pengumpulan data , data yang digunakan dalam penelitian ini adalah data yang bersumber dari UCI learning yaitu Data IRIS. Data Iris adalah salah satu data yang paling sering digunakan dalam proses pengembangan perhitungan klasifikasi data yang dapat diakses oleh siapapun karena bersifat publik pada uci repository.Dalam data iris ada 4 atribut yang dapat mempengaruhi klasifikasi yaitu: sepal length, sepal width, petal length, serta petal width. Atribut tujuan atau label dari data iris memiliki 3 kelas yaitu: iris setosa, iris versicolour, serta iris virginica. [8]. Pengolahan Data merupakan teknik yang digunakan agar data tersebut disesuaikan dengan metode pengklasifikasian yang akan diterapkan.

Setelah pengelolaan data maka diperlukan nya normalisasi data merupakan bagian daripada praposes data, pada normalisasi dilakukan terhadapa nilai- nilai yang tersimpan pada dataset sehingga proses pengelolaan akan menjadi mudah. Adanya rentang nilai yang terlalu jauh sehingga diperlukan normalisasi pada atribut- atribut tersebut sehingga diperlukannya normalisasi. Adapun normalisasi terbagi beberapa metode salah satunya adalah Z-score yang akan digunakan pada penelitian [9] [10]. Dimana Z-score normalization dapat dihitung dengan menggunakan rumus berikut :

$$Z= \frac{x-\bar{x}}{\sigma} \quad (1)$$

Ditahap selanjutnya adalah dengan menentukan rasio antara data latih dan data uji pada data yang akan digunakan sehingga mempermudah untuk proses perhitungan klasifikasi data.

2.2 Cross Validation

Crossvalidation atau dapat disebut estimasi rotasi merupakan suatu teknik daripada model validasi untuk menilai bagaimana hasil statistic analisis akan menggeneralisasi kumpulan data independent [M.J.Hartmann],[K.cranmer]. Teknik tersebut dapat dilakukan dengan melakukan prediksi pada model dan memperkirakan seberapa akurat suatu model

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 6, Oktober 2023 | Hal 442-447
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i6.290

prediktif ketika dijalankan kedalam prakteknya. Adapun salah satu satu teknik dari suatu validasi silang adalah k-fold cross validation, yang berfungsi memecahkan data menjadi K dibagian set data dengan ukuran yang sama.

2.3. Naïve Bayes

Naïve Bayes merupakan metode yang dapat memprediksi pada suatu class dari suatu objek yang mana kelasnya tidak diketahui dari masing – masing kelompok atribut yang ada sehingga penentuan class yang paling optimal dapat didasarkan pada pengaruh didapat di hasil pengamatan. [11] Naïve Bayes diasumsikan pada penyederhanaan nilai atribut secara kondisional saling bebas apabila diberikan nilai output. Sehingga nilai output probabilitas mengamai bagaimana cara suatu data dari probabilitas individu. [12] . Kelebihan dari naïve bayes adalah dalam metode ini hanya memerlukan jumlah data pelatihan (training data) yang bernilai minimun hingga penentuan estimasi parameter yang diperlukan pada pengklasifikasian. Sering kali naïve bayes bekerja jauh lebih efektif pada beberapa hasil kalkulasi dalam penelitian secara kompleks sesuai yang diharapkan. [13] Adapun persamaan dari naïve bayes sebagai berikut :

$$P(A|B)\frac{p(B|A)p(A)}{p(B)} \tag{2}$$

2.4. Optimasi

Pada tahap optimasi pada data yang akan dioptimasikan dapat dilakukan dengan menggunakan dua metode optimasi, yaitu Fungsi Select + Optimize Select sehingga dapat meiningkan weight pada atribut sehingga menghasilkan tingkat akurasi yang baik daripada sebelumnya.. [14]

2.5 Analisis Performa

Pada tahap analisis 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 [15] [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]

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\text{b. a} = 1/2(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}) \tag{4}$$

Proses klasifikasi multiclass, perhitungan performa presisi, recall, dan f-measure dapat diterapkan pada setiap label secara independen. [18] Presisi menggunakan persamaan 3, persamaan 4 menunjukan perhitungan performa recall dan persamaan 5 untuk f-measure [19].

$$recall = \frac{TP}{TP+FP} \tag{5}$$

$$f - measure = \frac{\text{precision} * \text{recall}}{\text{precison} + \text{recall}} \tag{6}$$

3. HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan data Iris yang diambil dari UCI Machine Learning. Data yang digunakan dibagi menjadi dua bagian yaitu data training dan data testing. Data tersebut terdiri dari 150 dataset yang terbagi menjadi 3 kelas dan 4 atribut, yaitu:

