

Segmentasi dan Klasifikasi Risiko Perdarahan dengan Algoritma *K-Means* dan *Naive Bayes* Berdasarkan Data Klinis dan Transfusi

Aryanti Aryanti^{1*}, Fadhilah Dwi Wulandari², Muhammad Rafiif³,
Faris Alghaniyyu⁴, Naris Kirana⁵, M. Nawval Alfazri⁶

^{1,2,3,4,5,6}Program Studi Teknik Telekomunikasi, Jurusan Teknik Elektro, Politeknik Negeri Sriwijaya,
Palembang, Sumatra Selatan, Indonesia

e-mail: aryanti@polsri.ac.id^{1*}, dwfadhilah93@gmail.com², muhammadrafiif.id@gmail.com³,
farisalghaniyyu@gmail.com⁴, narisputri13@gmail.com⁵, mnawvalalfazri19@gmail.com⁶

Informasi Artikel

Article History:

Received : 8 Juli 2025
Revised : 17 September 2025
Accepted : 22 Oktober 2025
Published : 29 Oktober 2025

*Korespondensi:

aryanti@polsri.ac.id

Keywords:

Bleeding risk, *K-Means* clustering,
Naive Bayes, clinical data, transfusion
data, risk classification

Hak Cipta ©2025 pada Penulis.
Dipublikasikan oleh Universitas
Dinamika



Artikel ini open access di bawah
lisensi [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/).

 10.37802/joti.v7i2.1124

**Journal of Technology and
Informatics (JoTI)**

P-ISSN 2721-4842

E-ISSN 2686-6102

[https://e-](https://e-journals.dinamika.ac.id/index.php/joti)

[journals.dinamika.ac.id/index.php/joti](https://e-journals.dinamika.ac.id/index.php/joti)

Abstract:

Bleeding risk assessment is essential in clinical decision-making, especially for patients undergoing frequent blood transfusions. This study presents a machine learning approach combining K-Means clustering and Gaussian Naive Bayes classification to assess bleeding risk based on clinical and transfusion history data. Patients were categorised into K-Means clusters, with the ideal number of clusters established by the Elbow Method and Silhouette Score. PCA visualisation demonstrated distinct distinctions among clusters. Cluster 0 contained patients with higher transfusion volume and frequency, showing significantly higher bleeding risk. Subsequently, the Naive Bayes classifier was trained on clinical features to predict bleeding risk and categorized into two risk levels. The model achieved 85.45 percent accuracy on training data and 86.67 percent on testing data, with the highest predictive accuracy observed in Cluster 0 (95.65 percent). These results highlight the potential of combining unsupervised and supervised learning techniques to enhance bleeding risk stratification and support better transfusion management.

PENDAHULUAN

Seiring dengan meningkatnya kasus penyakit menular dan kronis di masyarakat, kebutuhan akan transfusi darah turut mengalami peningkatan. Transfusi darah merupakan prosedur medis penting yang berfungsi sebagai terapi penunjang pada berbagai kondisi, baik hematologis maupun non-hematologis seperti trauma, pembedahan, syok, sepsis, serta gangguan pembentukan sel darah merah seperti talasemia [1]. Transfusi darah diperlukan

untuk menggantikan kehilangan darah, meningkatkan kapasitas oksigenasi, serta memperbaiki proses hemostasis. Namun, terapi ini juga memiliki risiko efek samping seperti reaksi hemolitik, alergi, TRALI, serta potensi transmisi penyakit menular [2].

Dalam praktiknya, transfusi darah tidak boleh dilakukan sembarangan. Untuk mencegah penggunaan sumber daya yang tidak efektif dan menghindari efek negatif terhadap pasien dan sistem pelayanan kesehatan, indikasi klinis harus dipertimbangkan [3]. Oleh karena itu, penting untuk mengembangkan pendekatan prediktif yang dapat membantu mengelompokkan dan mengklasifikasikan pasien berdasarkan tingkat risiko perdarahan, terutama bagi pasien dengan kondisi kronis yang membutuhkan transfusi darah berulang seperti talasemia.

