JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i3.4326

PEMODELAN KLASIFIKASI ANEMIA APLASTIK MENGGUNAKAN TEKNIK OVERSAMPLING DAN K-NEAREST NEIGHBORS

Fahreza Ananda Kusuma[1*], Budi Prasetyo[2]

[1,2]Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang

Received: 20 April 2024
Accepted: 31 Juli 2024
Published: 7 Agustus 2024

Keywords:
Anemia Aplastik, K-Nearest Neighbor, Machine Learning

Corespondent Email:
fahrezananda@students.unnes.ac.id

Abstrak. Anemia aplastik adalah kondisi medis langka yang ditandai oleh produksi sumsum tulang yang tidak memadai dari sel darah merah, sel darah putih, dan platelet. Kondisi ini dapat menyebabkan berbagai komplikasi serius dan memerlukan pengelolaan yang cermat. Dalam penelitian ini, kami menyelidiki metode klasifikasi untuk mengidentifikasi dan memprediksi keberadaan anemia aplastik berdasarkan profil klinis pasien. Kami menerapkan teknik oversampling menggunakan Synthetic Minority Over-sampling Technique (SMOTE) untuk menangani ketidakseimbangan kelas pada dataset. Selanjutnya, kami menggunakan algoritma K-Nearest Neighbors (KNN) untuk memodelkan klasifikasi. Data yang digunakan dalam penelitian ini terdiri dari berbagai fitur klinis, termasuk usia, jenis kelamin, respons pengobatan, dan parameter hematologi. Hasil eksperimen menunjukkan bahwa penggunaan teknik oversampling bersama dengan KNN dapat menghasilkan model klasifikasi yang efektif untuk anemia aplastik, dengan akurasi mencapai 97.56%. Hasil evaluasi juga menunjukkan nilai F1-Score sebesar 0.9767 dan recall sebesar 0.9545. Temuan ini menunjukkan bahwa pendekatan kami memiliki potensi dalam mendukung diagnosis dan manajemen pasien dengan anemia aplastik.

Abstract. *Aplastic anemia is a rare medical condition characterized by inadequate bone marrow production of red blood cells, white blood cells, and platelets. This condition can lead to various serious complications and requires careful management. In this study, we investigated classification methods to identify and predict the presence of aplastic anemia based on patients' clinical profiles. We applied oversampling techniques using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset. Subsequently, we utilized the K-Nearest Neighbors (KNN) algorithm to model the classification. The data used in this study consisted of various clinical features, including age, gender, treatment response, and hematologic parameters. The experimental results show that the use of oversampling techniques along with KNN can yield an effective classification model for aplastic anemia, with an accuracy of 97.56%. The evaluation results also indicate an F1-Score of 0.9767 and a recall of 0.9545. These findings suggest that our approach has the potential to support the diagnosis and management of patients with aplastic anemia.*

1. PENDAHULUAN

Anemia aplastik merupakan kondisi di mana terjadi penurunan produksi sel-sel darah oleh sumsum tulang yang ditandai dengan pansitopenia (penurunan jumlah semua jenis sel darah di sirkulasi perifer) akibat kelainan

primer pada sumsum tulang dalam bentuk aplasia atau hipoplasia, tanpa adanya infiltrasi, supresi, atau pendesakan sumsum tulang oleh faktor lain, sehingga menyebabkan retikulositopenia, anemia, granulositopenia, monositopenia, dan trombositopenia [1]. Kondisi ini dapat mengancam jiwa dan memerlukan penanganan medis yang tepat. Dalam beberapa tahun terakhir, perkembangan teknologi dan pembelajaran mesin telah memungkinkan penerapan algoritma kecerdasan buatan dalam bidang kedokteran. Salah satu metode yang menjanjikan adalah K-Nearest Neighbors (KNN), yang dapat digunakan untuk menganalisis data klinis dan membantu dalam diagnosis penyakit [2]. Penelitian sebelumnya telah mengeksplorasi penggunaan algoritma ini untuk mengklasifikasikan berbagai kondisi medis berdasarkan data klinis, namun belum banyak yang fokus pada kasus anemia aplastik.

