Prediksi Penyakit Talasemia Alfa Menggunakan Metode *K-Nearest Neighbors* (KNN)

Muhammad Basil Musyaffa Amin[1], Muhammad Jilan Naufal[2], Muhammad Fajrul Alwan[3], Gibran Hakim[4]

[1,2,3,4]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]basilmusyaffa19@student.ub.ac.id, [2]mjilannaufal@student.ub.ac.id,
[3]fajrulalwan20@student.ub.ac.id, [4]gibranhakim@student.ub.ac.id

Abstrak

Talasemia Alfa adalah penyakit kelainan darah merah yang merupakan salah satu penyebab kematian dan kesakitan di Indonesia. Penyakit ini terjadi akibat kelainan genetik yang menyebabkan kegagalan sintesis rantai globin. Rantai globin merupakan salah satu penyusun hemoglobin sehingga kelainan pada susunan hemoglobin akan mengakibatkan kelainan elastisitas dan lisisnya eritrosit. Kelainan ini bermanifestasi klinis beragam mulai anemia, pucat, lemah, nyeri, kelainan pada tulang hingga ikterus dan hepatosplenomegali. Diagnosis talasemia pada umumnya hanya menggunakan pemeriksaan hematologi lengkap mencakup hitung jumlah eritrosit, kadar hemoglobin, kadar hematokrit, MCV dan MCH, pemeriksaan darah samar serta elektroforesis hemoglobin. Pemeriksaan genetik dalam pendekatan diagnosis berbasis molekuler merupakan pemeriksaan yang akan memberikan hasil perubahan sekuens gen atau mutasi genetik yang akan berpengaruh pada perbedaan gambaran manifestasi klinis dan tingkat keparahan pasien thalasemia. Diagnosis molekuler dapat dilakukan dalam upaya peningkatan pelaksanaan yang efektif dan kualitas hidup pasien. Makalah ini menyajikan sebuah cara memprediksi penyakit talasemia alfa menggunakan metode KNN yang terdiri dari penggunaan pembelajaran mesin untuk memprediksi apakah penyakit talasemia alfa dapat terjadi.

Kata kunci: *Talasemia, KNN, diagnosis molekuler, pemeriksaan genetik*

Prediction of Alpha Thalassemia Disease Using K-Nearest Neighbors (KNN) Method

Abstract

Alpha Thalassemia is a red blood disorder which is one of the causes of death and illness in Indonesia. This disease occurs due to genetic disorder that affect the inability of a person to synthesize globin chains. A globin chain is one of the constituents of hemoglobin so that abnormality of composition of hemoglobin will cause abnormality of elasticity and lysis of the erythrocytes. This disorder has various clinical manifestations ranging from anemia, pale, fatigue, pain, abnormalities in the bone (thalassemic facie) to jaundice and hepatosplenomegaly. Generally, the diagnosis of thalassemia uses a complete blood count including calculating the number of erythrocytes, hemoglobin levels, hematocrit levels, MCV and MCH, blood smear examination and hemoglobin electrophoresis. The genetic examination on molecular-based diagnostic approach is an examination to get a result of changes in gene sequences or genetic mutations that will affect the difference of clinical manifestation and severity of thalassemia patients. Molecular diagnosis can be made in an effort to improve the effective management and the quality of life of patients. This paper presents a method of predicting alpha thalassemia disease Using KNN method which consists of using machine learning to predict whether Alpha Thalassemia disease may occur.

Keywords: *Thalassemia, KNN, genetic examination, molecular diagnosis*

---

1. PENDAHULUAN

Talasemia adalah sindrom gangguan yang diwariskan dan termasuk dalam kategori hemoglobinopati, yaitu kelainan yang disebabkan oleh gangguan dalam sintesis hemoglobin akibat mutasi di dalam atau dekat gen globin. Anak-anak yang mengalami talasemia menunjukkan tanda dan gejala seperti kelemahan, perkembangan fisik yang terhambat, penurunan berat badan, ketergantungan pada transfusi darah untuk bertahan hidup, deformitas wajah, anemia, splenomegali, perubahan tulang wajah, dan hepatomegali.