Beberapa penelitian terdahulu telah memanfaatkan metode pembelajaran mesin seperti *K-Means* dan *Naive Bayes* dalam bidang medis [4]. *K-Means* adalah algoritma *clustering non-hierarki* yang sering digunakan untuk mempartisi data ke dalam berbagai kelompok berdasarkan kemiripan karakteristik [5], [6]. Metode ini telah diterapkan untuk segmentasi citra medis seperti MRI guna efisiensi penyimpanan data dan pengelompokan wilayah geografis berdasarkan risiko bencana [7].

Sementara itu, algoritma *Naive Bayes* termasuk dalam supervised learning yang mengklasifikasikan objek berdasarkan probabilitas fitur yang dimiliki [8]. *Naive Bayes* dikenal efisien dan sederhana karena mengasumsikan independensi antar fitur dalam proses klasifikasi [9], [10]. Beberapa studi menggabungkan kedua algoritma ini untuk meningkatkan akurasi prediksi dalam sistem peringatan dini atau diagnosis medis [11]. Di sisi lain, penelitian juga menunjukkan bahwa faktor usia dapat meningkatkan risiko komplikasi seperti retinopati diabetik melalui mekanisme penurunan fungsi endotel dan peningkatan permeabilitas pembuluh darah [12].

Dalam konteks transfusi darah pada pasien talasemia, standar praktik transfusi telah dikembangkan untuk menjaga kadar hemoglobin, menentukan volume darah yang ditransfusikan, serta frekuensi pemberian transfuse [13]. Namun, belum banyak penelitian yang memanfaatkan gabungan *K-Means* dan *Naive Bayes* untuk mengelompokkan pasien berdasarkan data klinis dan histori transfusi guna memprediksi risiko perdarahan secara lebih tepat.

Meskipun telah ada penelitian yang memanfaatkan metode *K-Means* untuk segmentasi citra dan pengelompokan data serta *Naive Bayes* untuk klasifikasi, belum ada studi yang secara khusus menggabungkan kedua metode ini untuk melakukan segmentasi dan klasifikasi risiko perdarahan berdasarkan data klinis dan rekam medis transfusi pasien. Penelitian ini menjadi relevan mengingat pentingnya pendekatan prediktif dalam pelayanan kesehatan yang presisi dan berbasis data. Dengan membangun model yang mampu mengelompokkan pasien menurut tingkat risiko serta mengklasifikasikan kondisi mereka, diharapkan sistem pengambilan keputusan klinis dapat lebih efektif dan responsif terhadap kebutuhan transfusi yang kompleks. Selain itu, pemanfaatan data klinis sebagai variabel utama dalam pengklasifikasian risiko medis juga masih minim diimplementasikan secara sistematis, terutama pada kasus perdarahan. Beberapa penelitian yang ada cenderung fokus pada data citra atau geografis, bukan variabel laboratorium, usia, frekuensi transfusi, dan parameter klinis lainnya [14].

Penelitian ini bertujuan untuk membuat model klasifikasi dan segmentasi risiko perdarahan berdasarkan data klinis dan histori transfusi pasien menggunakan kombinasi algoritma *K-Means* dan *Naive Bayes*. Permasalahan utama yang ingin dipecahkan adalah bagaimana mengelompokkan dan mengklasifikasikan pasien secara otomatis untuk

membantu pengambilan keputusan medis dalam menentukan kebutuhan transfusi secara rasional, aman, dan efisien.

METODE

Studi ini dilakukan dalam dua tahap utama: (1) segmentasi pasien menggunakan algoritma K-Means Clustering dan (2) klasifikasi risiko perdarahan menggunakan algoritma Gaussian Naive Bayes. Tahapan ini bertujuan untuk mengelompokkan pasien berdasarkan karakteristik klinis dan perilaku transfusi, serta memprediksi tingkat risiko berdasarkan data historis.

