

Perbandingan Implementasi Metode K-Nearest Neighbor menggunakan Jarak Euclidean dan Manhattan pada Analisa Klasifikasi Penyakit Anemia

Sekar Buana Prameswary¹ Devany Agustianingsih², Anggit Gusti Nugraheni³
Ilmu Komputer, Ilmu Komputer², Sains Data³
email: sekarbuana04@gmail.com¹

ABSTRAK

Anemia merupakan masalah kesehatan global yang memerlukan penanganan efektif melalui klasifikasi yang akurat. Penelitian ini menggunakan algoritma K-Nearest Neighbor (KNN) untuk menganalisis dan mengklasifikasikan jenis penyakit anemia berdasarkan data hematologi dari 1.281 sampel. Proses klasifikasi melibatkan perhitungan jarak Euclidean dan Manhattan, dengan akurasi masing-masing 60% dan 64,31%. Hasil menunjukkan bahwa metode Manhattan lebih efektif dalam mendeteksi jenis anemia dengan kesalahan prediksi lebih sedikit dibandingkan Euclidean. Model yang dikembangkan diharapkan dapat mendukung diagnosis dini, meningkatkan efisiensi layanan kesehatan, dan memberikan solusi bagi wilayah dengan keterbatasan fasilitas diagnostik. Penelitian ini memberikan kontribusi signifikan dalam penerapan teknologi untuk klasifikasi data medis dan mendukung pengambilan keputusan klinis.

Kata Kunci: Anemia; Machine Learning; K-Nearest Neighbor, Euclidean, Manhattan

ABSTRACT

Anemia is a global health issue that requires effective handling through accurate classification. This study employs the K-Nearest Neighbor (KNN) algorithm to analyze and classify anemia types based on hematological data from 1,281 samples. The classification process involves the calculation of Euclidean and Manhattan distances, yielding accuracies of 60% and 64.31%, respectively. The results indicate that the Manhattan method is more effective in detecting anemia types with fewer prediction errors compared to Euclidean. The developed model is expected to support early diagnosis, enhance healthcare efficiency, and provide solutions for regions with limited diagnostic facilities. This research makes a significant contribution to the application of technology in medical data classification and clinical decision-making support.

Keywords: Anemia; Machine Learning; K-Nearest Neighbor, Euclidean, Manhattan

PENDAHULUAN

Anemia adalah kondisi medis yang umum dan menjadi masalah kesehatan global yang serius. Berdasarkan *World Health Organization* (WHO, 2021), anemia memengaruhi lebih dari 1,62 miliar orang di seluruh dunia, atau sekitar 24,8% populasi global. Anemia ditandai dengan rendahnya kadar hemoglobin dalam darah, yang menyebabkan tubuh mengalami kesulitan mendistribusikan oksigen ke jaringan. Kondisi ini dapat disebabkan oleh berbagai faktor, seperti kekurangan zat besi, gangguan genetik, infeksi kronis, dan gangguan sistemik lainnya. Dampaknya mencakup penurunan produktivitas, gangguan pertumbuhan pada anak-anak, dan peningkatan risiko mortalitas, terutama pada ibu hamil dan anak-anak (Kassebaum et al., 2014).

Anemia dapat diklasifikasikan berdasarkan etiologi atau karakteristik hematologis menjadi anemia mikrositik, normositik, dan makrositik. Identifikasi jenis anemia sangat penting untuk menentukan intervensi yang tepat dan mencegah komplikasi kesehatan yang serius. Namun, proses klasifikasi anemia sering kali membutuhkan pemeriksaan laboratorium yang kompleks, seperti analisis morfologi darah, pengukuran kadar hemoglobin, dan tes diagnostik tambahan. Metode ini memakan waktu, biaya, dan sumber daya yang signifikan, sehingga tidak selalu dapat diakses oleh populasi di daerah dengan fasilitas kesehatan terbatas (Bhutta et al., 2013). Oleh karena itu, diperlukan pendekatan alternatif berbasis teknologi untuk meningkatkan efisiensi dan akurasi klasifikasi anemia.