Dataset Talasemia Alfa adalah sebuah dataset yang efektif untuk mendeteksi pembawa penyakit talasemia alfa dan sangat penting untuk mencegahnya. Terdapat banyak tantangan untuk penyaringan yang efektif, terutama pada penggunaan sumber daya yang rendah. Mempertimbangkan talasemia alfa, pengujian genetik diperlukan untuk mendiagnosa kembali dari

pembawa, dimana hal tersebut mahal dan tidak tersedia secara luas. Dengan menggunakan model pembelajaran mesin (*machine learning*), maka dapat bertindak sebagai alat pendukung keputusan yang mudah diterapkan.

Dataset ini berasal dari database 288 kasus dari Human Genetics Unit (HGU) Fakultas Kedokteran, Kolombo, Sri Lanka. Data ini dikumpulkan dari anak-anak yang membawa talasemia alfa dan anggota keluarga mereka yang terdiagnosa dari tahun 2016 hingga 2020.

Sistem yang menjadi fokus dari dataset ini adalah cara membedakan penanda bahwa seseorang merupakan pembawa atau bukan dari hemoglobin individu tersebut. Terdapat dua kelas dalam dataset ini, yang pertama adalah alpha-carrier, yaitu individu yang positif membawa talasemia alfa. Kelas yang kedua adalah normal, yakni merupakan sebuah kelas yang berisi individu normal yang tidak membawa talasemia alfa.

Terdapat variabel-variabel independen yang terdapat dalam dataset ini yang merupakan variabel dengan angka desimal, diantaranya:

- Konsentrasi Hemoglobin dalam gram per desiliter - g/dL (hb)
- Volume sel darah merah dalam 10^12/L (rbc)
- Volume sel darah merah rata-rata dalam femtoliter - fl (mcv)
- Hemoglobin sel darah merah rata-rata dalam picogram - pg (mch)
- Konsentrasi hemoglobin sel darah merah rata-rata dalam gram per desiliter - g/dL (mchc)
- Lebar sebaran sel darah merah dalam persen - % (rdw)
- Jumlah total sel darah putih dalam 10^6/L (wbc) dengan tipe sel darah putih dalam persen - (neutrofil, limfosit)
- Jumlah total trombosit dalam 10^6/L (plt)
- Persentase Hemoglobin A, A2, dan F dari pengujian HPLC (hba, hba2, hbf)

2. METODE PENELITIAN

2.1. Jenis Dataset

Terdapat banyak variasi data yang dapat digunakan dalam penelitian. Namun, tidak semua jenis data tersebut relevan untuk penelitian yang sedang dilakukan. Oleh karena itu, penting bagi kita untuk memahami jenis-jenis data yang dapat digunakan. Jenis dataset dapat dikelompokkan ke dalam beberapa kategori berdasarkan skala pengukuran yang digunakan. Berdasarkan skala pengukuran, data dapat dibagi menjadi empat jenis, yaitu nominal, ordinal, interval, dan rasio. Data nominal dan ordinal termasuk dalam kategori data kategorikal, sementara data interval dan rasio adalah jenis data numerik.