1. Segmentasi Pasien dengan K-Means Clustering

Tahap pertama penelitian ini adalah segmentasi pasien berdasarkan fitur numerik, yaitu usia, kadar hemoglobin (HGB), jumlah trombosit (PLT), frekuensi transfusi, dan volume darah transfusi. Sebelum dilakukan klusterisasi, data diimputasi menggunakan rata-rata (mean imputation) dan dinormalisasi menggunakan Standard Scaler untuk menghindari pengaruh skala fitur yang berbeda.

Klusterisasi menggunakan algoritma K-Means dengan tujuan meminimalkan jarak kuadrat ke centroid. Dua metode evaluasi digunakan untuk menemukan jumlah kluster yang ideal:

a. Elbow Method

Metode ini memanfaatkan nilai Within-Cluster Sum of Squares (WCSS), yang diperoleh menggunakan Persamaan [15], [16]:

Keterangan:

$$WCSS = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \quad (1)$$

Keterangan:

C_i : kluster ke- i

x_j : data ke- j dalam kluster tersebut

μ_i : centroid kluster C_i

b. Silhouette Score

Indikator ini mengevaluasi tingkat kecocokan setiap data terhadap kluster tempatnya berada, yang dihitung menggunakan persamaan [17], [18]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

Keterangan:

$a(i)$: nilai rata-rata jarak suatu data terhadap anggota lain dalam kelompok yang sama

$b(i)$: nilai jarak rata-rata dari data tersebut terhadap kluster lain yang paling dekat

Visualisasi hasil kluster dilakukan menggunakan Principal Component Analysis (PCA) untuk reduksi dimensi ke dua komponen utama (PC1 dan PC2).

2. Klasifikasi Risiko dengan Gaussian Naive Bayes

Tahap ini bertujuan memprediksi risiko perdarahan sistemik pasien. Sebelum pelatihan model, dilakukan tahapan preprocessing sebagai berikut:

1) Penambahan fitur:

$$\text{transfusion_rate} = \frac{\text{Monetary}}{\text{Frequency} + 1} \quad (3)$$

2) One-hot encoding untuk fitur kategorikal

3) Imputasi nilai hilang (mean)

4) Transformasi distribusi menggunakan PowerTransformer

5) Seleksi fitur dengan variansi rendah menggunakan VarianceThreshold

Model Gaussian Naive Bayes menghitung probabilitas posterior menggunakan Persamaan [19], [20], [21]:

$$P(y | x) = \frac{P(y) \cdot \prod_{i=1}^n P(x_i | y)}{P(x)} \quad (4)$$

Untuk fitur kontinu, digunakan asumsi distribusi Gaussian seperti pada Persamaan [19], [20], [21]:

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \quad (5)$$

Keterangan:

x_i : Nilai fitur ke- i

μ, σ : Rata-rata dan standar deviasi fitur pada kelas y

Tabel 1. Kategori Risiko Perdarahan Berdasarkan Rentang Probabilitas [22]:

Kategori Risiko	Rentang Probabilitas
Low Risk	≤ 0.3
Medium Risk	$0.3 < \text{Probabilitas} \leq 0.7$
High Risk	> 0.7

3. Evaluasi Model

Evaluasi model dilakukan pada data pelatihan dan pengujian dengan metrik berikut [12]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

Keterangan :

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

Model juga dievaluasi per klaster untuk mengamati sensitivitasnya terhadap karakteristik sub-populasi berbeda.