Meskipun algoritma KNN telah menunjukkan potensi dalam aplikasi medis, masih terdapat beberapa tantangan yang perlu diatasi. Salah satu masalah utama adalah ketidakseimbangan kelas dalam dataset, di mana jumlah kasus anemia aplastik (kelas minoritas) jauh lebih sedikit dibandingkan dengan kasus non-anemia aplastik (kelas mayoritas). Kondisi ini dapat menyebabkan bias dalam pelatihan model dan menurunkan akurasi klasifikasi untuk kelas minoritas. Untuk mengatasi masalah ini, teknik oversampling seperti Synthetic Minority Over-sampling Technique (SMOTE) dapat digunakan [3]. SMOTE bekerja dengan membuat data sintetis untuk kelas minoritas, sehingga meningkatkan jumlah sampel dan menyeimbangkan distribusi kelas dalam dataset.

Penelitian ini bertujuan untuk mengeksplorasi kemampuan algoritma KNN dalam mengklasifikasikan anemia aplastik berdasarkan data klinis. Kami juga akan menerapkan teknik oversampling SMOTE untuk mengatasi masalah ketidakseimbangan kelas dalam dataset. Secara khusus, tujuan penelitian ini adalah: (1) Membangun model klasifikasi menggunakan algoritma KNN untuk mendeteksi anemia aplastik berdasarkan data klinis pasien, (2) Menerapkan teknik oversampling SMOTE untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, (3) Mengevaluasi kinerja model dengan metrik evaluasi yang sesuai, seperti akurasi, presisi, recall, dan F1-score. Diharapkan bahwa penelitian ini akan memberikan kontribusi signifikan dalam pengembangan metode diagnosis yang lebih akurat dan efisien untuk anemia aplastik. Hal ini dapat membantu meningkatkan pengelolaan penyakit, meningkatkan kualitas hidup pasien, dan membuka jalan menuju terapi yang lebih tepat dan terpersonalisasi [4].

2. TINJAUAN PUSTAKA

2.1. Preprocessing Data

Preprocessing data merupakan serangkaian tahapan yang dilakukan untuk mempersiapkan data mentah agar siap digunakan dalam analisis atau pemodelan. Langkah-langkah ini mencakup pembersihan data dengan menghilangkan data yang tidak konsisten, mengandung noise, atau duplikat. Selain itu, preprocessing data juga melibatkan penggabungan data dari berbagai sumber menjadi satu kesatuan yang utuh. Selanjutnya, data ditransformasi menjadi format yang sesuai untuk digunakan. Reduksi data juga dilakukan dengan memilih dan mengekstraksi fitur-fitur yang relevan untuk mengurangi kompleksitas. Keseluruhan proses ini bertujuan untuk memastikan data yang akan dianalisis atau dimodelkan telah bersih, konsisten, terintegrasi, dan memiliki format serta fitur yang tepat untuk tugas selanjutnya [5][6].

2.2. K-Nearest Neighbor

KNN atau K-Nearest Neighbors adalah sebuah algoritma klasifikasi populer dalam machine learning dan data mining. Metode ini merupakan pendekatan non-parametrik yang bekerja dengan mencari kelompok k obyek dalam data training yang paling dekat (mirip) dengan obyek baru yang akan diklasifikasikan. Klasifikasi dilakukan dengan menghitung jumlah label kelas terbanyak di antara k tetangga terdekat dari obyek baru tersebut. KNN memiliki kelebihan karena sederhana, mudah diimplementasikan, dan cukup efektif untuk klasifikasi jika data training cukup banyak. Namun di sisi lain, KNN juga memiliki beberapa kelemahan seperti perlu menentukan nilai k yang optimal, komputasi klasifikasi yang mahal jika data training sangat besar, serta

sensitif terhadap fitur yang tidak relevan atau noise. Salah satu kelemahan utama KNN adalah ketergantungannya pada pemilihan nilai k yang tepat dan biaya komputasi tinggi pada saat klasifikasi karena seluruh komputasi dilakukan saat klasifikasi bukan saat pelatihan data (lazy learning) [7].

2.3. Confusion Matrix

Confusion matrix merupakan sebuah metode yang dimanfaatkan untuk menghitung tingkat akurasi dalam konsep data mining. Melalui evaluasi menggunakan confusion matrix, beberapa metrik seperti akurasi, presisi, dan recall dapat dihasilkan. Akurasi dalam klasifikasi merujuk pada persentase ketepatan catatan data yang diklasifikasikan dengan benar setelah pengujian dilakukan terhadap hasil klasifikasi tersebut [8].

