KOMPARASI ALGORITMA C4.5 DAN *K-NEAREST NEIGHBOR* (K-NN) PADA KLASIFIKASI PENYAKIT ANEMIA

Nur Avia Adenta Sari[*1], Ulfarida Miftakhul Jannah[2], Nurmalitasari[3]
[123]Program Studi Sistem Informasi, Universitas Duta Bangsa Surakarta
*e-mail : [*1] 210101076@mhs.udb.ac.id, [2] 210101084@mhs.udb.ac.id, [3] nurmalitasari@udb.ac.id*

Pada penelitian ini, akan menguji pendekatan algoritma Decision Tree dan K-Nearest Neighbor (K-NN) untuk mengklasifikasikan jenis anemia. Tujuan dari penelitian ini adalah untuk mengidentifikasi algoritma mana yang lebih akurat dan memiliki kinerja yang lebih baik dalam hal uji akurasi, presisi, dan recall. Pendekatan pengujian K-fold cross validation pada algoritma C4.5 mencapai akurasi (99,75%), presisi (100%), dan recall (99,50%) yang paling tinggi. Sementara itu, metode K-Nearest Neighbor (K-NN) memperoleh akurasi 89,20%, presisi 86,51%, dan recall 92,88%. Pendekatan Decision Tree C4.5 lebih unggul dari algoritma K-Nearest Neighbor (K-NN) dalam hal mengkategorikan anemia dengan menggunakan K-fold cross validation.

***Kata Kunci*— Anemia, Algoritma C4.5, Klasifikasi, K-Nearest Neighbor, Komparasi.**

I. PENDAHULUAN

Anemia merupakan kondisi jumlah sel darah merah seseorang berada dibawah tingkat yang dianggap normal. Hal ini dapat terjadi karena kekurangan hemoglobin dalam tubuh yang mengakibatkan sedikit sel darah merah yang terbentuk [1]. Anemia juga menyebabkan berbagai masalah kesehatan, seperti gangguan daya tahan tubuh, fokus, prestasi belajar, tidak bugar, dan produktivitas. Anemia pada remaja putri dapat meningkatkan kemungkinan kematian ketika melahirkan, kelahiran prematur, dan berat badan bayi rendah [2]. Berdasarkan pengaruh anemia, deteksi dini perlu dilakukan untuk mencegah anemia sejak dini [3]. Dalam melakukan deteksi dini dapat dilakukan dengan cara klasifikasi [4].

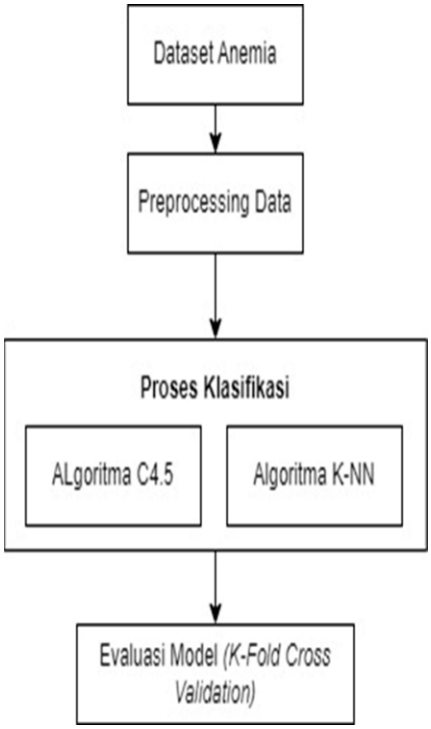
Klasifikasi bertujuan untuk mengenali model dari dataset penyakit anemia menggunakan analisis data pelatihan untuk memprediksi atau mengklasifikasikan seseorang [5]. Beberapa algoritma yang seringkali digunakan untuk mendeteksi dini serta pengklasifikasian adalah Decision Tree C4.5, Support Vector Machine (SVM), dan K-Nearest Neighbour (KNN) [5].