Salah satu pendekatan yang potensial untuk mengatasi masalah ini adalah penerapan algoritma K-Nearest Neighbor (KNN) dalam klasifikasi anemia. Metode ini telah terbukti efektif dalam berbagai aplikasi klasifikasi data karena kemampuannya dalam mengidentifikasi pola berdasarkan kedekatan data baru dengan data pelatihan. Dalam konteks anemia, data hematologi seperti kadar hemoglobin, hematokrit, dan indeks eritrosit dapat digunakan sebagai parameter untuk membangun model klasifikasi yang akurat. Penelitian sebelumnya menunjukkan bahwa KNN memberikan akurasi tinggi dalam klasifikasi data kesehatan, termasuk dalam aplikasi diagnosis penyakit. Sebagai contoh, penelitian oleh Jasri et al. (2024) melaporkan bahwa KNN mampu mencapai akurasi hingga 92% dalam diagnosis penyakit berbasis data medis.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi anemia menggunakan metode K-Nearest Neighbor (KNN) dengan memanfaatkan data hematologi sebagai input utama. Melalui pengembangan ini, penelitian berupaya mengevaluasi performa model dalam mengklasifikasikan jenis anemia secara akurat, serta memberikan rekomendasi penggunaan model sebagai alat bantu dalam mendukung diagnosis anemia di layanan kesehatan. Penelitian ini juga memanfaatkan dua metode perhitungan jarak dalam algoritma KNN, yaitu jarak Euclidean dan Manhattan. Jarak Euclidean mengukur garis lurus terpendek antara dua titik data, sementara jarak Manhattan menghitung jarak dengan mengikuti jalur grid. Kedua metode ini dipertimbangkan karena masing-masing memiliki keunggulan dalam mengolah data yang berbeda karakteristiknya.

Sejumlah penelitian sebelumnya menunjukkan efektivitas KNN dalam aplikasi medis. Misalnya, penelitian oleh Chen et al. (2022) melaporkan bahwa KNN memberikan akurasi hingga 89% dalam klasifikasi data hematologi pasien dengan penyakit darah. Selain itu, Al-Khazraji et al. (2023) melaporkan bahwa metode ini mencapai tingkat sensitivitas hingga 91% dalam klasifikasi anemia mikrositik dan normositik, menunjukkan potensi besar dalam mendukung proses diagnosis anemia. termasuk untuk menentukan nilai keteladanan pengurus di Pondok Pesantren Nurul Jadid, menghasilkan akurasi tinggi sebesar 97,70% dengan nilai k optimal 4 (Jasri et al., 2024).

Penelitian oleh Suwanda et al. (2020) menganalisis performa Manhattan Distance dan Euclidean Distance dalam algoritma K-Means dengan menggunakan dataset Iris. Hasil penelitian menunjukkan bahwa Manhattan Distance memberikan kinerja yang lebih baik dibandingkan Euclidean Distance, dengan jumlah iterasi yang lebih rendah untuk nilai centroid 3 dan 4, yaitu masing-masing hanya 4 iterasi dibandingkan Euclidean yang memerlukan 6 dan 8 iterasi pada kondisi yang sama

Selain itu, Sharma et al. (2016) melakukan analisis perbandingan antara Manhattan Distance dan Euclidean Distance pada algoritma A*. Hasilnya menunjukkan bahwa Manhattan Distance menghasilkan waktu pencarian yang lebih cepat untuk jarak lintasan yang sama. Pada kasus dengan panjang lintasan 34, Manhattan Distance hanya memerlukan waktu 14 ms dibandingkan Euclidean Distance yang membutuhkan 21 ms. Ini membuktikan bahwa Manhattan Distance lebih efisien dalam menghitung lintasan terpendek berdasarkan parameter waktu eksekusi

Dengan demikian, penelitian ini diharapkan dapat menghasilkan model klasifikasi anemia yang dapat digunakan untuk mendukung diagnosis dini secara cepat, akurat, dan efisien. Penelitian ini juga bertujuan untuk membantu tenaga medis dalam memberikan intervensi yang lebih baik dan meningkatkan akses terhadap layanan kesehatan, khususnya di wilayah dengan keterbatasan fasilitas diagnostik.