3. METODE PENELITIAN

3.1. Pengumpulan Data

Proses pengumpulan data dimulai dengan pembuatan data dummy menggunakan fungsi generate_dummy_data, yang menghasilkan data simulasi dengan berbagai atribut seperti usia, jenis kelamin, respons pengobatan, dan parameter medis lainnya. Setiap atribut memiliki rentang nilai yang ditentukan secara acak.

Selanjutnya, untuk menyimulasikan data yang tidak lengkap, sekitar 20% dari data numerik seperti hemoglobin, jumlah trombosit, jumlah leukosit, dan persentase reticulocyte ditandai sebagai nilai yang hilang (NaN). Ini dilakukan dengan memilih secara acak 20% dari total baris data dan mengubah nilai numeriknya menjadi NaN.

Hasilnya adalah DataFrame Pandas, sebuah pustaka berlisensi BSD dan sumber terbuka, menyediakan struktur data khusus untuk analisis data dalam Python [9]. Isinya mencakup informasi lengkap tentang setiap catatan data, termasuk apakah individu memiliki anemia aplastik atau tidak.

3.2. Prepocessing Data

3.2.1. Label Encoding

Langkah ini dilakukan untuk mengubah nilai-nilai kategorikal dalam fitur seperti Jenis Kelamin dan Respons Pengobatan menjadi nilai numerik [10]. Hal ini diperlukan karena sebagian besar algoritma pembelajaran mesin memerlukan input dalam bentuk numerik. Metode yang digunakan adalah Label Encoding, di mana setiap nilai kategorikal diberi label unik dalam bentuk angka.

3.2.2. Missing Value

Kehilangan nilai dapat menyebabkan penurunan akurasi data dan menurunkan kualitasnya secara keseluruhan, terutama saat dilakukan proses pengolahan data lanjutan seperti klasifikasi [11].

Pada langkah ini, data yang memiliki nilai yang hilang atau NaN diidentifikasi. Ini dilakukan untuk fitur-fitur numerik seperti Hemoglobin, Jumlah Trombosit, Jumlah Leukosit, dan Persentase Reticulocyte. Metode yang digunakan untuk mengisi nilai yang hilang adalah imputasi mean dengan SimpleImputer. SimpleImputer adalah sebuah alat atau fungsi dalam pustaka scikit-learn yang digunakan untuk mengisi nilai yang hilang dalam dataset dengan strategi yang telah ditentukan [12], seperti mean, yang berarti nilai rata-rata dari setiap fitur numerik digunakan untuk menggantikan nilai yang hilang.

3.2.3. Normalisasi

Normalisasi dilakukan untuk memastikan bahwa semua fitur numerik memiliki skala yang seragam [13]. Ini penting karena beberapa algoritma pembelajaran mesin, seperti K-Nearest Neighbors (KNN), sangat sensitif terhadap perbedaan skala. Dalam kode ini, digunakan RobustScaler untuk normalisasi, yang secara efektif menangani pencilan (outliers) dengan menormalkan data berdasarkan median dan rentang interkuartil.

3.2.4. Oversampling

Langkah ini dilakukan untuk menangani ketidakseimbangan kelas dalam dataset. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu kelas jauh lebih sedikit daripada kelas lainnya [14]. Dalam kasus ini, metode SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk menghasilkan sampel sintetis dari kelas minoritas (dalam hal ini, kelas yang memiliki Anemia Aplastik) agar seimbang dengan kelas mayoritas.

3.3. Pemodelan Algoritma KNN

Dalam langkah-langkah pemodelan algoritma K-Nearest Neighbors (KNN), data

resampled dipisahkan menjadi dua bagian utama: data latih dan data uji. Proses ini menggunakan metode train_test_split dengan proporsi 80:20, di mana 80% data digunakan untuk melatih model, sementara 20% data digunakan untuk menguji performanya [15]. Model KNN kemudian dibuat dengan menentukan jumlah tetangga terdekat yang akan dipertimbangkan dalam klasifikasi, dalam contoh ini sebanyak 5 tetangga terdekat. Setelah model dibuat, langkah selanjutnya adalah melatih model tersebut dengan menggunakan data latih. Dengan model yang telah dilatih, prediksi dilakukan pada data uji untuk menentukan label kelasnya.

