PROPOSAL RISET INFORMATIKA

Klasifikasi Anemia Berbasis CatBoost dan Analisis Interpretabilitas Model dengan SHAP

Disusun Oleh :

Vika Rafi Ana 22081010009

PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR

2025

BAB I

PENDAHULUAN

1.1 Latar Belakang

Anemia merupakan salah satu masalah kesehatan global yang masih menjadi perhatian utama, terutama di negara berkembang seperti Indonesia. Berdasarkan penelitian oleh Djaafara et al. (2024), prevalensi anemia di Indonesia cukup tinggi dan memiliki penyebab yang kompleks, mulai dari kekurangan zat besi hingga gangguan genetik. Selain itu, studi yang dilakukan oleh Indriyani et al. (2024) menunjukkan bahwa prevalensi anemia di kalangan wanita pekerja di Indonesia mencapai sekitar 31,2%, menandakan bahwa anemia masih menjadi tantangan kesehatan yang serius. Pada kelompok ibu hamil, penelitian oleh Yuliana et al. (2024) menunjukkan peningkatan signifikan kasus anemia berdasarkan hasil Survei Kesehatan Dasar (Riskesdas), dari 37,1% menjadi 48,9%.

Deteksi dini anemia umumnya dilakukan melalui pemeriksaan *Complete Blood Count (CBC)* yang menghasilkan berbagai parameter darah seperti kadar hemoglobin, hematokrit, dan jumlah eritrosit. Namun, proses interpretasi hasil CBC secara manual bergantung pada keahlian tenaga medis, sehingga rentan terhadap subjektivitas dan keterbatasan waktu (Sharma et al., 2023). Oleh karena itu, dibutuhkan sistem otomatis yang mampu membantu proses klasifikasi anemia secara cepat dan akurat.

Perkembangan teknologi *machine learning* memungkinkan proses klasifikasi dilakukan dengan algoritma yang mampu mengenali pola kompleks pada data medis. Salah satu algoritma yang terbukti efektif dalam menangani data tabular adalah CatBoost, yang dikembangkan oleh tim Yandex (Dorogush, Ershov and Gulin, 2018). CatBoost memiliki kemampuan mengolah data kategorikal secara langsung serta menghindari *overfitting*,

menjadikannya unggul dibandingkan beberapa algoritma lain seperti XGBoost dan LightGBM.

Namun, penerapan *machine learning* di bidang medis masih menghadapi tantangan besar, yaitu sulitnya menjelaskan hasil prediksi model. Untuk menjawab hal tersebut, digunakan metode SHAP (Shapley Additive Explanations) yang dapat memberikan interpretasi mengenai pengaruh masing-masing fitur terhadap hasil klasifikasi (Lundberg and Lee, 2017). Pendekatan ini membantu menjadikan sistem *machine learning* lebih transparan dan dapat dipercaya oleh praktisi medis.

Berdasarkan uraian di atas, penelitian ini akan berfokus pada pengembangan model klasifikasi anemia berbasis CatBoost serta analisis interpretabilitas model menggunakan SHAP, dengan tujuan menghasilkan sistem yang akurat dan mudah dipahami oleh tenaga medis.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana penerapan algoritma CatBoost dalam mengklasifikasikan penyakit anemia berdasarkan data *Complete Blood Count (CBC)*?
2. Bagaimana hasil evaluasi performa model klasifikasi CatBoost dibandingkan dengan model pembanding lainnya?
3. Bagaimana interpretasi hasil klasifikasi menggunakan metode SHAP untuk mengetahui faktor-faktor dominan dalam prediksi anemia?

1.3 Identifikasi Masalah

Berdasarkan uraian pada latar belakang, dapat diidentifikasi beberapa permasalahan yang menjadi dasar penelitian ini, yaitu:

1. Masih tingginya angka kejadian anemia di Indonesia pada berbagai kelompok masyarakat, seperti wanita pekerja dan ibu hamil, menunjukkan perlunya sistem pendeteksi dini yang efisien.
2. Proses identifikasi anemia masih dilakukan secara manual berdasarkan hasil pemeriksaan darah (*Complete Blood Count/CBC*), sehingga membutuhkan waktu dan tenaga medis yang berpengalaman.
3. Belum optimalnya penerapan algoritma pembelajaran mesin (machine learning) untuk mengklasifikasikan anemia secara otomatis dengan akurasi tinggi.
4. Kurangnya transparansi dan interpretabilitas model machine learning yang digunakan di bidang medis, sehingga hasil prediksi sulit dijelaskan kepada tenaga kesehatan.
5. Belum adanya sistem klasifikasi anemia yang menggabungkan model berperforma tinggi seperti CatBoost dengan metode interpretasi SHAP agar hasilnya tidak hanya akurat tetapi juga dapat dijelaskan.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menerapkan algoritma CatBoost untuk melakukan klasifikasi penyakit anemia berdasarkan data CBC.
2. Mengevaluasi performa model menggunakan metrik akurasi, presisi, recall, dan f1-score.
3. Menganalisis interpretabilitas model menggunakan metode SHAP untuk menjelaskan pengaruh fitur terhadap hasil prediksi.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penerapan algoritma CatBoost yang dikombinasikan dengan metode Shapley Additive Explanations (SHAP) telah banyak digunakan pada berbagai bidang penelitian, termasuk kesehatan dan analisis data prediktif, karena kemampuannya memberikan akurasi tinggi sekaligus interpretabilitas model yang baik.

Penelitian oleh Praptiwi, Kurnia, Fitrianto, dan Ernawati (2024) menerapkan CatBoost untuk mendeteksi faktor risiko anemia pada anak usia 5–12 tahun. Model ini memperlihatkan performa yang stabil dalam menangani data tidak seimbang melalui pembobotan kelas dan menghasilkan interpretasi yang jelas menggunakan SHAP, yang membantu mengidentifikasi variabel dominan seperti kadar hemoglobin dan hematokrit terhadap diagnosis anemia.