Class:
1. Iris Sentosa
2. Iris VersiColor
3. Iris Virginica

Atribut:
1. Sepal Length
2. Sepal Width
3. Petal Length
4. Petal Width

Metode yang digunakan adalah metode data mining naïve bayes yang di eksperimenkan dengan Tools Rapidminer. Proses penggabungan dengan algoritma particle Optimize Weight (forward) merupakan upaya untuk meningkatkan hasil akurasi yang dilakukan dengan menaikkan partikel itarasi partikel, jumlah iterasi, bobot inersia, konstanta kecepatan 1 dan 2. Jumlah partikel digunakan untuk menentukan banyaknya popsize pada algoritma tersebut. Partikel direpresentasikan dengan bobot tiap atribut yang akan dioptimasi. Kemudian setiap data training dan data testing akan di kalikan dengan bobot.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 6, Oktober 2023 | Hal 442-447
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i6.290

3.1 Penerapan Metode Naïve Bayes

Berdasarkan asumsi penyederhaan Naive Bayes suatu nilai atirbut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, nilai putput yang probabilitasnnya secara kondisional saling bebas jika diberikan nilai ouput. Maka pemberian nilai output , probabilitas akan mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naïve Bayes yaitu metode ini membutuhkan jumlah data pelatihan (Training Data) yang kecil agar dapat menentukan estimasi dari parameter yang akan digunakan dalam proses klasifikasinya. Didalam metode naïve bayes data string yang bersifat konstan dibedakan dengan data numerik yang bersifat kontinyu, sehingga perbedaan ini akan terlihat pada penentuan nilai probabilitas setiap kriteria baik itu dengan nilai data string maupun kriteria dengan nilai data numerik.

3.2 Dataset

Implementasi metode naïve bayes pada tahap ini merupakan contoh perhitungan manual mulai dari penetapan data latih dan data uji, implementasi metode naïve bayes hingga perhitungan performa. Berikut ini terdiri dari 150 dataset yang terbagi menjadi 3 kelas dan 4 atribut [20] Berikut tabel 1 [21] menunjukkan sample datanya.

Tabel 1. Data Sample

sepal length	sepal width	petal length	petal width	class
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa

3.3 Hasil Pengujian

Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan dari nilai accuracy, nilai Class Recall serta nilai evaluasi dengan confusion matrix yang akan menghasilkan nilai true positif atau true positif dan true negative atau true negative, lalu dengan membagikan data menjadi 2 yaitu data testing dan data latih yang dibagi menjadi 2 ratio yaitu 80% dan 20% maka hasil yang didapatkan dari confusion matrix metode naïve bayes adalah pada table 2 berikut ini:

Tabel 2. Confussion Matrix

	true Iris-setosa	true Iris-versicolor	true Iris-virginica
Pred.Iris-setosa	50	0	0
Pred.Iris – versicolor	0	48	2
Pred.Iris-Virginica	0	2	18
Class recall	100%	96%	90%

Tabel diatas merupakan hasil dari akurasi yang tertinggi pada pengujian data yang menggunakan cross validation secara linear serta particle weights (forward) dengan mengoptimalkan nilai fitness 1 , maka perbandingan data tersebut dapat dilihat nilai accuracy dari klasifikasi dengan menggunakan nilai cross validation dituangkan pada tabel 3.

Tabel 3. Accuracy

k-fold	Accuracy
2	41.67%
3	81.67%
4	81.75%
5	83.33%
6	83.33%
7	96.64%
8	96.67%
9	96.64%

Dari hasil pengujian menyimpulkan bahwa pada k= 8 dengan menghasilkan nilai akurasi sebesar 96.67% dengan menggunakan metode naïve bayes yang menggunakan algoritma particle weights (forward).

4. KESIMPULAN

Dari pembahasan-pembahasan diatas maka dapat ditarik kesimpulan bahwa Penelitian dengan particle weights (forward) untuk pemilihan atribut pada metode naïve bayes dialkukan agar dapat meningkatkan akurasi prediksi dalam dataset iris. Hasil penelitian untuk nilai akurasi metode naive bayes dengan algoritma particle weight (forward) senilai 96,67%. Dari 150 dataset yang terbagi menjadi 3 kelas dan 4 atribut yang dilakukan dengan ratio 80% dan 20% sebagai data uji dan data latih. Maka dapat disimpulkan bahwa penerapan Teknik optimasi weight mampu meningkatkan nilai akurasi dengan class recall yang rata – rata di atas 90% pada setiap atribut yang ada pada dataset. Berdasarkan proses pengujian dan kesimpulan yang telah dilakukan, agar penelitian ini dapat ditingkatkan dimukakan saran – saran yang diusulkan yaitu: Penelitian ini diharapkan mampu digunakan sebagai perkembangan agar melahirkan metode – metode baru yang lebih meningkatkan akurasi dengan menggunakan metode dan algoritma yang lain. Menambahkan jumlah data yang lebih besar lagi dan atribut yang lebih banyak sehingga hasil pengukuran yang akan didapatkan lebih baik lagi. Menggunakan metode optimasi lainnya seperti Genetik Algorthm, Bagging dan metode optimasi lainnya. Melakukan pengembangan dengan menggunakan metode feature selection yang lain seperti forward selection, genetic algorthm dan metode feature selection lainnya untuk menyeleksi atribut yang berpengaruh kuat, sehingga atribut yang dipakai hanya sedikit namun tidak mengurangi akurasi dari algoritma yang digunakan. Penelitian ini dapat dikembangkan lagi dengan membandingkan algoritma data mining lainnya misal supprot vector machine dengan mengoptimalkan parameter menggunakan genetic algrithm atau dengan menggunakan pangujian model yang sama untuk dataset public sebagai data sekunder dan data hasil riset sebagai data primer.

