Perbandingan Algoritma Klasifikasi untuk Prediksi Penyakit Anemia Menggunakan Orange

Bagus Adi Prabowo
Magister Teknik Informatika, Universitas Pamulang
E-Mail: bagusadiprabowo46@gmail.com

Abstrak*—Penelitian ini melakukan analisis terhadap membandingkan kinerja beberapa algoritma klasifikasi yang berbeda dalam prediksi penyakit Anemia. Data yang digunakan dalam penelitian ini mencakup atribut Blue Pixel, Sex, Red Pixel, Green Pixel, dan Hb. Anemia atau yang secara awam dikenal dengan kurang darah, merupakan suatu keadaan dimana terjadi penurunan kadar hemoglobin (Hb) di dalam sel darah merah yang berfungsi untuk mengangkut oksigen ke seluruh tubuh sehingga kebutuhan oksigen jaringan tidak terpenuhi. Dalam penelitian ini, telah dilakukan perbandingan algoritma klasifikasi untuk memprediksi penyakit anemia menggunakan aplikasi Orange. Hasil menunjukkan bahwa algoritma Decision Tree, Logistic Regression, dan Naive Bayes memberikan nilai akurasi terbaik sebesar 96,2% untuk dataset anemia dari Kaggle. Di sisi lain, algoritma k-Nearest Neighbors (kNN) memiliki performa terendah dengan akurasi sebesar 94,2%. Kesimpulan ini menegaskan efektivitas beberapa algoritma dalam prediksi anemia, dengan model terbaik adalah Decision Tree, Logistic Regression, dan Naive Bayes.

***Kata Kunci*— Prediksi anemia, *Logistic Regression*, *Decision Tree*, Naive Bayes, kNN.**

I. PENDAHULUAN

Anemia atau yang secara awam dikenal dengan kurang darah, merupakan suatu keadaan dimana terjadi penurunan kadar hemoglobin (Hb) di dalam sel darah merah yang berfungsi untuk mengangkut oksigen ke seluruh tubuh sehingga kebutuhan oksigen jaringan tidak terpenuhi mengakibatkan kekurangan suplai oksigen dapat mengganggu fungsi organ tubuh [1].

Kurang darah atau anemia adalah gangguan darah yang ditandai dengan jumlah sel darah merah yang rendah atau ketika sel darah merah tidak berfungsi dengan baik. Sel darah merah memiliki kandungan protein ber-zat besi yang disebut hemoglobin, hemoglobin berfungsi untuk mengikat dan menyalurkan oksigen untuk sel-sel tubuh. Pada kondisi anemia jumlah sel darah merah dan hemoglobin berkurang sehingga oksigen tidak tersuplai dengan baik dan penderita mengeluh lemas dan pucat. Normalnya, orang dewasa menderita anemia apabila kadar hemoglobin darahnya di bawah 14 gram per desiliter pada laki-laki dan 12 gram per desiliter untuk wanita [2].

Gejala anemia dapat berupa kelelahan, kulit pucat, sesak napas, pusing, limbung, atau detak jantung cepat. Pengobatan tergantung pada diagnosis utama. Suplemen zat besi dapat digunakan untuk kekurangan zat besi. Suplemen vitamin B dapat digunakan untuk kadar vitamin rendah. Transfusi darah dapat digunakan untuk kehilangan darah. Obat untuk mendorong pembentukan darah dapat digunakan jika produksi darah tubuh berkurang [3].

Pada artikel ini, kami akan menjelaskan hasil penelitian yang bertujuan untuk membandingkan kinerja berbagai algoritma klasifikasi dalam konteks prediksi anemia. Kami akan menganalisis algoritma seperti Logistic Regression, Decision Trees, Naive Bayes, dan K-Nearest Neighbors (KNN). Hasil penelitian ini diharapkan dapat memberikan informasi berharga bagi para profesional kesehatan, peneliti dan praktisi kesehatan dalam memilih algoritma yang paling tepat untuk memprediksi diagnosis anemia berdasarkan karakteristik pasien. Seiring dengan kemajuan teknologi dan pemrosesan data, pendekatan ini dapat menjadi alat penting dalam memerangi anemia dan meningkatkan kualitas hidup pasien.