4. Analisis Fitur Berpengaruh

Untuk mengetahui fitur yang paling berkontribusi dalam klasifikasi, digunakan perbedaan rata-rata distribusi Gaussian antar kelas:

$$\text{Feature Importance} = \theta_1 - \theta_0 \quad (10)$$

Keterangan:

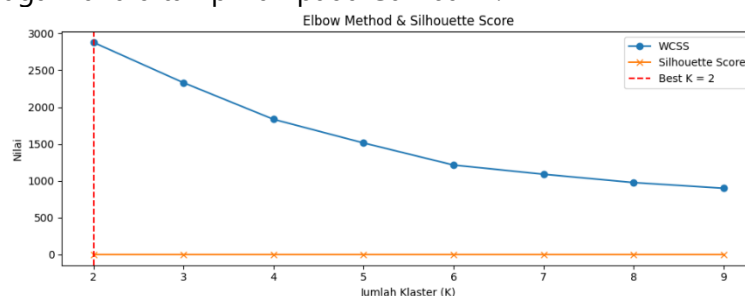
θ_1 : log-mean fitur pada kelas risiko

θ_0 : log-mean fitur pada kelas non-risiko.

HASIL DAN PEMBAHASAN

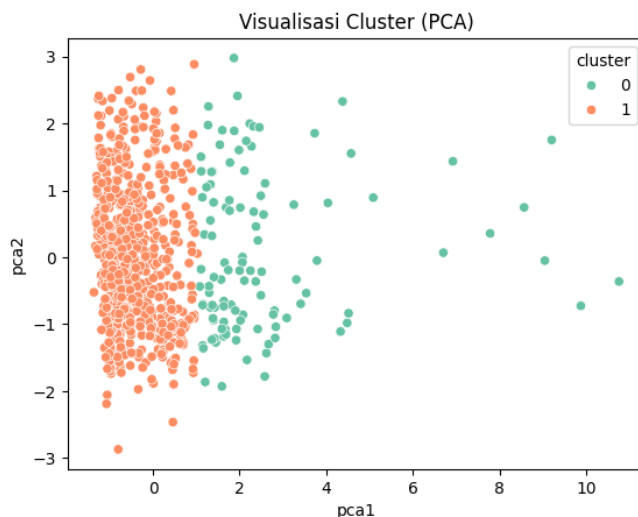
1. Segmentasi Pasien dengan *K-Means*

Proses segmentasi pasien menggunakan metode *K-Means* menghasilkan dua klaster utama berdasarkan data klinis dan riwayat transfusi. Metode Elbow dan nilai Silhouette Score digunakan untuk menentukan jumlah klaster yang ideal. Hasil terbaik diperoleh pada jumlah klaster $K=2$, sebagaimana ditampilkan pada Gambar 1.



Gambar 1. Penentuan Jumlah Klaster Menggunakan Elbow Method dan Silhouette Score

Untuk memvisualisasikan distribusi klaster dalam ruang berdimensi lebih rendah, dilakukan transformasi menggunakan Principal Component Analysis (PCA). Kedua klaster yang dihasilkan memiliki pemisahan yang cukup jelas, seperti yang ditunjukkan oleh hasil visualisasi dua dimensi yang ditunjukkan pada Gambar 2, mengindikasikan bahwa metode *K-Means* mampu mengelompokkan pasien secara efektif berdasarkan fitur klinis dan transfusi.



Gambar 2. Visualisasi Kluster Menggunakan PCA pada Ruang 2 Dimensi

Tabel 2 berikut menunjukkan karakteristik masing-masing kluster:

Tabel 2. Rata-Rata Karakteristik Pasien per Kluster

Klaster	Jumlah Pasien	Usia (tahun)	HGB	PLT	Frekuensi Transfusi	Volume Darah (cc)	Risiko Perdarahan (%)
0	112	54,7	12,5	219	16,0	4011,2	94,6%
1	636	50,3	12,1	216	3,7	915,1	26,7%