Evaluasi model dilakukan dengan menghitung akurasi, serta membuat classification report dan confusion matrix untuk memahami performa model dalam mengklasifikasikan data uji yang belum pernah dilihat sebelumnya. Evaluasi ini memberikan wawasan yang penting dalam menilai keandalan model KNN dalam konteks klasifikasi data.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data dummy yang dibuat untuk menyimulasikan data pasien dengan anemia aplastik memiliki total 200 baris data. Setiap baris data mewakili satu entitas pasien dan terdiri dari beberapa atribut yang relevan dengan kondisi medis, seperti usia, jenis kelamin, respons pengobatan, serta parameter laboratorium seperti hemoglobin, jumlah trombosit, jumlah leukosit, dan persentase retikulosit. Pembuatan data dummy dilakukan dengan memperhatikan rentang nilai yang realistis untuk setiap atribut agar representatif dengan situasi medis yang mungkin terjadi. Selain itu, sebagian kecil dari data juga diperkenalkan nilai-nilai yang hilang secara acak untuk merepresentasikan masalah umum dalam pengumpulan data medis. Total 200 baris data tersebut disiapkan sebagai dataset untuk analisis lanjutan dalam pemodelan dan evaluasi algoritma klasifikasi.

4.2. Prepocessing Data

4.2.1. Label Encoding

Terdapat fitur kategorikal "Jenis Kelamin" dengan nilai "Laki-laki" dan "Perempuan", label encoding akan mengubah nilai-nilai ini menjadi 0 dan 1 secara berturut-turut.

	Usia	Jenis Kelamin	Respons Pengobatan
0	32	1	0
1	77	0	0
2	60	0	2
3	39	1	1
4	64	1	1
..	...	...	...
195	39	1	0
196	41	1	1
197	40	1	0
198	51	1	0
199	40	0	1

Gambar 1. Label Encoding untuk Jenis Kelamin dan Respons Pengobatan

4.2.2. Missing Value

Ditemukan nilai yang hilang berjumlah 40 atau 20% dari data pada bagian hemoglobin, jumlah trombosit, jumlah leukosit dan persentase reticulocyte.

	Jumlah Trombosit (/µL)	Jumlah Leukosit (/µL)	\
0	488835.0	NaN	
1	154558.0	NaN	
2	NaN	5058.0	
3	NaN	8641.0	
4	151583.0	2814.0	
..	...	...	
195	232107.0	5228.0	
196	461692.0	4771.0	
197	347161.0	4363.0	
198	309864.0	8619.0	
199	223154.0	NaN	

Gambar 2. Data dengan nilai hilang

Untuk menangani dilakukan metode SimpleImputer dengan mean untuk mengisi nilai yang hilang.

	Jumlah Trombosit (/µL)	Jumlah Leukosit (/µL)
0	488835.000000	6076.505882
1	154558.000000	6076.505882
2	263798.541176	5058.000000
3	263798.541176	8641.000000
4	151583.000000	2814.000000

Gambar 3. SimpleImputer dengan Mean

4.2.3. Normalisasi

Normalisasi data menggunakan RobustScaler bertujuan untuk mengubah skala atau rentang nilai dari fitur-fitur numerik dalam dataset agar memiliki distribusi yang lebih seragam dan konsisten.

Data Setelah Normalisasi:

	Usia	Hemoglobin (g/dL)	Jumlah Trombosit (/µL)
0	-0.561983	-1.256210	1.009146
1	0.925620	-0.221534	-0.489874
2	0.363636	-0.885076	0.000000
3	-0.330579	0.000000	0.000000
4	0.495868	0.000000	-0.503215

Gambar 4. Data Ternormalisasi

4.2.4. Oversampling

Oversampling dilakukan setelah preprocessing data untuk mengatasi ketidakseimbangan kelas pada dataset.

```
Data setelah oversampling:
       Usia  Hemoglobin (g/dL)  Jumlah Trombosit (/µL)
0 -0.561983          -1.256210                1.009146
1  0.925620          -0.221534               -0.489874
2  0.363636          -0.885076                0.000000
3 -0.330579           0.000000                0.000000
4  0.495868           0.000000               -0.503215
```

Gambar 5. Oversampling menggunakan SMOTE

4.3. Pemodelan Algoritma KNN

Setelah dilakukan semuanya, maka selanjutnya adalah menguji data dengan model algoritma KNN. Didapatkan hasil yaitu akurasi 97.56% dengan f1-score 0.9767 dan recall sebesar 0.9545.

```
Accuracy: 0.975609756097561

Classification Report:
              precision    recall  f1-score   support

       False       0.95      1.00      0.97        19
        True       1.00      0.95      0.98        22

    accuracy                           0.98        41
   macro avg       0.97      0.98      0.98        41
weighted avg       0.98      0.98      0.98        41


Confusion Matrix:
[[19  0]
 [ 1 21]]
```