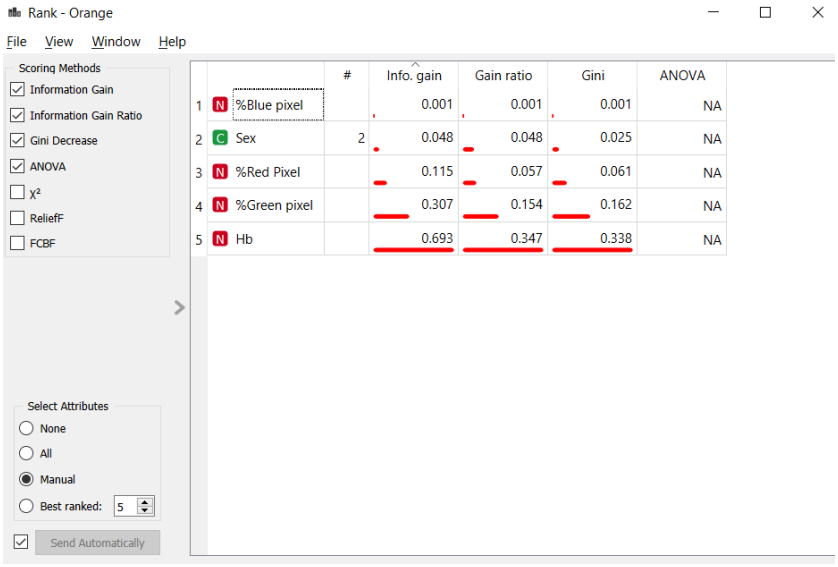
II. METODE PENELITIAN

Metode yang digunakan dalam tulisan ini adalah: (1) pengumpulan data (2) pembuatan model dan pengukuran metrik di aplikasi Orange (3) evaluasi model.

A. Pengumpulan Data

Kumpulan data ini berasal dari referensi sumber media internet dunia Kesehatan Penyakit Anemia. Tujuan dari kumpulan data ini adalah untuk memprediksi secara diagnostik apakah seorang pasien menderita anemia atau tidak, berdasarkan pengukuran diagnostik tertentu yang disertakan dalam kumpulan data tersebut. Beberapa kendala ditempatkan pada pemilihan contoh-contoh ini dari database yang lebih besar. Secara khusus, semua pasien di sini adalah wanita berusia minimal 21 tahun keturunan India Pima. Data yang digunakan sudah dalam kondisi bersih dan tervalidasi, sehingga tidak diperlukan tahapan pemrosesan data.

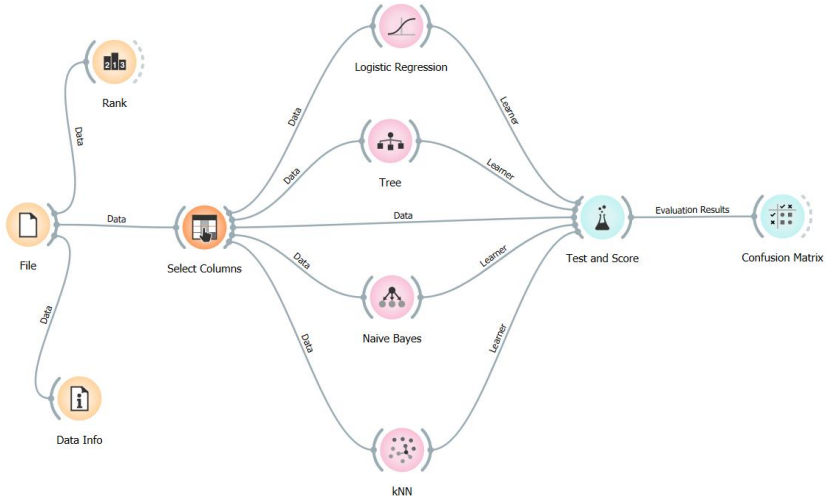
Jumlah sampel sebanyak 105 data dan terdiri dari 7 atribut yaitu : Sex, %Red Pixel, %Green Pixel, %Blue Pixel, Hb, Anaemic, dan Name. Setelah dilakukan seleksi atribut, hanya akan diambil 5 besar atribut yang paling berpengaruh yaitu : %Blue Pixel, Sex, %Red Pixel, %Green Pixel, dan Hb berdasarkan *#, Information Gain, Gain Ratio, Gini dan ANOVA.*