METODE

Penelitian yang kami lakukan kali ini adalah analisa klasifikasi penyakit anemia dengan dataset yang digunakan terdiri dari 1.281 sampel data, yang dibagi menjadi data latih dan data uji. Dataset mencakup parameter seperti White Blood Cell (WBC), Lymphocyte Percentage (LYMp), Neutrophil Percentage (NEUTp), Lymphocyte Number (LYMn), Neutrophil Number (NEUTn), Red Blood Cell (RBC), Hemoglobin (HGB), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Platelet Count (PLT), Platelet Distribution Width (PDW), dan persentase retikulosit (7PCT).

Dataset mencakup klasifikasi ke dalam tujuh kategori penyakit, yaitu Healthy (sehat), Iron Deficiency Anemia (anemia defisiensi besi), Leukemia, Macrocytic Anemia, Normocytic Anemia, Other Microcytic Anemia, dan Thrombocytosis. Setiap kategori merepresentasikan kondisi hematologi tertentu yang didiagnosis berdasarkan parameter dalam dataset.

Tabel 1. Data Latih

KODE	WBC	LYMp	NEUTp	LYMn	NEUTn	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	PDW	7PCT	Diagnosis
LT01	8.845	25.811	77.511	1.88076	5.14094	4.95	15.2	46.1526	89.7	30.6	34.2	27.9	14.31251157	0.26028	Healthy
LT02	8	25.845	77.511	1.88076	5.14094	5.1	13.5	46.1526	90	29	32	35.0	17.3	0.26028	Healthy
LT03	7.5	25.845	77.511	1.88076	5.14094	5.4	13.8	46.1526	92	30	32	32.0	13.5	0.26028	Healthy
LT04	8.3	25.845	77.511	1.88076	5.14094	5.2	14.2	46.1526	91	30	33	36.0	12.3	0.26028	Healthy
LT05	9	25.845	77.511	1.88076	5.14094	5.7	15.1	46.1526	96	32	33	38.0	16.2	0.26028	Healthy
LT06	8.2	25.845	77.511	1.88076	5.14094	5.3	14.8	46.1526	91	30	33	36.0	16.3	0.26028	Healthy
LT07	8.1	25.845	77.511	1.88076	5.14094	5.5	14.7	46.1526	93	31	33	36.0	17.7	0.26028	Healthy
LT08	8	25.845	77.511	1.88076	5.14094	5	13.2	46.1526	90	29	32	35.0	14.31251157	0.26028	Healthy
LT09	8.3	25.845	77.511	1.88076	5.14094	5.2	14.2	46.1526	91	30	33	36.0	14.31251157	0.26028	Healthy
LT10	8.2	25.845	77.511	1.88076	5.14094	5.3	14.8	46.1526	91	30	33	36.0	16.1	0.26028	Healthy
...
LT1026	6.4	27.6	63.4	1.8	4	4.66	13.6	41	88.1	29.1	33.1	14.0	13.1	0.12	Thrombocytopenia

