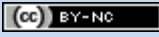


Analisis performa metode *Gaussian Naïve Bayes* untuk klasifikasi citra tulisan tangan karakter arab

Nurul A'yunnisa^{a,1}, Yulita Salim^{a,2}, Huzain Azis^{a,3}

^a Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM. 5, Makassar dan 90231, Indonesia

¹ nurulaayunnisa.labfik@umi.ac.id; ² yulita.salim@umi.ac.id; ³ huzain.azis@umi.ac.id;

INFORMASI ARTIKEL	ABSTRAK
Diterima : 09-10-2022 Direvisi : 13-11-2022 Diterbitkan : 31-12-2022 Kata Kunci: <i>Gaussian Naïve Bayes</i> Klasifikasi Analisis Performa Akurasi Presisi Recall F-Measure	Klasifikasi objek pada gambar secara umum menjadi salah satu masalah dalam visi komputer, bagaimana sebuah komputer dapat mencontoh kemampuan manusia dalam memahami informasi gambar, mengenali objek layaknya manusia, seperti mengenali tulisan tangan atau mengenali pola tertentu pada sebuah gambar. Setiap orang memiliki cara penulisan dan tulisan tangan yang unik dan tidak selaras satu sama lain. Pada artikel ini peneliti mencoba menggunakan dataset tulisan tangan karakter arab. Penelitian ini bertujuan untuk menghitung performa metode (akurasi, presisi, <i>recall</i>, dan <i>f-measure</i>) <i>Gaussian Naïve Bayes</i>. Berdasarkan hasil perhitungan performa menunjukkan tingkat akurasi tertinggi sebesar 12%, presisi 10%, <i>recall</i> 12%, dan <i>f-measure</i> 8%.  

I. Pendahuluan

Tulisan tangan adalah hasil menulis, dengan tangan (bukan ketikan) [1]. Setiap orang memiliki cara penulisan dan tulisan tangan yang unik dan tidak selaras satu sama lain. Penulisan huruf atau karakter Arab pun memiliki perbedaan dengan penulisan bahasa lain. Hal ini terlihat pada tingkat kesulitan dan prosedur tertulis [2]. Bahasa Arab adalah jenis bahasa Semit yang digunakan di negara-negara Timur Tengah sebagai bahasa ibu oleh jutaan orang. Secara umum, alfabet Arab terdiri dari dua puluh delapan karakter alfabet [3]. Pengenalan huruf-huruf arab atau huruf *hijayah* tersebut memerlukan proses klasifikasi.

Klasifikasi objek pada gambar secara umum menjadi salah satu masalah dalam visi komputer, bagaimana sebuah komputer dapat mencontoh kemampuan manusia dalam memahami informasi gambar, mengenali objek layaknya manusia, seperti mengenali tulisan tangan atau mengenali pola tertentu pada sebuah gambar. Bagi manusia hal ini menjadi pekerjaan yang sangatlah sederhana dan mudah, tetapi pada kenyataannya menjadi pekerjaan yang sukar bagi komputer, karena komputer hanya melihat nilai piksel dan data piksel sehingga sulit untuk diproses. Apalagi dengan berbagai variasi dari gambar sangat mempengaruhi pelatihan sehingga untuk mendapatkan hasil yang baik menjadi lebih sulit dan mempengaruhi akurasi. Diharapkan komputer dapat melakukan pengenalan objek layaknya otak biologis manusia walaupun dengan bentuk dan cara kerja yang berbeda [4]. Proses mengenali pola dan gaya tulisan manusia yang berbeda, penulis mencoba menerapkan metode *Gaussian Naïve Bayes* [2][5].

Teorema Bayes, yang juga dikenal sebagai aturan Bayes, adalah alat yang berguna untuk menghitung probabilitas bersyarat. Peluang bersyarat dari A ketika B dilambangkan dengan $P(A|B)$. Distribusi gaussian adalah salah satu metode paling umum dan penting dalam menghitung probabilitas dan statistik [6]. Data pelatihan dipisahkan berdasarkan kelas, kemudian mean dan standar deviasi dari setiap kelas akan dihitung [7].