REFERENCES

[1] A. Mabrur and R. Lubis, "Penerapan Data Mining untuk Memprediksi Kriteria Nasabah," Jurnal Komputer dan Informatika (KOMPUTA), Vols. 1, No.1, pp. 53-57, 2012.

[2] S. Mujiasih, "Pemanfaatan Data Mining Untuk Prakiraan Cuaca," Jurnal Meteorologi dan Geofisika, vol. 12, no. No.2, pp. 189-195, 2017.

[3] U. F. A. W. Service, in Definition of Terms and Phrases, http://www.fws.gov/stand/devterms.html, , 8 February 2013.

[4] M. Bramer, in Principles of Data Mining. , London, 2007.

[5] Bustami, "Penerapan Naive Bayes Classifier," Jurnal Informatika, vol. VIII, pp. 1-5, 2014.

[6] E. Mutiara, "ALGORITMA KLASIFIKASI NAIVE BAYES BERBASIS PARTICLE SWARM," JURNAL SWABUMI, vol. 8, no. No.1, pp. 46-58, 2020.

[7] H. Muhammad, A. C. Prasojo, S. A. Nur, L. Surtiningsih and I. Cholissodin, "OPTIMASI NAÏVE BAYES CLASSIFIER DENGAN MENGGUNAKAN PARTICLE," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) , vol. 4, no. No.3, pp. 180-184, 2017.

[8] Y. Verawati and M. Said Hasibuan, "Perbandingan Data Set IRIS Dengan Aplikasi Rapid," Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat , Vols. ISSN: 2598-0256, E-ISSN: 2598-0238, pp. 158-163, 2021.

[9] C.Luo, Z. J, X.Xue, W. L, R.Ren and Y. Q, "Cosine normalization: Using cosine similarity instead of dot product," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11139 LNCS, no. doi: 10.1007/978-3-030-01418-6_38., pp. 382-391, 2018.

[10] D. Nasution, H.H.Khotimah and C. N, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan," Comput. Eng. Sci. Syst. J, vol. 4 No.1, no. doi: 10.24114/cess.v4i1.11458, p. 78, 2019.

[11] J. Lieng, I. Kencana and T. Oka, "Analisis Sentimen Menggunakan Metode Naive Bayes Classifier dengan Seleksi Fitur Chi Square," Jurnal Matematika , vol. 3, pp. 92-99, 2014.

[12] M. Ridwan, H. Suyono and M. Sarosa, "Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Jurnal EECCIS, vol. 1 No.7, pp. 59-64, 2013.

[13] S. Pattekari and A. Parveen, "Prediction System for Heart Disease Using Naive," International Journal of Advanced Computer and Mathematical Sciences, vol. 3 No.3, no. ISSN 2230-9624, pp. 290-294, 2012.

[14] M. Rizal, Z. M. Syahaf, S. Rully Priyambodo and Y. Ramadhani, "Optimasi algoritma Naive Bayes Menggunakan Forward Selection untuk klasifikasi penyakit ginjal kronis," NARATIF : Jurnal Ilmiah Nasional Riset Aplikasi dan Teknik Informatika, vol. 05 No.1, no. P-ISSN: 2656-7377 || E-ISSN: 2714-8467, Juni 2023.

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

[16] N. L and A. H, "Perancangan Sistem Pendukung Keputusan Untuk Proses Kenaikan Jabatan," Semin. Nas. Teknol. Inf. dan Multimed, pp. 6-7, 2016.

[17] T. A, "Classification assessment methods," Appl. Comput. Informatics, no. doi: 10.1109/ICPR.2010.764, 2018.

[18] H. T, S.Rosset, Z. J and Z. H, "Multi-class AdaBoost," Stat. Interface, vol. 2 No.3, no. doi: 10.4310/sii.2009.v2.n3.a8, pp. 349-360, 2009.

[19] "The advantages of the Matthews correlation coefficient(MCC) over F1 score and Accuracy in binary classification evaluation," BMC Genomics, vol. 21 No.1, no. doi: 10.1186/s12864-019-6413-7, pp. 1-13, 2020.

[20] H. Azis, Purnawansyah, F. Fattah and I. Pratiwi Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data," ILKOM Jurnal Ilmiah , vol. 12 No.2, no. E-ISSN 2548-7779, pp. 81-86, 2020.

[21] Fisher, "https://archive.ics.uci.edu/," 1936. [Online]. Available: https://archive.ics.uci.edu/dataset/53/iris. [Accessed 28 Oktober 2023].