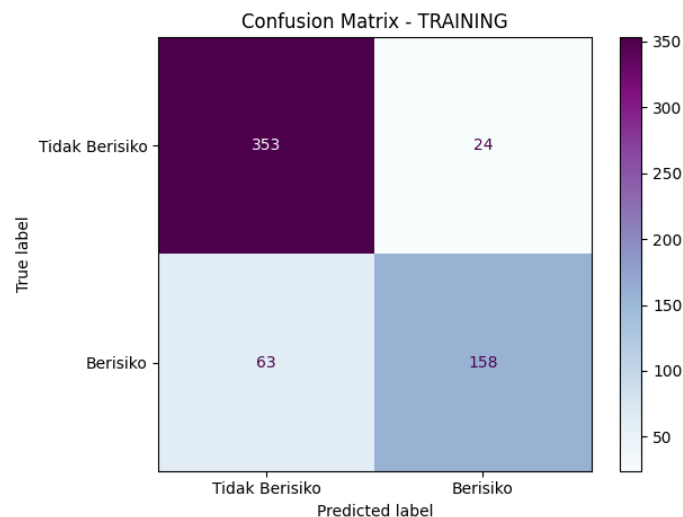
Klaster 0 menunjukkan pasien dengan frekuensi transfusi lebih tinggi dan volume darah yang besar, serta tingkat risiko perdarahan yang sangat tinggi (94,6%). Sebaliknya, Klaster 1 didominasi pasien dengan riwayat transfusi rendah dan tingkat risiko yang jauh lebih kecil (26,7%). Hasil ini yang menyatakan bahwa frekuensi transfusi dan total volume darah merupakan indikator utama dalam penilaian risiko perdarahan.

2. Klasifikasi Risiko Perdarahan dengan *Naive Bayes*

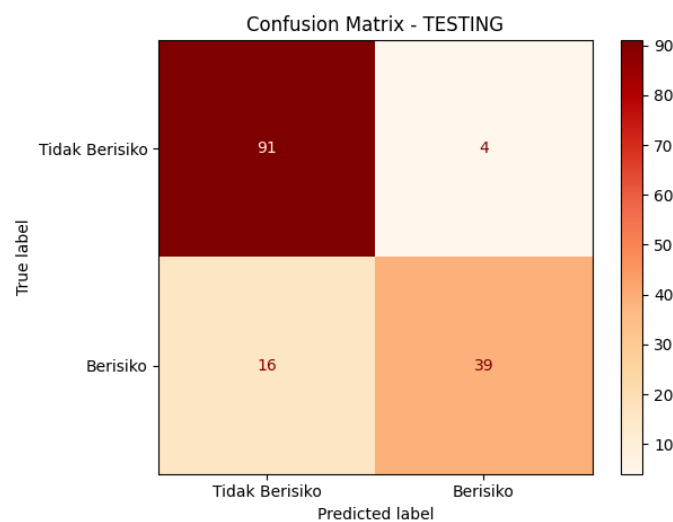
Berdasarkan fitur klinis dan transfusi, model klasifikasi algoritma *Naive Bayes* digunakan untuk memprediksi kelompok risiko perdarahan. Evaluasi model dilakukan pada data pelatihan dan pengujian, dengan hasil ditampilkan pada Tabel 2.

Tabel 3. Evaluasi Kinerja Model *Naive Bayes*

Dataset	Akurasi	Precision	Recall	F1-Score
Training	85,45%	86,81%	71,49%	78,41%
Testing	86,67%	90,70%	70,91%	79,59%



Gambar 3. *Confusion Matrix Training*



Gambar 4. *Confusion Matrix Testing*

Kinerja model menunjukkan akurasi yang tinggi dan stabil antara data pelatihan dan pengujian. Meskipun nilai recall berkisar 70%, yang menunjukkan masih adanya kasus false negative, secara keseluruhan model cukup andal dalam mendeteksi kelompok pasien dengan risiko perdarahan tinggi.