Perbandingan algoritma klasifikasi penyakit anemia telah dilakukan sebelumnya, dan nilai performa algoritma C4.5 dengan akurasi 99,29% menunjukkan bahwa algoritma ini lebih baik dibanding dengan Support Vector Machine (SVM) dalam klasifikasi penyakit anemia [3]. Pada penelitian tentang perbandingan algoritma C4.5 dan K-Nearest Neighbour (K-NN), telah diimplementasikan dengan perolehan bahwa performa yang lebih unggul ditunjukkan oleh algoritma C4.5 dengan akurasi tertinggi sebesar 96,49%, sedangkan nilai akurasi sebesar 94,73% diperoleh oleh K-Nearest Neighbour (K-NN) [6]. Dengan hasil yang ditunjukkan dalam penelitian tersebut, dapat disimpulkan bahwa algoritma C4.5 mengungguli dibandingkan algoritma lainnya [7].

Penelitian ini melakukan perbandingan algoritma Decision Tree dan KNN dalam mengklasifikasikan kategori penyakit anemia. Tujuan penelitian ini untuk menentukan dari dua algoritma yang menghasilkan performa terbaik dalam hal akurasi, presisi, dan recall. Penelitian ini akan menghasilkan klasifikasi dalam dua kelas, yaitu anemia dan tidak anemia.

II. METODE PENELITIAN

Pendekatan penelitian sangat penting untuk mencapai hasil yang konsisten dengan hasil yang direncanakan. Gambar 1 menggambarkan berbagai langkah penelitian.



Gambar 1. Langkah penelitian untuk mencapai tujuan dan hasil pada penelitian.

A. Dataset Anemia

Penelitian ini memanfaatkan set data publik yang didapatkan dari web Kaggle melalui https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset Data tersebut berformat CSV dan berisi informasi mengenai penyakit anemia. Dataset ini memiliki 1.421 entri dengan 5 atribut dan satu label, "Hasil" (0 menunjukkan tidak ada anemia dan 1 menunjukkan anemia). Sebanyak 801 titik data menunjukkan tidak ada anemia, sedangkan 620 menunjukkan anemia. Tabel 1 menunjukkan atribut yang digunakan.

Tabel 1.
Informasi Atribut

Atribut	Deskripsi	Tipe Data	Range
Gender	Jenis Kelamin	Numerik	0=laki-laki, 1=perempuan
Hemoglobin	Kadar hemoglobin	Numerik	6.6 – 16.9
MCH	Parameter Mean Corpuscular Hemoglobin (MCH)	Numerik	16 – 30
MCHC	Parameter dari Mean Corpuscular Hemoglobin Concentration (MCHC)	Numerik	27.8 – 32.5
MCV	Parameter dari Mean Corpuscular Volume (MCV)	Numerik	69.4 – 101.6
Result	Kategori klasifikasi Anemia atau tidak Anemia	Numerik	0=tidak anemia, 1=anemia

B. Preprocessing Data

Preprocessing data melibatkan menghilangkan data duplikat, memeriksa inkonsistensi data, dan memperbaiki kesalahan seperti kesalahan cetak [8]. Penelitian ini menemukan bahwa tidak ada atribut yang perlu dihapus atau dibuang, sehingga digunakan total 5 atribut untuk melakukan klasifikasi. Langkah selanjutnya adalah memeriksa nilai-nilai kosong dari berbagai variabel pada tabel 2.

Tabel 2.
Hasil Variabel yang Bernilai Null

Variabel	Jumlah Null
Gender	0
Hemoglobin	0
MCH	0
MCHC	0
MCV	0
Result	0

Berdasarkan Tabel 2, dapat dapat disimpulkan bahwa tidak ada nilai yang hilang (missing value), menunjukkan bahwa data sudah siap untuk lanjut ke tahap berikutnya.

C. Decision Tree C4.5

Algoritma C4.5 biasa diimplementasikan dalam data mining untuk mengklasifikasikan atau mengelompokkan dengan menggunakan beberapa aturan yang dihasilkan dari pohon keputusan [6]. Beberapa kelebihan dimiliki oleh algoritma C4.5, seperti kemampuan dalam memproses data numerik maupun data diskrit dengan baik, penanganan nilai atribut yang tidak lengkap, penciptaan aturan yang mudah dipahami dan kecepatan eksekusi lebih cepat dibandingkan dengan algoritma lainnya dihasilkan oleh algoritma C4.5 [9]. [10] menyatakan bahwa Algoritma pohon keputusan C4.5 dapat dibuat dalam berbagai langkah, yaitu dengan menentukan atribut sebagai akar, membuat cabang pada setiap nilai, memisahkan setiap kasus di cabang tersebut, lalu ulangi proses tiap-tiap cabang sampai seluruh kasus mempunyai kelas yang sama. Rumus penyelesaian dapat diperoleh di bawah ini :