1. Skala nominal adalah tingkat pengukuran yang paling dasar. Angka-angka dalam skala nominal digunakan untuk mengelompokkan kategori yang berbeda, tetapi tidak untuk mengukur seberapa besar atau kecilnya suatu nilai. Sebagai contoh, ketika mengkodekan jenis kelamin, menggunakan kode 1 untuk laki-laki dan 0 untuk perempuan hanya digunakan untuk membedakan antara kategori laki-laki dan perempuan, tanpa memberikan nilai yang lebih tinggi atau rendah. Jika kode tersebut dibalik, maka maknanya tidak akan berubah.
2. Skala ordinal mirip dengan skala nominal, namun memiliki urutan tingkatan. Pada skala ordinal, angka-angka memiliki nilai yang lebih tinggi atau rendah tergantung pada urutannya. Meski begitu, jarak antara 0 dan 1 pada skala ordinal tidak dapat dijelaskan. Sebagai contoh, tingkat kepuasan dalam analisis kinerja dapat menggunakan skala ordinal: sangat puas (5), puas (4), cukup puas (3), tidak puas (2), dan sangat tidak puas (1). Angka-angka ini memiliki makna bahwa 3 lebih tinggi daripada 2, 4 lebih tinggi daripada 3 dan 2, dan seterusnya. Namun, jarak atau selisih antara 1 dan 2, 2 dan 3, dan seterusnya, tidak memiliki makna tertentu.
3. Pada skala interval (atau skala selang), angka-angka mewakili tingkatan tertentu, dan angka-angka yang berurutan memiliki jarak yang sama. Skala interval memiliki karakteristik bahwa tidak ada titik dasar (nol) yang mutlak, sehingga perbandingan tidak dapat dilakukan. Sebagai contoh, pengukuran suhu dengan menggunakan derajat Celsius (°C) merupakan contoh skala interval. Suhu 40 derajat dan 20 derajat memiliki selisih yang sama dengan suhu 80 derajat dan 60 derajat, yaitu 20 derajat. Namun, suhu 40 derajat tidak berarti dua kali lebih panas daripada suhu 20 derajat. Demikian pula, suhu 0 derajat tidak berarti tidak ada panas sama sekali.
4. Skala rasio adalah tingkat pengukuran yang paling tinggi. Selain dapat membedakan kategori, skala rasio juga menunjukkan tingkatan dan memiliki interval yang sama antara dua nilai yang berurutan. Yang membedakan skala rasio adalah memiliki titik dasar (nol) mutlak, sehingga memungkinkan perbandingan langsung. Contoh penggunaan skala rasio termasuk berat badan, jumlah barang, tinggi produksi, dan sebagainya.

Dataset Talasemia Alfa menggunakan skala rasio pada variabelnya sehingga memiliki tingkat pengukuran yang tinggi untuk membantu keakuratan dari prediksi yang akan penulis lakukan.

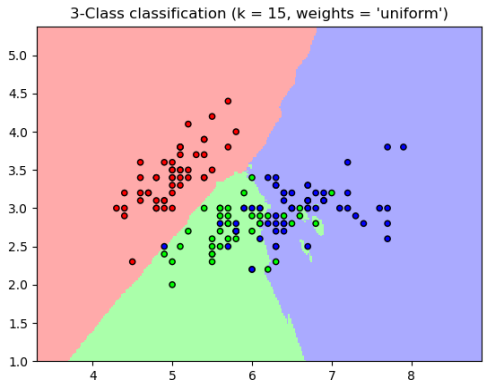
2.2. Pemodelan

Pemodelan yang digunakan pada percobaan kali ini akan menggunakan metode *K-Nearest Neighbors* (KNN). KNN adalah sebuah metode dalam *supervised learning* yang digunakan untuk klasifikasi. Konsep klasifikasi dengan menggunakan KNN melibatkan pencarian tetangga terdekat dari data yang akan diklasifikasikan.

Ide dasar dari KNN adalah mencari K data terdekat dari data yang akan diklasifikasikan berdasarkan jaraknya. Jarak dapat dihitung menggunakan berbagai metrik seperti Euclidean distance atau Manhattan distance. Setelah mendapatkan K data terdekat, mayoritas kelas dari tetangga-tetangga tersebut akan ditetapkan sebagai kelas prediksi untuk data yang akan diklasifikasikan.

Metode KNN memiliki keuntungan dalam kemampuannya untuk bekerja dengan dataset berdimensi tinggi. Namun, KNN memerlukan penghitungan jarak antara data yang akan diklasifikasikan dengan semua data dalam dataset, sehingga dapat menjadi komputasi yang intensif.

Dengan menggunakan metode KNN, kita dapat mencari tetangga terdekat dari data yang akan diklasifikasikan dan menggunakan mayoritas kelas tetangga tersebut sebagai prediksi kelas untuk data tersebut. Metode KNN dapat menjadi alternatif yang efektif dalam pemodelan klasifikasi dengan dataset berdimensi tinggi.