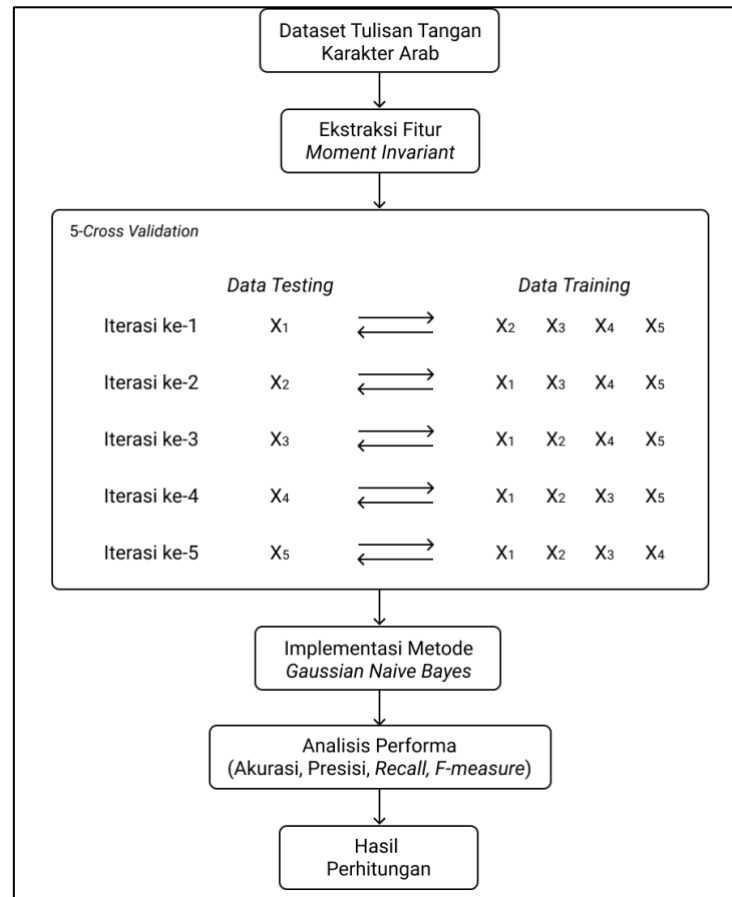
Sebuah penelitian membahas tentang perbandingan tingkat akurasi dari metode *Artificial Neural Network* (ANN) dan *Gaussian Naïve Bayes* (GNB) pada pengenalan angka tulisan tangan [8][9][6]. Penelitian tersebut, diketahui bahwa tingkat akurasi metode GNB lebih tinggi yaitu 28,33% dan metode ANN sebesar 11,67% dengan jumlah data yang digunakan yaitu 200 data.

Berdasarkan penelitian yang telah dipaparkan sebelumnya, peneliti mencoba mengangkat kembali metode yang diterapkan dengan menggunakan dataset yang berbeda. Dataset yang digunakan adalah tulisan tangan karakter Arab, terdiri dari 16.800 citra yang ditulis oleh 60 orang, memiliki rentang usia antara 19 sampai 40 tahun, dan 90% dari 60 orang tersebut merupakan pengguna tangan kanan. Setiap orang menulis sepuluh kali dari masing-masing karakter (dari alif sampai ya). Hasil tulisan tangan tersebut dipindai pada resolusi 300 dpi.

Adapun dataset tersebut telah disitasi oleh dua artikel internasional yaitu artikel pada *International Conference on Information and Communication Systems (ICICS)* dan *Springer International Publishing*.