Gambar 1.
Model *Rank*

B. Pembuatan Model di Aplikasi Orange dan Pengukuran Metrik

Algoritma yang akan dibandingkan adalah: (1) Logistic Regression (2) kNN (3) Naive Bayes (4) Decision Tree. *Logistic Regression* adalah algoritma pembelajaran mesin yang digunakan untuk memecahkan masalah klasifikasi dengan memodelkan probabilitas kejadian suatu peristiwa dengan memperhatikan faktor-faktor prediktor yang berkaitan. Algoritma *kNN* adalah algoritma klasifikasi yang mendasarkan pada mayoritas kategori dari k- tetangga terdekat. Algoritma *Naive Bayes* adalah algoritma yang mempelajari probabilitas suatu objek dengan ciri-ciri tertentu yang termasuk dalam kelompok/kelas tertentu. *Decision Tree* merupakan salah satu cara data processing dalam memprediksi masa depan dengan cara membangun klasifikasi atau regresi model dalam bentuk struktur pohon.



Gambar 2.
Model Aplikasi Data Orange

Metrik yang dipergunakan untuk pengukuran adalah *Accuracy*. Metrik ini mengukur % kebenaran prediksi tiap algoritma terhadap jumlah keseluruhan data. Hasil akan dimasukkan ke dalam empat kategori table matrix, yakni TP (True Positif), FP (False Positif), FN (False Negatif) dan TN (True Negatif). Perhitungan akurasi data akan diperoleh dari nilai yang didapatkan pada confusion matrix dengan ketentuan persamaan, Rumus dari Accuracy ini adalah:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

C. Analisa metrik

Analisa Metrik merupakan hal utama dalam penulisan ini. Terdapat beberapa definisi dari metrik, salah satunya dari Fenton dan Pfleeger: "sebuah proses dimana angka atau simbol ditempatkan sebagai atribute dari sebuah entity pada dunia nyata dengan sedemikian rupa untuk menggambarkan entity tersebut berdasarkan aturan yang terdefinisi dengan jelas" Entity adalah sebuah

obyek, seperti modul perangkat lunak, sedangkan attribute merupakan properti yang dapat diukur dari sebuah obyek. Entity dapat dibedakan menjadi tiga kategori: produk, proses, dan sumber daya. Proses merupakan kegiatan yang berhubungan dengan pengembangan perangkat lunak, sebuah produk merupakan artefak yang dihasilkan dalam pengembangan perangkat lunak. Sedangkan sumber daya merupakan orang, hardware, ataupun software yang dibutuhkan dalam proses tersebut. Atribut merupakan properti dari entity, seperti tinggi badan. Atribut dapat dikelompokan menjadi dua kategori utama, internal dan eksternal. Internal atribut diukur secara langsung dari entity, sedangkan eksternal atribut merupakan atribut yang tidak secara langsung diukur dari entity, berupa hasil perhitungan ataupun diturunkan dari internal atribut. Melalui metrik, produk dan proses dari perangkat lunak dapat diukur dan dibandingkan secara objektif.

III. HASIL DAN PEMBAHASAN

A. *Confusion Matrix*

Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Evaluasi dengan confusion matrix menghasilkan nilai accuration, precision, recall dan f-measure. Accuration dalam klasifikasi adalah persentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi.

Tabel 1.
Jaringan siap uji coba

Aktual	**Cassified as**	
	+	-
+	True Positive (TP)	True Negative (TN)
-	False Positive (FP)	False Negative (FN)

- True Positive (TP)
  Interpretasi: Anda memprediksi positif dan itu benar.
- True Negative (TN):
  Interpretasi: Anda memprediksi negatif dan itu benar.
- False Positive (FP): (Kesalahan Tipe 1)
  Interpretasi:Anda memprediksi positif dan itu salah.
- False Negative (FN): (Kesalahan Tipe 2, kesalahan tipe 2 ini sangat berbahaya)
  Interpretasi : Anda memprediksi negatif dan itu salah.