Gambar 1. Contoh Visualisasi KNN

Karakteristik dari metode *K-Nearest Neighbors* (KNN) adalah sebagai berikut:

1. Secara prinsip, KNN merupakan metode klasifikasi yang berbasis pada jarak.
2. KNN tidak menerapkan strategi Structural Risk Minimization (SRM) seperti yang dilakukan oleh SVM.
3. Prinsip kerja KNN tidak terbatas pada klasifikasi dua kelas, tetapi dapat digunakan untuk klasifikasi multikelas.
4. KNN tidak memerlukan proses pelatihan yang kompleks. Pada tahap prediksi atau pengujian, KNN langsung mencari tetangga terdekat dari data yang akan diklasifikasikan.
5. KNN tidak menghasilkan model yang dapat disimpan, tetapi melakukan prediksi berdasarkan mayoritas kelas tetangga terdekat.
6. KNN mampu memisahkan data dengan distribusi kelas baik yang linier maupun non-linier, karena klasifikasi dilakukan berdasarkan jarak dengan tetangga terdekat.
7. KNN tidak dipengaruhi oleh dimensi data yang tinggi, sehingga tidak memerlukan proses reduksi dimensi.
8. Penggunaan memori dalam KNN dipengaruhi oleh jumlah data yang ada, karena KNN harus menyimpan seluruh dataset dalam memori untuk mencari tetangga terdekat. Besarnya dimensi data juga dapat mempengaruhi waktu komputasi yang dibutuhkan.
9. KNN merupakan metode non-parametrik, yang berarti tidak membuat asumsi tertentu tentang distribusi data. Hal ini memungkinkan KNN untuk bekerja dengan baik pada dataset yang memiliki distribusi yang kompleks atau tidak terduga.

2.3 Normalisasi MinMax

Normalisasi Min-Max adalah teknik yang digunakan dalam pengolahan data untuk mengubah nilai-nilai dari fitur-fitur dalam dataset ke dalam rentang yang ditentukan. Tujuan normalisasi Min-Max adalah untuk membawa semua nilai dalam dataset ke dalam rentang yang seragam, biasanya antara 0 dan 1, tetapi dapat disesuaikan dengan rentang lainnya jika diperlukan.

Proses normalisasi Min-Max melibatkan dua langkah utama:

A. Menentukan rentang target: Pertama, kita harus menentukan rentang target yang diinginkan untuk data yang akan dinormalisasi. Rentang ini biasanya adalah 0 hingga 1, tetapi dapat disesuaikan dengan kebutuhan tertentu.
B. Mengaplikasikan formula normalisasi: Setelah rentang target ditentukan, langkah selanjutnya adalah menerapkan formula normalisasi pada setiap nilai dalam dataset.

Dengan menerapkan formula normalisasi Min-Max, semua nilai dalam dataset akan berada dalam rentang target yang ditentukan, dengan nilai minimum akan menjadi 0 dan nilai maksimum

menjadi 1. Hal ini membantu dalam menganalisis data dengan lebih mudah, serta memastikan bahwa setiap fitur memiliki pengaruh yang seimbang pada analisis yang dilakukan.

Normalisasi Min-Max sering digunakan dalam berbagai bidang, seperti machine learning, data mining, dan analisis data, untuk mempersiapkan data sebelum dilakukan proses seperti clustering, klasifikasi, atau regresi.

2.4 Oversampling SMOTE

Dataset Talasemia termasuk dataset yang tidak seimbang. Terdapat perbedaan yang sangat signifikan antara kelas 0 (alpha-carrier) dan kelas 1 (normal). Ada beberapa hal yang bisa dilakukan untuk mengatasi data yang tidak seimbang. Untuk mengatasi dataset ini, penulis menggunakan teknik oversampling, yaitu SMOTE.

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah metode yang digunakan untuk mengatasi ketidakseimbangan jumlah data antara kelas minoritas dan mayoritas dengan menghasilkan data sintetis. Data sintetis ini dibuat berdasarkan k-tetangga terdekat (*k-nearest neighbor*).