3. Evaluasi Model Berdasarkan Klaster

Untuk mengevaluasi efektivitas segmentasi terhadap kinerja klasifikasi, dilakukan pengujian akurasi model pada masing-masing klaster. Hasil menunjukkan bahwa model lebih akurat dalam memprediksi risiko pada Klaster 0 (risiko tinggi) dengan akurasi sebesar 95,65%, dibandingkan Klaster 1 (risiko rendah) sebesar 89,06%. Hal ini menunjukkan bahwa model lebih sensitif dalam mengenali pola klinis pasien dengan kecenderungan perdarahan sistemik yang tinggi.

KESIMPULAN DAN SARAN

Menurut penelitian ini, kombinasi algoritma *K-Means* dan *Naive Bayes* efektif dalam segmentasi dan klasifikasi risiko perdarahan berdasarkan data klinis dan transfusi. Segmentasi dengan *K-Means* berhasil mengelompokkan pasien ke dalam dua kluster dengan perbedaan signifikan pada frekuensi dan volume transfusi, yang berkorelasi dengan tingkat risiko perdarahan. Klasifikasi menggunakan *Naive Bayes* menunjukkan akurasi tinggi (86,67%) dan performa stabil, terutama dalam mendeteksi kelompok risiko tinggi. Temuan ini menunjukkan bahwa pendekatan berbasis machine learning dapat mendukung pengambilan keputusan klinis yang lebih tepat. Namun, penelitian ini masih terbatas pada jumlah fitur dan model yang digunakan. Oleh karena itu, pengembangan selanjutnya disarankan untuk memperluas variabel klinis, menguji algoritma pembandingan, serta menerapkan validasi pada data real-time agar sistem yang dibangun lebih akurat dan aplikatif dalam praktik medis.

DAFTAR PUSTAKA

- [1] S. Oktavia and A. Qoiriah, "Pengembangan Aplikasi Penentu Kelayakan Calon Pendonor Darah Berbasis Website Menggunakan Algoritma Naïve Bayes," *J. Inform. Comput. Sci. JINACS*, vol. 6, no. 01, pp. 1–10, May 2024, doi: 10.26740/jinacs.v6n01.p1-10.
- [2] W. I. Rahayu, C. Prianto, and E. A. Novia, "Perbandingan algoritma K-Means dan Naive Bayes untuk memprediksi prioritas pembayaran tagihan rumah sakit berdasarkan tingkat kepentingan pada PT. Pertamina (Persero)," *J. Tek. Inform.*, vol. 13, no. 2, pp. 1–8, Apr. 2021.
- [3] "Patient Blood Management and Clinical Use of Blood." Accessed: Oct. 16, 2025. [Online]. Available: <https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/blood-and-products-of-human-origin/patient-blood-management-and-clinical-use-of-blood>
- [4] S. S. Nurhadiva, A. Aryanti, and S. Sarjana, "Blood Pressure and Heart Rate Measurement for Hypertension Classification Using the K-Nearest Neighbors Method Based on IoT," *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 12, no. 2, pp. 373–382, Sept. 2024, doi: 10.33558/piksel.v12i2.9824.
- [5] R. Z. Megantara, P. I. Faot, R. H. Ito, K. F. Annabel, and O. Karnalim, "Pemanfaatan Machine Learning dalam Prediksi Rating: Studi Kasus pada Data Abstrak Publikasi Ilmiah," *J. Technol. Inform. JoTI*, vol. 7, no. 1, pp. 62–75, Apr. 2025, doi: 10.37802/joti.v7i1.999.
- [6] K. Auliasari and M. Kertaningtyas, "Studi komparasi klasifikasi pola tekstur citra digital menggunakan metode K-Means dan Naive Bayes," *J. Inform.*, vol. 18, no. 2, pp. 175–185, Dec. 2018, doi: 10.30873/ji.v18i2.1220.
- [7] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.
- [8] J. Lemantara, "Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil," *J. Technol. Inform. JoTI*, vol. 4, no. 1, pp. 31–40, Oct. 2022, doi: 10.37802/joti.v4i1.265.
- [9] D. Nurmadewi, C. P. Gusti, E. C. Hernanto, F. Ananda, M. A. Hutagalung, and N. Hikmah, "Analisis Sentimen untuk Melihat Respon Masyarakat Terhadap Vaksin Pfizer," *J. Technol. Inform. JoTI*, vol. 4, no. 2, pp. 77–80, Apr. 2023, doi: 10.37802/joti.v4i2.322.
- [10] "View of Article Review: Risk Factors Affecting the Progression of Diabetic Retinopathy in Type 2 Diabetes Mellitus Patients." Accessed: Oct. 16, 2025. [Online]. Available: <https://journalofmedula.com/index.php/medula/article/view/958/755>