II. Metode

Secara garis besar tahapan penelitian yang dilakukan pada penelitian ini yaitu mengidentifikasi masalah, pengumpulan data, perancangan sistem, pengkodean, dan pengambilan kesimpulan. Adapun rincian tahapan penelitian tersebut adalah sebagai berikut:



Gambar 1. Tahapan desain penelitian sistem yang dilakukan

Berdasarkan [Gambar 1](#), tahap-tahap dari desain penelitian sistem yang dilakukan adalah sebagai berikut:

1. Tahap pertama, menyiapkan dataset citra tulisan tangan karakter arab yang telah tersegmentasi.
2. Tahap kedua, dilakukan proses ekstraksi fitur dengan menggunakan ekstraksi fitur *moment invariant*. Pada tahap ini dilakukan proses konversi data citra menjadi data numerik. Hasil konversi tersebut berupa nilai array yang diberi label H1 sampai H7 dan Target, kemudian diekspor dalam file format .csv.
3. Tahap ketiga, dataset dibagi menjadi *data training* dan *data testing*. Dilakukan pengulangan sebanyak 5 (lima) kali sesuai dengan *k-fold cross validation* yang telah ditentukan. Proses validasi ini dilakukan dengan cara membagi data menjadi 5 bagian dengan cara melakukan pengacakan data. Setiap pengulangan terdiri dari 80% *data training* dan 20% *data testing*, dimana data akan bergiliran menjadi *data testing*.
4. Tahap keempat, dilakukan perhitungan nilai rata-rata dari *data training*, perhitungan standar deviasi, dan perhitungan nilai *gaussian naive bayes*. Adapun rumus yang digunakan terdapat pada persamaan (5), (6), dan (7).
5. Tahap terakhir, hasil perhitungan performa berupa akurasi, presisi, *recall*, dan *f-measure* dan melihat seberapa besar nilai yang diperoleh.

A. Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan [10]. Data Mining digunakan untuk ekstraksi informasi penting yang tersembunyi dari dataset yang besar. Dengan adanya data mining maka akan didapatkan suatu permata berupa pengetahuan di dalam kumpulan data-data yang banyak jumlahnya [11].

Klasifikasi merupakan fungsi data mining yang menetapkan item dalam koleksi ke kategori atau kelas target. Tujuan klasifikasi adalah untuk memprediksi kelas target secara akurat untuk setiap kasus dalam data [12].

B. Moment Invariant

Moment Invariant merupakan salah satu metode ekstraksi ciri bentuk yang nilainya tidak berubah terhadap perlakuan rotasi, translasi, pencerminan, dan penskalaan. Pada metode ini dihasilkan tujuh nilai moment yang dapat menggambarkan suatu objek berdasarkan posisi, orientasi dan parameter-parameter lainnya.

Sekelompok momen merepresentasikan karakteristik global dari bentuk citra dan menyediakan informasi tipe-tipe geometri citra. *Moment invariant* telah diperkenalkan oleh Hu pada tahun 1961, Hu memperkenalkan *moment invariant* untuk citra digital dengan ukuran $M \times N$ piksel, dihitung dengan menggunakan persamaan:

$$m_{pq} = \sum_{x=1}^M \sum_{y=1}^N x^p y^q f(x, y) \quad (1)$$

dengan $f(x, y)$ merupakan nilai piksel pada koordinat (x, y) .

Invarian translasi dapat dihitung dengan menggunakan *central moment* yang didefinisikan dengan persamaan:

$$\mu_{pq} = \sum_{x=1}^M \sum_{y=1}^N (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (2)$$

dengan $\bar{x} = \frac{m_{10}}{m_{00}}$, dan $\bar{y} = \frac{m_{01}}{m_{00}}$.

Central moment yang dinormalisasi didefinisikan dengan persamaan:

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^\lambda} \quad (3)$$

dengan $\lambda = \frac{(i+j)}{2} + 1$.