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\, p \quad (1)$$

$$Gain(S, A) = Entropy\,(S) - \sum_{t=i}^{n} \frac{|S_i|}{|S|} * Entropy\,(S_i) \quad (2)$$

D. K-Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor (K-NN) adalah pendekatan pengelompokan yang belajar dari data yang telah dikategorikan sebelumnya [11]. Algoritma K-NN mempunyai kelebihan dalam memperoleh data yang jelas dan tidak membingungkan, serta sangat efektif digunakan pada dataset yang besar [12]. Proses perhitungan dalam Algoritma k-NN menggunakan Euclidean Distance, yaitu metode untuk mencari jarak antara dua titik variabel dengan mengukur seberapa dekat dan mirip kedua titik tersebut, semakin kecil jarak di antara keduanya semakin dekat dan mirip titik-titik tersebut [13]. Berikut ini rumus algoritma K-Nearest Neighbor (K-NN) [14] :

$$d_i = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Keterangan:
d = Jarak
x = Data Pelatihan
y = Data Pengujian
n = Dimensi data
i = Variabel Data

E. Evaluasi Hasi K-Fold Cross Validation

Teknik K-Fold Cross Validation menghitung tingkat rata-rata sistem dengan menguji sejumlah besar parameter input secara acak [9]. Penelitian ini menggunakan nilai k sebesar 10, selanjutnya data dipisahkan menjadi 10 dengan 1 komponen data latih pada awal dan yang lainnya diuji secara bergantian. Setelah penerapan teknik pengujian *k-fold cross validation,* langkah berikutnya menghitung nilai akurasi presisi dan recall. Performa dari pengelompokan berbasis matriks ditampilkan dalam

sebuah tabel yang disebut confusion matrix, di mana hasil klasifikasi prediktif dikategorikan sebagai True Positive, True Negative, False Positive, dan False Negative [8]. Bentuk confusion matrix untuk klasifikasi 2 kelas disajikan pada gambar 2.

Kelas		Nilai Aktual	
		Negative	***Positive***
Nilai Prediksi	***Negative***	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Gambar 2. Confusion Matrix

Gambar 2 menunjukkan, True Positive (TP) didefinisikan sebagai proporsi data positif yang dikategorikan secara akurat sebagai positif. False Negative (FN) didefinisikan sebagai total data negatif yang secara tidak tepat dilabeli sebagai positif. False Positive (FP) didefinisikan sebagai total data positif secara tidak tepat dilabeli sebagai negatif. True Negative (TN) didefinisikan sebagai total data negatif dengan kategori benar sebagai negatif.

Ukuran evaluasi kinerja dalam klasifikasi didasarkan pada nilai confusion matrix, seperti akurasi, presisi, dan recall.

Rumus untuk menentukan akurasi dapat dilihat pada persamaan (4)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Rumus untuk menentukan presisi dapat dilihat pada persamaan (5)

$$Presisi = \frac{TP}{TP+FP} \quad (5)$$

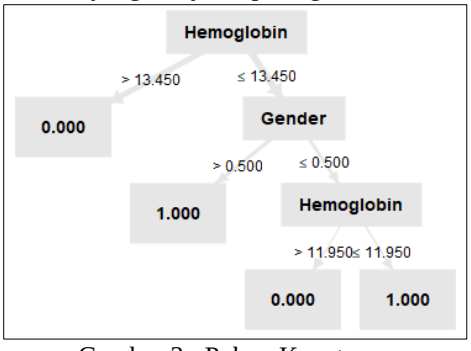
Rumus untuk menentukan recall dapat dilihat pada persamaan (6)

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

III. HASIL DAN PEMBAHASAN

A. Algoritma C4.5

Tahap pertama dalam algoritma C4.5 adalah menentukan gain mana yang menghasilkan keuntungan terbesar dari data. Setelah menjalankan program, ditemukan bahwa gain dengan perolehan terbesar adalah hemoglobin. Hasil pohon keputusan menunjukkan bahwa tingkat hemoglobin yang tinggi mengindikasikan risiko penyakit anemia yang lebih rendah, sedangkan tingkat hemoglobin yang rendah mengindikasikan risiko penyakit anemia yang lebih tinggi. Sebuah pohon keputusan algoritma C4.5 yang disajikan pada gambar 3.