Tabel 2. Data Uji

KODE	WBC	LYMp	NEUTp	LYMn	NEUTn	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	PDW	7PCT
UJ01	7.15	25.845	77.511	1.88076	5.14094	5.34	13.1	46.1526	81.3	24.5	30.2	233	14.31251157	0.26028
UJ02	8	28.2	63.8	2.3	5.1	5.36	13	43.3	80.8	24.2	30	173	15.4	0.17
UJ03	8.3	21.9	68.8	1.8	5.7	5.38	15.1	47.8	88.9	28	31.5	153	14.1	0.14
UJ04	8.2	34.5	56.3	2.8	4.6	5.29	14.4	45.2	85.5	27.2	31.8	205	15.1	0.21
UJ05	7.97	25.845	77.511	1.88076	5.14094	5.53	13.9	46.1526	83.5	25.1	30.1	186	14.31251157	0.26028
UJ06	8.27	25.845	77.511	1.88076	5.14094	5.17	14.6	46.1526	91.7	28.2	30.8	251	14.31251157	0.26028
UJ07	5.7	25.845	77.511	1.88076	5.14094	4.14	13.6	46.1526	88.5	33.4	37.6	262	14.31251157	0.26028
UJ08	6.1	25.845	77.511	1.88076	5.14094	4.8	13.8	46.1526	90.4	28.8	31.8	200	14.31251157	0.26028
UJ09	8.1	25.845	77.511	1.88076	5.14094	4.73	13	46.1526	88.3	29.6	33.5	252	14.31251157	0.26028
UJ10	8.4	25.845	77.511	1.88076	5.14094	4.75	13.2	46.1526	89.2	28.9	32.2	157	14.31251157	0.26028
...
UJ255	9.8	16.1	75.6	1.6	7.4	5.34	15.6	47	88.2	29.2	33.1	117	14.1	0.11

Penelitian ini menggunakan algoritma K-Nearest Neighbor (KNN), yang merupakan metode supervised learning untuk klasifikasi berbasis kedekatan antar data. Proses dimulai dengan pra-pemrosesan data untuk memastikan bahwa semua atribut berada dalam skala yang seragam. Dataset dibagi menjadi dua bagian yaitu data latih untuk membangun model dan data uji untuk mengevaluasi model.

Metode KNN bekerja dengan menghitung jarak antara data uji dan data latih menggunakan jarak Euclidean dan Manhattan. Rumus jarak Euclidean dinyatakan sebagai

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

Jarak Manhattan

$$d(x, y) = \left| \sum_{i=1}^n (x_i - y_i) \right|$$

Di mana

$d(x, y)$ adalah jarak antara dua titik x dan y ,

x_i dan y_i adalah nilai atribut ke- i dari masing-masing titik,

n adalah jumlah atribut.

Setelah jarak dihitung, nilai k , yaitu jumlah tetangga terdekat, ditentukan melalui validasi silang. Prediksi dilakukan dengan mengambil kelas mayoritas dari k tetangga terdekat sebagai kelas hasil.

Perhitungan yang dihasilkan memerlukan proses data mining, terutama karena tujuan kami adalah mengimplementasikan metode K-Nearest Neighbor (KNN) untuk menganalisis dan mengklasifikasikan penyakit anemia berdasarkan dataset. Proses data mining membantu kami mendapatkan wawasan dari data, membuat prediksi, dan menghasilkan model klasifikasi yang akurat akan dilakukan evaluasi performa model menggunakan confusion matrix, yaitu tabel yang menyajikan hubungan antara hasil prediksi model dengan label sebenarnya. Komponen dalam confusion matrix adalah

True Positive (TP) untuk kasus positif yang diprediksi dengan benar.

False Positive (FP) untuk kasus negatif yang salah diprediksi sebagai positif.

True Negative (TN) untuk kasus negatif yang diprediksi dengan benar.

False Negative (FN) untuk kasus positif yang salah diprediksi sebagai negatif.

Bentuk umum confusion matrix pada klasifikasi dataset anemia ini adalah tabel matriks $n \times n$, di mana n adalah jumlah kelas. Setiap baris mewakili kelas sebenarnya, dan setiap kolom mewakili kelas prediksi. Dalam penelitian ini, terdapat tujuh kelas sehingga confusion matrix akan berbentuk 7×7 . Untuk perhitungan akurasi, komponen-komponen dari confusion matrix dijumlahkan sebagai berikut

$$\text{Akurasi} = \frac{\sum_{i=1}^n CM(i, i)}{\sum_{i=1}^n \sum_{j=1}^n CM(i, j)}$$

Di mana

$CM(i, i)$ adalah elemen diagonal utama confusion matrix (prediksi benar untuk kelas ke- i),

$\frac{\sum_{i=1}^n CM(i, i)}{\sum_{i=1}^n \sum_{j=1}^n CM(i, j)}$ adalah jumlah seluruh elemen dalam confusion matrix (total data).