Berdasarkan momen ternormalisasi di atas, Hu memperkenalkan tujuh invarian yang diberikan dalam persamaan [13]:

$$\begin{aligned} M_1 &= \eta_{20} + \eta_{02} \\ M_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ M_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ M_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ M_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ M_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} - \eta_{12})(\eta_{21} + \eta_{03}) \\ M_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} + 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (4)$$

C. Gaussian Naïve Bayes

Teorema Bayes, juga dikenal sebagai aturan Bayes, adalah alat yang berguna untuk menghitung probabilitas bersyarat. Peluang bersyarat dari A ketika B dilambangkan dengan $P(A | B)$. Distribusi Gaussian

adalah salah satu metode paling umum dan penting dalam menghitung probabilitas dan statistik. Distribusi Gaussian adalah:

$$P(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}} \quad (5)$$

Dimana, μ adalah rata-rata δ adalah standar deviasi. Untuk mendapatkan nilai μ dan δ menggunakan persamaan (6) dan (7) [6]:

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

$$\delta^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n-1} \quad (7)$$

D. Cross Validation

Cross validation atau validasi silang adalah metode statistik yang digunakan untuk mengevaluasi dan membandingkan sekumpulan data dengan membagi data menjadi dua bagian, yaitu data latih dan data uji. 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 cross validation* untuk menghilangkan bias pada data. Pelatihan dan pengujian dilakukan sebanyak k kali [14].

E. Confusion Matrix

Confusion matrix merupakan visualisasi untuk mengevaluasi dari kinerja model klasifikasi. Untuk melakukan klasifikasi evaluasi komparatif, maka dalam penelitian ini menggunakan *confusion matrix*. *Confusion matrix* ini meliputi informasi tentang kelas yang sebenarnya dan kelas prediksi. Hal ini akan ditemukan pada kolom matriks yang mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kejadian pada kelas tersebut. *Confusion matrix* adalah salah satu alat ukur berbentuk matriks 2x2 yang digunakan untuk memaparkan jumlah ketepatan algoritma yang dipakai. *Confusion matrix* disajikan pada Tabel 1. di bawah ini [15].

Tabel 1. *Confusion Matrix*

		Actual Class	
		Class = Yes	Class = No
Class Predicted	Class = Yes	TP (True Positive)	FP (False Positive)
	Class = No	FN (False Negative)	TN (True Negative)

Keterangan:

True Positive (TP) : Jika data yang diprediksi bernilai positif dan sesuai dengan nilai aktual (positif).

False Positive (FP) : Jika yang di tidak sesuai dengan nilai aktual.

False Negative (FN) : Jika yang diprediksi bernilai negatif dan aktualnya positif.

True Negative (TN) : Jika benar antara prediksi negatif dan kenyataannya negatif.

Untuk mengukur *performance* dari hasil *data mining* menggunakan akurasi, presisi, *recall*, dan *f-measure*, adapun rumusnya bisa dilihat di bawah ini [15]:

$$\text{Akurasi} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (8)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

$$F\text{-measure} = 2 \frac{\text{presisi} \times \text{recall}}{\text{presisi} + \text{recall}} \quad (11)$$

III. Hasil dan Pembahasan

Berdasarkan penelitian yang telah dilakukan, hasil penelitian yang diperoleh adalah sebagai berikut:

A. Hasil Pengimplementasian Kestraksi Fitur

Hasil dari pengimplementasian ekstraksi fitur moment invariant pada dataset citra tulisan tangan karakter arab ditunjukkan pada Tabel 2.