Gambar 3. Pohon Keputusan

Berdasarkan gambar 3, algoritma C4.5 pada pohon keputusan klasifikasi penyakit anemia menghasilkan aturan linguistik sebagai berikut:

- Jika hemoglobin melebihi 13,4, hasilnya adalah 0. Hal ini menunjukkan bahwa kadar hemoglobin melebihi atau sama dengan 13,4, jawabannya adalah tidak.
- Jika kadar hemoglobin kurang dari 13,4 dan jenis kelamin adalah 0, maka hasilnya adalah 1. Hal ini menunjukkan bahwa kadar hemoglobin kurang dari 13,4, jenis kelamin laki-laki, dan kadar hemoglobin kurang dari 11,9, maka jawabannya adalah ya.
- Jika hemoglobin < 13,4 dan jenis kelamin = 0, dan hemoglobin >= 11,9, hasilnya adalah 0. Hal ini menunjukkan bahwa hemoglobin kurang dari 13,4, jenis kelamin adalah laki-laki, dan jika hemoglobin lebih besar atau sama dengan 11,9, jawabannya adalah tidak.
- Jika hemoglobin < 13,4 dan jenis kelamin = 1, hasilnya adalah 1. Hal ini menunjukkan bahwa hemoglobin kurang dari 13,4 dan jenis kelaminnya perempuan, jawabannya adalah ya.

Untuk memperoleh nilai akurasi, presisi, dan recall untuk dataset anemia ditentukan dengan menggunakan teknik pelatihan K-fold cross validation. Tabel 3 menyatakan hasil perhitungan confusion matrix algoritma C4.5.

Tabel 3.
Confusion Matrix

K-fold cross validation			
Kelas		**Nilai Aktual**	
		0	**1**
Nilai Prediksi	**0**	801	0
	1	4	797

Tabel 3 menunjukkan bahwa 797 titik data diidentifikasi secara tepat sebagai kategori anemia menggunakan *K-fold cross validation*. Tidak ada kesalahan data ditemukan untuk diklasifikasikan sebagai anemia yang seharusnya data tersebut termasuk dalam kategori non-anemia. Selain itu, 801 data diidentifikasi dengan tepat sebagai kategori anemia. 4 data yang semestinya berada di kategori anemia tetapi teridentifikasi sebagai kategori non-anemia.

B. Algoritma K-Nearest Neighbor (K-NN)

Model yang diterapkan adalah algoritma K-NN. Confusion matrix dibangun menggunakan pendekatan pelatihan K-fold cross validation, yang menghasilkan nilai akurasi, presisi, dan recall untuk klasifikasi dataset anemia. Tabel 4 menunjukkan hasil perhitungan confusion matrix dengan metode K-NN.

Tabel 4.
Confusion Matrix

K-fold cross validation			
Kelas		**Nilai Aktual**	
		0	**1**
Nilai Prediksi	**0**	685	116
	1	57	744

Dari Tabel 4 menunjukkan bahwa dalam teknik pengujian *K-fold cross validation* menunjukkan bahwa 744 data diperkirakan dengan akurat sebagai kategori anemia. Sebanyak 116 data yang seharusnya berada di kategori anemia tetapi salah diklasifikasikan sebagai non-anemia. Selain itu, 685 data diidentifikasi secara tepat sebagai non-anemia. Sebanyak 57 data diidentifikasi sebagai anemia padahal seharusnya diklasifikasikan sebagai non-anemia.

C. Evaluasi Hasil dan Perbandingan Hasil Kedua Metode

Setelah dilakukan pemodelan dengan algoritma C4.5 dan K-NN, serta penerapan teknik pada pengujian *k-fold cross validation* menghasilkan perbandingan nilai akurasi, presisi, dan recall dari dua algoritma tersebut, seperti yang disediakan tabel 5.

Tabel 5.
Perbandingan Perfoma Algoritma

Algoritma	Akurasi	Presisi	Recall
C4.5	99.75%	100%	99.50%
K-NN	89.20%	86.51%	92.88%

Tabel 5 menunjukkan bahwa algoritma C4.5 unggul dibandingkan algoritma K-NN dalam hal akurasi, presisi, dan recall pada uji K-fold cross validation dengan nilai 99,75%, 100%, dan 99,50% secara berurutan. Algoritma K-NN hanya mencapai akurasi 89,20%, presisi 86,51%, dan recall 92,88%.