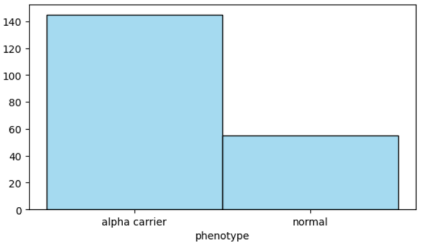
Jumlah k-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan implementasinya. Pembangkitan data sintetis untuk data berjenis numerik berbeda dengan data berjenis kategorikal. Pada data numerik, jarak antara data sintetis dengan tetangganya diukur menggunakan jarak Euclidean, sedangkan pada data kategorikal, nilai modus digunakan sebagai metode yang lebih sederhana.

3. EKSPERIMEN DAN HASIL

Ada beberapa langkah *preprocessing* yang penulis lakukan sebelum memproses dataset Talasemia Alfa, diantaranya:

A. Mengecek Missing Value.
B. Visualisasi Data menggunakan Seaborn, untuk melihat apakah dataset Talasemia Alfa memiliki data yang tidak seimbang atau tidak.
C. Encoding, dengan mengubah data dengan huruf alphabet, menjadi data nominal.
D. Normalisasi Data.

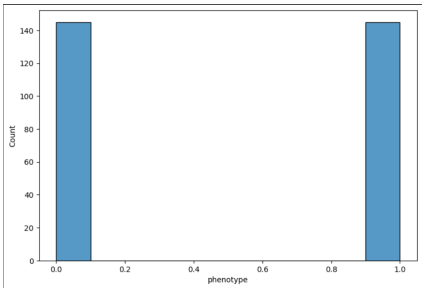
Setelah langkah-langkah diatas dilakukan, penulis menemukan bahwa dataset Talasemia Alfa memiliki data yang tidak seimbang



Gambar 2. Data Sebelum Dioversampling

Dapat dilihat bahwa jumlah data alpha-carrier memiliki jumlah data yang lebih besar daripada jumlah data normal, dan perbedaan jumlah data ini termasuk dengan perbedaan yang signifikan

Penulis melakukan oversampling SMOTE untuk mengatasi ketidakseimbangan data yang muncul.



Gambar 3. Data Setelah Dioversampling

Sebagai keterangan, angka 0 mewakili label alpha carrier, dan angka 1 mewakili kelas normal. Dapat dilihat bahwa data pada kelas phenotype sudah seimbang, sehingga dataset Talasemia Alfa bisa diproses.

Untuk menentukan metode apa yang akan digunakan untuk memproses dataset Talasemia Alfa, penulis melakukan perbandingan dari 3 pemodelan, yaitu KNN, Naive Bayes dan SVM

```
============================== SVC ===================
              precision    recall  f1-score   support

           0       0.44      0.96      0.60        26
           1       0.00      0.00      0.00        32

    accuracy                           0.43        58
   macro avg       0.22      0.48      0.30        58
weighted avg       0.20      0.43      0.27        58

============================================================
============================== Gaussian ================
              precision    recall  f1-score   support

           0       0.67      0.38      0.49        26
           1       0.63      0.84      0.72        32

    accuracy                           0.64        58
   macro avg       0.65      0.61      0.60        58
weighted avg       0.65      0.64      0.62        58

============================================================
============================== KNN ===================
              precision    recall  f1-score   support

           0       0.86      0.46      0.60        26
           1       0.68      0.94      0.79        32

    accuracy                           0.72        58
   macro avg       0.77      0.70      0.69        58
weighted avg       0.76      0.72      0.70        58
```

Gambar 4. Perbandingan Akurasi 3 Pemodelan (KNN, Naive Bayes, dan SVM)

Dapat dilihat, bahwa jika menggunakan metode SVM, akan mendapatkan akurasi sebesar 43%, jika menggunakan metode Naive Bayes, akan mendapatkan akurasi sebesar 64%, dan jika menggunakan metode KNN, akan mendapatkan akurasi sebesar 72%. Berdasarkan hal tersebut, pemodelan KNN yang dipilih untuk memproses dataset Talasemia Alfa, karena akurasinya yang cukup tinggi.