Tabel 2. Hasil implementasi ekstraksi fitur *moment invariant*

Id	H1	H2	H3	H4	H5	H6	H7	Target
0	1,80E-03	2,44E+10	1,28E+07	5,31E+05	4,33E-04	7,67E+02	5,71E-06	1
1	1,25E-03	1,06E+09	1,53E+06	4,57E+04	3,80E-06	4,55E+02	-3,77E-06	1
2	1,87E-03	3,01E+10	5,60E+04	1,64E+05	3,88E-07	1,40E+01	-3,13E-07	1
3	1,81E-03	2,76E+09	2,91E+05	1,74E+06	3,91E-05	2,85E+03	-4,21E-06	1
4	1,78E-03	2,45E+09	1,77E+06	1,03E+07	1,39E-02	1,58E+04	1,35E-03	1
...	...	...	...	...	...	...	...	...
16795	1,62E-03	1,38E+09	9,57E+05	3,99E+04	1,34E-05	1,03E+01	7,69E-06	28
16796	1,09E-03	9,51E+07	1,91E+05	3,69E+05	2,09E-05	4,29E+00	-2,28E-05	28
16797	1,52E-03	3,21E+09	1,85E+06	2,02E+06	4,64E-04	4,27E+01	1,15E-05	28
16798	1,07E-03	6,80E+07	1,09E+05	1,27E+05	1,49E-07	1,23E+00	1,69E-07	28
16799	1,09E-03	1,86E+09	8,28E+03	1,02E+05	-2,71E-06	-4,38E-01	1,20E-07	28

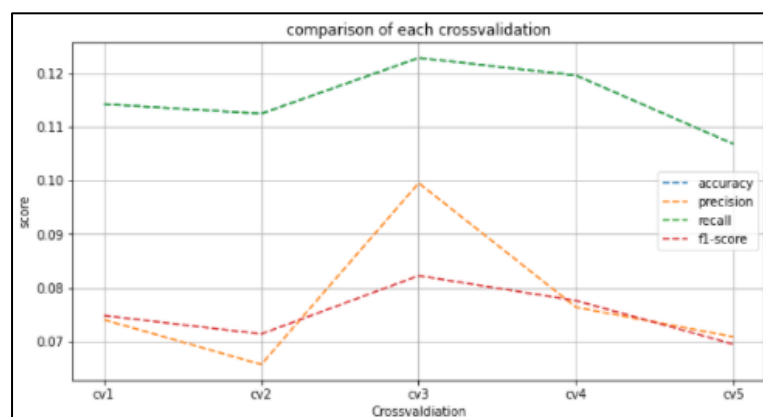
B. Hasil Perhitungan Performa Metode Gaussian Naïve Bayes

Tabel 3. merupakan hasil pengujian performa pada dataset tulisan tangan karakter arab menggunakan metode *Gaussian Naïve Bayes* dengan nilai *k-fold cross validation* 5.

Tabel 3. Hasil perhitungan performa metode *gaussian naïve bayes*

Performa	Cross Validation				
	Fold I	Fold II	Fold III	Fold IV	Fold V
Akurasi	0.11428571	0.1125	0.12291667	0.11964286	0.10684524
Presisi	0.07402946	0.06567434	0.09954625	0.07636267	0.0708444
Recall	0.11428571	0.1125	0.12291667	0.11964286	0.10684524
F-measure	0.07481008	0.07136704	0.08224536	0.07760225	0.06946864

Gambar 2. adalah diagram yang menampilkan perbandingan dari hasil pengujian *cross validation* dengan metode *Gaussian Naïve Bayes*.



Gambar 2. Perbandingan hasil pengujian *cross validation*

IV. Kesimpulan

Perhitungan performa (akurasi, presisi, *recall*, dan *f-measure*) dengan ekstraksi fitur *moment invariant* menggunakan metode *Gaussian Naïve Bayes* dilakukan pada dataset citra tulisan tangan karakter arab yang berjumlah 16.800 data. Berdasarkan hal tersebut menunjukkan nilai yang diperoleh dari perhitungan performa metode *Gaussian Naïve Bayes* pada dataset tulisan tangan karakter arab yaitu akurasi sebesar 12%, presisi 10%, *recall* 12%, dan *f-measure* 8%.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Pembimbing Utama Ibu Ir. Yulita Salim, S.Kom., M.T., MTA dan Pembimbing Pendamping Bapak Ir. Huzain Azis, S.Kom., M.Cs., MTA yang telah memberikan masukan-masukan dalam penyelesaian artikel ini.