Metrik ini memberikan gambaran keseluruhan tentang kemampuan model dalam membuat prediksi yang benar. Evaluasi akurasi akan dilakukan pada data uji untuk menentukan seberapa baik metode K-Nearest Neighbor mengklasifikasikan jenis penyakit anemia.

HASIL DAN PEMBAHASAN

Langkah perhitungan pertama adalah mencari jarak terdekat pada masing-masing data uji dan mengelompokkan dengan jarak lima terdekat karena menggunakan $K=5$. Pada gambar 3 dibawah, merupakan hasil data uji yang kami cantumkan dari perhitungan masing-masing kategori.

Tabel 3. Hasil Pehitungan Dataset Penyakit Anemia menggunakan Jarak Euclidean dan Manhattan

KODE	True Label	Prediction (Euclidean)	Prediction (Manhattan)
UJ01	Healthy	Normocytic hypochromic anemia	Normocytic hypochromic anemia
UJ68	Iron deficiency anemia	Iron deficiency anemia	Iron deficiency anemia
UJ105	Leukemia	Normocytic normochromic anemia	Leukemia
UJ114	Leukemia with thrombocytopenia	Thrombocytopenia	Thrombocytopenia
UJ116	Macrocytic anemia	Healthy	Normocytic hypochromic anemia
UJ119	Normocytic hypochromic anemia	Normocytic hypochromic anemia	Normocytic hypochromic anemia
UJ174	Normocytic normochromic anemia	Normocytic hypochromic anemia	Normocytic hypochromic anemia
UJ227	Other microcytic anemia	Other microcytic anemia	Other microcytic anemia
UJ238	Thrombocytopenia	Thrombocytopenia	Thrombocytopenia

Kami mengambil UJ01 sebagai hasil uji coba hitung tercantum.

Tabel 4. Hasil Pehitungan UJ01 menggunakan Jarak Euclidean

Rank	KODE	Distance	Diagnosis
------	------	----------	-----------

1	LT70	2.64	Healthy
2	LT357	3.9	Iron deficiency anemia
3	LT545	4.6	Normocytic hypochromic anemia
4	LT530	5.45	Normocytic hypochromic anemia
5	LT846	6.39	Normocytic normochromic anemia

Tabel 5. Hasil Pehitungan UJ01 menggunakan Jarak Manhattan

Rank	KODE	Distance	Diagnosis
1	LT70	5.7	Healthy
2	LT357	8.16	Iron deficiency anemia
3	LT545	9.15	Normocytic hypochromic anemia
4	LT530	9.71	Normocytic hypochromic anemia
5	LT846	14.9	Normocytic normochromic anemia

Berdasarkan hasil hitung dari 255 dataset, didapatkan 153 untuk prediksi tepat dan 102 untuk prediksi tidak tepat menggunakan jarak euclidean juga 164 untuk prediksi tepat dan 91 prediksi tidak tepat menggunakan jarak manhattan. Setelah didapatkan hasil tersebut, kami melakukan dengan *data mining* dengan *confusion matrix* menggunakan akurasi dengan diketahui pada Tabel 6 dibawah.