Sebelum penulis mengaplikasikan pemodelan KNN pada dataset, penulis membagi dataset Talasemia Alfa menjadi data latih sebanyak 160 data dan data uji menjadi 40 data.

Hasil pemodelan KNN diaplikasikan pada dataset Talasemia Alfa adalah sebagai berikut.

```
Untuk K = 3 hasilnya adalah 23
Untuk K = 7 hasilnya adalah 30
Untuk K = 9 hasilnya adalah 30
Untuk K = 13 hasilnya adalah 29
```

Gambar 5. Hasil Pemodelan KNN

Dapat dilihat, bahwa jika kita memilih k = 3, jumlah prediksi benar terhadap data uji sebanyak 23 data, jika kita memilih k = 7 dan k = 9, jumlah prediksi benar sebanyak 30 data, dan jika kita memilih k = 13, jumlah prediksi benar sebanyak 29 jumlah data.

$$Akurasi = \frac{Jumlah\ prediksi\ benar}{jumlah\ data\_uji}$$

Gambar 6. Rumus Menentukan Akurasi

Dengan menggunakan rumus sederhana diatas untuk menghitung akurasi, dengan jumlah data uji sebanyak 40 data, maka pada k = 3, akan memiliki akurasi sebesar 57,5%. Hal tersebut diaplikasikan pada k = 7, k = 9 dan k = 13. Hasilnya, k = 7 dan k = 9 memiliki akurasi sebesar 75% dan k = 9 memiliki akurasi sebesar 72,5%.

Dapat dilihat bahwa dari k = 7 dan k = 9 memiliki akurasi lebih besar daripada k = 3, namun jika kita menaikkan jumlah k, akurasinya akan menurun. Sehingga penulis menyimpulkan bahwa KNN dengan k dengan rentang 7 hingga 9 merupakan pemodelan yang paling optimal.

4. KESIMPULAN

Dataset Talasemia Alfa adalah sebuah dataset yang efektif untuk mendeteksi pembawa penyakit talasemia alfa, dengan menggunakan beberapa variabel seperti konsentrasi hemoglobin, volume sel darah merah, volume sel darah merah, hemoglobin sel darah merah dan variabel lainnya. Jenis dataset tersebut termasuk dalam jenis dataset yang menggunakan skala rasio.

Pemodelan yang penulis gunakan adalah metode *K-Nearest Neighbors* (KNN) yang merupakan salah satu metode *supervised learning* yang digunakan untuk klasifikasi. KNN melakukan klasifikasi dengan mencari tetangga terdekat dari data yang akan diklasifikasikan berdasarkan jaraknya. Penulis juga menggunakan Oversampling SMOTE (*Synthetic Minority Over-sampling Technique*) yang merupakan metode yang digunakan untuk mengatasi ketidakseimbangan jumlah data antara kelas minoritas dan mayoritas dengan menghasilkan data sintetis.

Setelah melakukan *preprocessing* dan data pada dataset Talasemia Alfa telah seimbang, penulis memproses dataset menggunakan metode KNN. Dari pemrosesan tersebut, penulis mendapatkan hasil bahwa pemodelan KNN menggunakan k dengan rentang 7 hingga 9 adalah pemodelan paling optimal dengan akurasi sebesar 75%.

DAFTAR PUSTAKA

Astarani & Gustava Siburian. (2016, Juli). GAMBARAN KECEMASAN ORANG TUA PADA ANAK DENGAN THALASEMIA. *Jurnal STIKES*, 9, p.3.

Azmatul Barro, Dina Sulvianti, & Mochamad Afendi. (2013). PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) TERHADAP DATA TIDAK SEIMBANG PADA PEMBUATAN MODEL KOMPOSISI JAMU. Departemen Statistika FMIPA IPB, 1–2.

Istighfarizky F, dkk. (2022, Agustus). Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur. *Jurnal Elektronik Ilmu Komputer Udayana*.