Dari contoh di atas dapat digambarkan bahwa:

- Nilai Prediksi adalah keluaran dari program dimana nilainya Positif dan Negatif
- Nilai Aktual adalah nilai sebenarnya dimana nilainya True dan False

Tabel 2 merupakan tampilan tabel pada *confusion matrix* untuk melakukan perhitungan *accuration*, *precision*, *recall* dan f-*measure*. Dari model yang dibangun dengan mempergunakan masing-masing algoritma, didapatkan confusion matrix sebagai berikut:

Tabel 2.
Model *Confussion Matrix*

Logistic Regression

Actual \ Predicted	No	Yes	Σ
No	77	1	78
Yes	3	23	26
Σ	80	24	104

Naive Bayes

Actual \ Predicted	No	Yes	Σ
No	76	2	78
Yes	2	24	26
Σ	78	26	104

Decision Tree

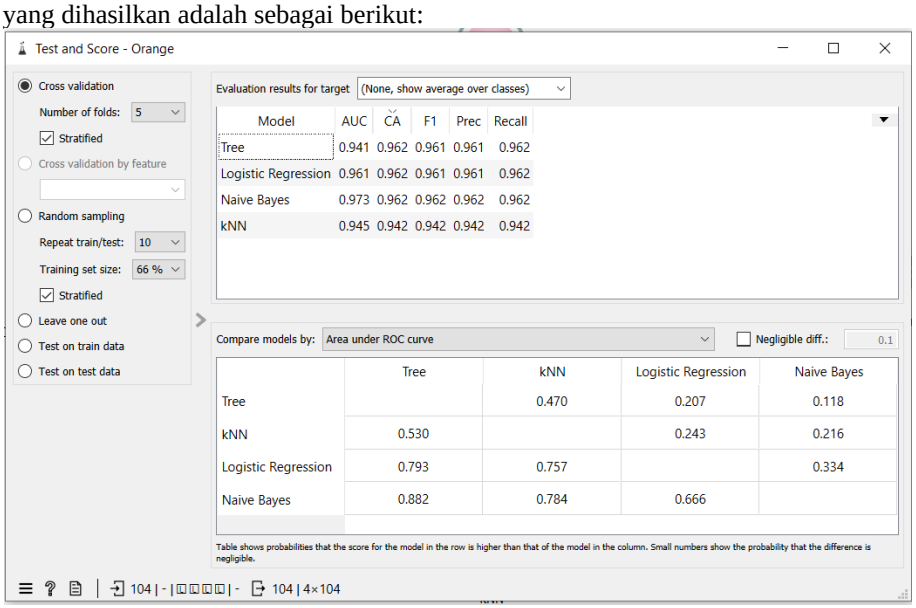
Actual \ Predicted	No	Yes	Σ
No	77	1	78
Yes	3	23	26
Σ	80	24	104

kNN

Actual \ Predicted	No	Yes	Σ
No	76	2	78
Yes	4	22	26
Σ	80	24	104

B. Confusion Matrix

Confusion Metrix yang dihasilkan adalah sebagai berikut:



Gambar 3.
Model *Test and Score*

Dari metriks di atas didapatkan bahwa jika diurutkan berdasarkan % kebenaran prediksi / *Accuracy*/ CA yang dihasilkan masing-masing model:

a. #1 adalah ***Decision Tree*** dengan nilai CA 0.962 atau 96.2%.
b. #2 adalah ***Logistic Regression*** dengan nilai CA 0.962 atau 96.2%.
c. #3 adalah **Naïve Bayes** dengan nilai CA 0.962 atau 96.2 %.
d. #4 adalah **kNN** dengan nilai CA 0.942 atau 94.2%.

IV. KESIMPULAN

Kesimpulan yang diambil dalam penelitian ini adalah:

1. Beberapa dari algoritma klasifikasi tersebut memberikan nilai Accuracy yang sama yaitu *Decision Tree*, *Logistic Regression* dan *Naive Bayes* untuk domain permasalahan
2. Model terbaik dalam memprediksi penyakit anemia berdasarkan dataset dari kaggle adalah *Decision Tree, Logistic Regression* dan *Naive Bayes* dengan nilai Accuracy 96.2%.
3. Model terburuk dalam memprediksi penyakit anemia berdasarkan dataset dari kaggle adalah **kNN** dengan nilai Accuracy 94.2%.

DAFTAR PUSTAKA

[1] S. Hospitals, “Apa Itu Anemia?,” https://www.siloamhospitals.com/informasi-siloam/artikel/apa-itu-anemia.
[2] H. N. Prabaningsih, J. Farizal, G. Baruara, H. Laksono, and P. W. Welkriana, “Gambaran Kadar Hemoglobin Sebelum dan Sesudah Donor Darah di UTD PMI Kota Bengkulu Tahun 2022,” Poltekkes Kemenkes Bengkulu, 2022.
[3] A. Bros, “Kenali Jenis Anemia,” https://awalbros.com/patologi-klinik/kenali-jenis-anemia/.