Gambar 6. Hasil Model KNN

5. KESIMPULAN

a. Model K-Nearest Neighbors (KNN) mampu memberikan akurasi yang tinggi dalam memprediksi kelas target, dengan nilai akurasi mencapai 97.56%.
b. Teknik oversampling menggunakan SMOTE efektif dalam menangani ketidakseimbangan kelas dalam data, sehingga meningkatkan performa model dalam memprediksi kelas minoritas.
c. Meskipun demikian, ada kelemahan dalam model KNN, yaitu sensitif terhadap jumlah tetangga (parameter k) yang digunakan. Pemilihan nilai k yang tidak tepat dapat menghasilkan prediksi yang kurang akurat.
d. Pengembangan selanjutnya dapat dilakukan dengan mengeksplorasi model klasifikasi lainnya dan melakukan tuning parameter untuk meningkatkan performa model.

DAFTAR PUSTAKA

[1] T. G. Dharmayuda, S. PD-KHOM, N. M. I. Pratiwi, dan P. N. Tediantini, "ANEMIA APLASTIK".

[2] C.-Y. L. Hafiz Abbad Ur Rehman dan Z. Mushtaq, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease," *Journal of the Chinese Institute of Engineers*, vol. 44, no. 1, hlm. 77–87, 2021, doi: 10.1080/02533839.2020.1831967.

[3] R. Siringoringo, "Klasifikasi data tidak Seimbang menggunakan algoritma SMOTE dan k-nearest neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, 2018.

[4] B. Wiranti, "Urgensi Aspek Psikodermatologi dalam Perawatan Kulit: Memahami Keterkaitan Emosi dan Kesehatan Kulit," *Mutiara: Jurnal Ilmiah Multidisiplin Indonesia*, vol. 2, no. 1, hlm. 224–244, 2024.

[5] I. N. Simbolon, "PREDIKSI KUALITAS AIR SUNGAI DI JAKARTA MENGGUNAKAN KNN YANG DIOPTIMALISASI DENGAN PSO," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, 2024.

[6] L. Wohlrab dan J. Fürnkranz, "A review and comparison of strategies for handling missing values in separate-and-conquer rule learning," *J Intell Inf Syst*, vol. 36, no. 1, hlm. 73–98, 2011, doi: 10.1007/s10844-010-0121-8.

[7] H. and B. D. and B. Y. and G. K. Guo Gongde and Wang, "KNN Model-Based Approach in Classification," dalam *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Z. and S. D. C. Meersman Robert and Tari, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, hlm. 986–996.

[8] P. Mayadewi dan E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," *SESINDO 2015*, vol. 2015, 2015.

[9] A. Putra dan H. Toba, "Pengembangan Gudang Data Pendukung Analisis Tren Penyewaan Peralatan Katering dengan Algoritma Apriori," *Journal of Information System and Technology (JOINT)*, vol. 1, no. 1, hlm. 5–14, 2020.

[10] H. Santoso, R. A. Putri, dan S. Sahbandi, "Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 13, no. 1, hlm. 62–72, 2023.

[11] M. Lutfi dan M. Hasyim, "Penanganan Data Missing Value Pada Kualitas Produksi Jagung Dengan Menggunakan Metode K-Nn Imputation Pada Algoritma C4. 5," *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, vol. 2, no. 2, hlm. 89–104, 2019.

[12] P. Ghadekar, P. Akolkar, D. Anand, P. Oswal, S. Dixit, dan N. Chandak, "Mergers and Acquisitions Prediction using Hybrid-Machine Learning and Deep Learning Approach," dalam *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2022, hlm. 65–70.

[13] D. A. Nasution, H. H. Khotimah, dan N. Chamidah, "Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma K-NN," *Comput. Eng. Sci. Syst. J*, vol. 4, no. 1, hlm. 78, 2019.

[14] L. Qadrini, H. Hikmah, dan M. Megasari, "Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, hlm. 386–391, 2022.

[15] D. Cahyanti, A. Rahmayani, dan S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indonesian Journal of Data and Science*, vol. 1, no. 2, hlm. 39–43, 2020.