Daftar Pustaka

- [1] R. Akbar and E. A. Sarwoko, "Studi Analisis Pengenalan Pola Tulisan Tangan Angka Arabic (Indian) Menggunakan Metode K-Nearest Neighbors dan Connected Component Labeling," vol. 12, no. 2, pp. 45–51, 2016.
- [2] A. Zahriyono, A. Suryan, and M. D. Suliiyo, "Implementasi Pembacaan Huruf Hijaiyyah Dan Karakter Angka Arab Dengan Menggunakan Jaringan Syaraf Tiruan LVQ (Learning Vector Quantization)," Universitas Telkom, 2013.
- [3] A. El-Sawy, M. Loey, and H. El-Bakry, "Arabic Handwritten Characters Recognition using Convolutional Neural Network," *2019 10th Int. Conf. Inf. Commun. Syst. ICICS 2019*, vol. 5, pp. 147–151, 2017, doi: 10.1109/IACS.2019.8809122.
- [4] R. D. Nurfita and G. Ariyanto, "Implementasi Deep Learning Berbasis Tensorflow untuk Pengenalan Sidik Jari," *Emit. J. Tek. Elektro*, vol. 18, no. 01, pp. 22–27, 2018, doi: 10.23917/emitor.v18i01.6236.
- [5] A. A. Mahran, R. K. Hapsari, and H. Nugroho, "Penerapan Naive Bayes Gaussian Pada Klasifikasi Jenis Jamur Berdasarkan Ciri Statistik Orde Pertama," *Netw. Eng. Res. Oper.*, vol. 5, no. 2, p. 91, 2020, doi: 10.21107/nero.v5i2.165.
- [6] Herman *et al.*, "Comparison of Artificial Neural Network and Gaussian Naïve Bayes in Recognition of Hand-Writing Number," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIconCIT 2018*, no. 1, pp. 276–279, 2018, doi: 10.1109/EIconCIT.2018.8878651.
- [7] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *Proc. 5th Int. Eng. Conf. IEC 2019*, pp. 165–170, 2019, doi: 10.1109/IEC47844.2019.8950650.
- [8] S. Mujahidin, B. Prasetyo, and M. C. C. Utomo, "Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode Gaussian naïve bayes," *Voteteknika (Vocational Tek. Elektron. dan Inform.)*, vol. 10, no. 3, p. 17, 2022, doi: 10.24036/voteteknika.v10i3.118299.
- [9] D. H. Sulaksono and A. C. P. Siregar, "Komputasi Penentuan Kualitas pada Fiber Optik Berdasarkan Rugi Daya dengan Gaussian Naive Bayes Menggunakan Teknologi CUDA," *J. IPTEK*, vol. 22, no. 2, pp. 35–42, 2019, doi: 10.31284/j.ipitek.2018.v22i2.322.
- [10] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.
- [11] R. Yanto and R. Khoiriah, "Implementasi Data Mining dengan Metode Algoritma Apriori dalam Menentukan Pola Pembelian Obat," *Creat. Inf. Technol. J.*, vol. 2, no. 2, pp. 102–113, 2015, doi: 10.24076/citec.2015v2i2.41.
- [12] G. Kesavaraj and S. Sukumaran, "A Study On Classification Techniques in Data Mining," 2013.
- [13] N. A. Haryono, W. Hapsari, A. Angesti, and S. Felixiana, "Penggunaan Momen Invariant, Eccentricity, Dan Compactness Untuk Klasifikasi Motif Batik Dengan K-Nearest Neighbour," *J. Inform.*, vol. 11, no. 2, pp. 107–115, 2016, doi: 10.21460/inf.2015.112.411.
- [14] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive

-
- Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
- [15] C. A. Sugianto, “Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk SMAN 1 Gibeber Untuk Siswa Baru Menggunakan Decision Tree,” *J. TEDC*, vol. 9, no. 1, pp. 39–43, 2015, doi: 10.31227/osf.io/vedu7.