Tabel 6. Tabel Nilai True dan False Akurasi

Metode	TP	TN	FP	FN
Euclidean	26	127	41	61
Manhattan	30	134	37	54

Berikut merupakan hasil perhitungan akurasi Jarak Euclidean

$$\text{Akurasi} = \frac{26 + 127}{26 + 127 + 41 + 61} \times 100\% = 60.0\%$$

Dan hasil perhitungan akurasi Jarak Manhattan

$$\text{Akurasi} = \frac{30 + 134}{30 + 134 + 37 + 54} \times 100\% = 64.31\%$$

Berdasarkan hasil perhitungan, metode K-Nearest Neighbor dengan jarak Manhattan memiliki akurasi lebih tinggi dibandingkan dengan jarak Euclidean, ditunjukkan oleh jumlah True Positives (30 vs. 26) dan True Negatives (134 vs. 127) yang lebih banyak, serta jumlah False Positives (37 vs. 41) dan False Negatives (54 vs. 61) yang lebih sedikit. Perbedaan ini menunjukkan bahwa jarak Manhattan lebih efektif dalam mengklasifikasikan data pada dataset ini, kemungkinan karena sifatnya yang lebih robust terhadap perbedaan skala atau distribusi fitur. Hal ini mengindikasikan bahwa pemilihan metrik jarak yang sesuai sangat penting untuk meningkatkan akurasi algoritma KNN pada kasus tertentu.

SIMPULAN

Berdasarkan hasil perhitungan dengan metode K-Nearest Neighbor menggunakan jarak Euclidean dan Manhattan, metode Manhattan menunjukkan akurasi yang lebih baik (64.31%) dibandingkan Euclidean (60.00%). Untuk metode Euclidean, terdapat 26 True Positives (TP), 127 True Negatives (TN), 41 False Positives (FP), dan 61 False Negatives (FN). Sementara itu, metode Manhattan memiliki 30 True Positives (TP), 134 True Negatives (TN), 37 False Positives (FP), dan 54 False Negatives (FN). Perbedaan ini menunjukkan bahwa Manhattan lebih efektif dalam mendeteksi data positif dan negatif dengan kesalahan prediksi yang lebih sedikit dibandingkan Euclidean. Oleh karena itu, metode Manhattan dapat dianggap lebih andal untuk dataset ini.

REFERENSI

- Bhutta, Z. A., Das, J. K., Rizvi, A., Gaffey, M. F., Walker, N., Horton, S., ... & Black, R. E. (2013). Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost?. *The lancet*, 382(9890), 452-477.
- Hirosawa, T., Harada, Y., Tokumasu, K., Ito, T., Suzuki, T., & Shimizu, T. (2024). Comparative study to evaluate the accuracy of differential diagnosis lists generated by gemini advanced, gemini, and bard for a case report series analysis: cross-sectional study. *JMIR Medical Informatics*, 12, e63010.
- Ilham, F., Sembiring, A., & Siregar, R. (2024). DATA MINING METODE K-NEAREST NEIGHBOUR UNTUK REGRESI DATA PENJUALAN KAIN TEKSTILE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(6), 12878-12885.
- Jasri, M., & Ghozali, A. F. (2024). Klasifikasi Nilai Pengurus Teladan Pondok Pesantren Nurul Jadid Menggunakan Metode K-Nearest Neighbor (KNN). *SMARTICS Journal*, 10(1), 13-20.
- Karima, R. A., & Fatah, Z. (2024). IMPLEMENTASI METODE K-NEAREST NEIGHBOR UNTUK KLASIFIKASI PENYAKIT PARU-PARU PADA ANAK. *Jurnal Ilmiah Multidisiplin Ilmu*, 1(6), 10-17.
- Kassebaum, N. J., Jasrasaria, R., Naghavi, M., Wulf, S. K., Johns, N., Lozano, R., ... & Murray, C. J. (2014). A systematic analysis of global anemia burden from 1990 to 2010. *Blood, the Journal of the American Society of Hematology*, 123(5), 615-624.
- Ogino, J., Wilson, M. L., Hofstra, T. C., & Chan, R. Y. (2024). A Novel Discriminating Tool for Microcytic Anemia in Childhood. *Clinical Pediatrics*, 00099228231221330.
- Sharma, S. K., & Kumar, S. (2016). Comparative analysis of Manhattan and Euclidean distance metrics using A* algorithm. *J. Res. Eng. Appl. Sci*, 1(4), 196-198.
- Suwanda, R., Syahputra, Z., & Zamzami, E. M. (2020, June). Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid K. In *Journal of Physics: Conference Series* (Vol. 1566, No. 1, p. 012058). IOP Publishing.