IV. KESIMPULAN

Penelitian ini membandingkan dua algoritma dalam mengklasifikasikan anemia. Dengan dataset anemia publik yang diperoleh dari Kaggle, penelitian ini akan menilai metode mana yang lebih akurat dalam mendeteksi anemia tergantung pada jenis kelamin, kadar hemoglobin, dan karakteristik hematologi lainnya.

Hasil penelitian ini menyimpulkan bahwa algoritma C4.5 lebih unggul daripada K-Nearest Neighbor (K-NN) dalam hal akurasi, presisi, dan recall. Dengan menggunakan uji K-fold cross validation, algoritma C4.5 memperoleh akurasi maksimum sebesar 99,75%, presisi 100%, dan recall 99,50%. Sebaliknya, algoritma K-Nearest Neighbor (K-NN) menghasilkan akurasi tertinggi yaitu 89,20%, presisi 86,51%, dan recall 92,88%.

Dari hasil penelitian ini, diambil kesimpulan bahwa algoritma C4.5 dianggap lebih unggul daripada algoritma K-Nearest Neighbor (K-NN) pada pengelompokan penyakit anemia dengan metode yang digunakan yaitu pengujian K-fold cross validation.

DAFTAR PUSTAKA

[1] P. Kemenkes, "Mengenal Gejala Anemia pada Remaja," *ayosehat.kemkes.go.id*, 2023. https://ayosehat.kemkes.go.id/mengenal-gejala-anemia-pada-remaja#:~:text=Anemia adalah suatu kondisi dimana,jumlah produksi sel darah merah. (accessed Jul. 08, 2024).

[2] Kemenkes, "Mengenal Dampak Anemia Pada Remaja," *KemenkesRI*, 2022. https://upk.kemkes.go.id/new/mengenal-dampak-anemia-pada-remaja (accessed Jul. 08, 2024).

[3] D. E. Yanti, L. Framesti, and A. Desiani, "JIP (Jurnal Informatika Polinema) PERBANDINGAN ALGORITMA C4.5 DAN SVM DALAM KLASIFIKASI PENYAKIT ANEMIA," pp. 427–434, 2022, [Online]. Available: https://www.kaggle.com/datasets/biswaranjanrao/an

[4] A. Desiani *et al.*, "Perbandingan Klasifikasi Penyakit Kanker Paru-Paru menggunakan Support Vector Machine dan K-Nearest Neighbor," *J. Process.*, vol. 18, no. 1, pp. 54–62, 2023, doi: 10.33998/processor.2023.18.1.700.

[5] A. Nurmasani and Y. Pristyanto, "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class," *Pseudocode*, vol. 8, no. 1, pp. 21–26, 2021, doi: 10.33369/pseudocode.8.1.21-26.

[6] Fahrurrozi and Wasilah, "Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor (KNN) Dan Decision Tree C-45," *Teknika*, vol. 17, no. 2, pp. 427–434, 2023, [Online]. Available: https://jurnal.polsri.ac.id/index.php/teknika/article/view/7565

[7] Y. F. Wijaya and A. Triayudi, "Perbandingan Algoritma Klasifikasi Data Mining Pada Prediksi Penyakit Diabetes," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 165–174, 2023, doi: 10.47065/josyc.v5i1.4614.

[8] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," *Simkom*, vol. 7, no. 2, pp. 104–110, 2022, doi: 10.51717/simkom.v7i2.96.

[9] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[10] H. Hasanah and Nurmalitasari, "Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C45 dalam Prediksi Penyakit Jantung," *Pros. Semin. Nas. Teknol. dan Sains*, vol. 2, pp. 13–18, 2023.

[11] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[12] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[13] A. Yudhana, S. Sunardi, and A. J. S. Hartanta, "Algoritma K-Nn Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon," *Transmisi*, vol. 22, no. 4, pp. 123–129, 2020, doi: 10.14710/transmisi.22.4.123-129.

[14] S. K. P. Loka and A. Marsal, "Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier untuk Klasifikasi Status Gizi Pada Balita," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 8–14, 2023, doi: 10.57152/malcom.v3i1.474.