- [11]Y. M. Nancy and D. Sumanti, "Latar Belakang Penyakit pada Penggunaan Transfusi Komponen Darah pada Anak," *Sari Pediatri*, vol. 13, no. 3, pp. 159–64, Nov. 2016, doi: 10.14238/sp13.3.2011.159-64.
- [12]D. Purbasari, E. Fadila, and M. N. Imani, "Status Nutrisi dan Hemoglobin Setelah Pemberian Transfusi pada Penderita Thalasemia," *J. Med. Nusantara*, vol. 1, no. 2, pp. 225–239, June 2023, doi: 10.59680/medika.v1i2.427.
- [13]A. K. F. Aidia, P. J. Amelia, and V. R. S. Nastiti, "Prediksi Jumlah Pasien Covid-19 Dengan Menggunakan Klasifikasi Algoritma Machine Learning," *SINTECH Sci. Inf. Technol. J.*, vol. 5, no. 2, pp. 165–172, Oct. 2022, doi: 10.31598/sintechjournal.v5i2.1163.
- [14]11451201604 RILA GUSTIA, "Penerapan Radial Basis Function Neural Network (RBFNN) untuk klasifikasi penyakit kanker serviks," skripsi, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2020. Accessed: Oct. 17, 2025. [Online]. Available: <https://repository.uin-suska.ac.id/24236/>
- [15]J. Cárcamo, A. Cuevas, and L. A. Rodríguez, "On uniqueness of the set of k-means," arXiv.org. Accessed: Oct. 17, 2025. [Online]. Available: <https://arxiv.org/abs/2410.13495v1>
- [16]D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J. Am. Stat. Assoc.*, vol. 105, no. 490, p. 713, June 2010, doi: 10.1198/jasa.2010.tm09415.
- [17]M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, p. 759, June 2021, doi: 10.3390/e23060759.
- [18]M. B. Al- Zoubi and M. A. Rawi, "An Efficient Approach for Computing Silhouette Coefficients," *J. Comput. Sci.*, vol. 4, no. 3, pp. 252–255, Mar. 2008, doi: 10.3844/jcssp.2008.252.255.
- [19]"(PDF) Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in The Detection of DDOS In Cloud Computing," *ResearchGate*, doi: 10.1109/ACCESS.2023.3328951.
- [20]W. Wang, L. Yan, F. Liu, and Y. Li, "Improving Gaussian Naive Bayes classification on imbalanced data through coordinate-based minority feature mining," *PeerJ Comput. Sci.*, vol. 11, p. e3003, July 2025, doi: 10.7717/peerj-cs.3003.
- [21]"(PDF) Human Activity Recognition Using Gaussian Naïve Bayes Algorithm in Smart Home," *ResearchGate*, Aug. 2025, Accessed: Oct. 17, 2025. [Online]. Available: https://www.researchgate.net/publication/345398584_Human_Activity_Recognition_Using_Gaussian_Naive_Bayes_Algorithm_in_Smart_Home
- [22]"Blood transfusion safety." Accessed: Oct. 17, 2025. [Online]. Available: <https://www.who.int/health-topics/blood-transfusion-safety>