Studi *scoping review* oleh Kario dan Rico Kurniawan (2024) meninjau berbagai penelitian lokal terkait penerapan *machine learning* dalam klasifikasi anemia. Mereka mencatat bahwa algoritma *gradient boosting* seperti CatBoost semakin banyak digunakan karena kemampuannya mengolah data tabular medis dengan efisien. Selain itu, metode SHAP menjadi pendekatan yang populer untuk meningkatkan transparansi hasil model di bidang medis.

Penelitian oleh Kurniawan dan Widodo (2023) menunjukkan bahwa kombinasi CatBoost + SHAP mampu menghasilkan model prediksi penyakit dengan akurasi tinggi serta memberikan pemahaman mendalam mengenai pengaruh setiap fitur terhadap hasil

klasifikasi. Meskipun topiknya adalah diabetes, pendekatan metodologinya menunjukkan potensi besar untuk diterapkan juga pada deteksi penyakit anemia.

Pada bidang non-medis, penelitian Wijaya dan Santoso (2023) membuktikan bahwa kombinasi CatBoost dan SHAP dapat digunakan untuk menjelaskan faktor-faktor utama yang memengaruhi popularitas game, seperti jumlah ulasan positif dan harga. Hasil penelitian ini memperkuat peran CatBoost sebagai model yang kuat dan fleksibel untuk data tabular serta SHAP sebagai alat interpretasi yang efektif di berbagai domain.

Penelitian internasional oleh Liu et al. (2023) dalam artikel juga menekankan keunggulan CatBoost + SHAP dalam bidang kesehatan. Studi ini menggunakan model CatBoost untuk memprediksi risiko rawat inap ulang pasien gagal jantung lansia selama satu tahun, dan SHAP digunakan untuk menilai kontribusi setiap fitur klinis terhadap risiko tersebut. Hasilnya menunjukkan bahwa pendekatan berbasis interpretabilitas mampu meningkatkan kepercayaan praktisi medis terhadap model prediktif.

Secara keseluruhan, berbagai penelitian tersebut membuktikan bahwa kombinasi CatBoost dan SHAP memberikan hasil yang kuat dalam hal performa prediksi sekaligus transparansi model. Penerapan kedua metode ini pada kasus klasifikasi anemia diharapkan dapat menghasilkan sistem yang tidak hanya akurat tetapi juga mudah diinterpretasikan oleh tenaga medis.

2.2 Landasan Teori

2.2.1 Anemia dan Pemeriksaan CBC

Anemia merupakan kondisi berkurangnya kadar hemoglobin dalam darah yang mengakibatkan pasokan oksigen ke jaringan tubuh menjadi tidak optimal (WHO, 2023). Pemeriksaan *Complete Blood Count (CBC)* meliputi pengukuran parameter seperti Hb, HCT, MCV, MCH, dan RBC yang menjadi dasar dalam menentukan jenis anemia (Kumar and Clark, 2022).

2.2.2 Algoritma CatBoost

CatBoost adalah algoritma *gradient boosting* yang dikembangkan oleh Yandex. Keunggulannya terletak pada kemampuannya mengolah data kategorikal tanpa proses *one-hot encoding* dan kemampuannya mengurangi *overfitting* (Dorogush, Ershov and Gulin, 2018).

2.2.3 Interpretabilitas Model dan SHAP

Interpretabilitas model (*model interpretability*) adalah kemampuan model untuk menjelaskan alasan di balik prediksi yang dihasilkan. SHAP (Shapley Additive Explanations) menghitung kontribusi setiap fitur berdasarkan teori permainan kooperatif (*cooperative game theory*) (Lundberg and Lee, 2017). Nilai SHAP dapat ditampilkan dalam bentuk visual seperti *summary plot* atau *force plot*, sehingga pengguna dapat memahami fitur mana yang paling berpengaruh.

2.2.4 Evaluasi Model Klasifikasi

Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan f1-score untuk menilai performa model secara menyeluruh (Han, Kamber and Pei, 2022).

BAB III

METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini bersifat eksperimen dengan pendekatan kuantitatif, karena melibatkan proses pengujian algoritma *machine learning* menggunakan data numerik dari hasil pemeriksaan darah. Pendekatan ini digunakan untuk menganalisis hubungan antara parameter hematologi dengan kondisi anemia melalui penerapan algoritma CatBoost dan analisis interpretabilitas menggunakan SHAP. Tujuan dari penelitian ini adalah untuk menghasilkan model klasifikasi anemia yang akurat, stabil, dan dapat dijelaskan berdasarkan kontribusi setiap parameter darah terhadap hasil prediksi.

Data yang digunakan dalam penelitian ini berasal dari dataset publik berjudul "Anemia Types Classification", yang tersedia di platform Kaggle (Ehab Aboelnaga, 2022). Dataset ini digunakan sebagai sumber utama dalam pengembangan dan pengujian model klasifikasi anemia berbasis algoritma CatBoost.

Dataset tersebut terdiri atas sekitar 1.282 baris data dengan 15 atribut, yang merepresentasikan berbagai parameter hasil pemeriksaan darah atau *Complete Blood Count (CBC)*. Setiap atribut berisi nilai-nilai laboratorium yang umum digunakan dalam diagnosis anemia, seperti kadar hemoglobin (HGB), hematokrit (HCT), jumlah sel darah merah (RBC), volume sel darah rata-rata (MCV), hemoglobin rata-rata per sel (MCH), konsentrasi hemoglobin rata-rata (MCHC), jumlah sel darah putih (WBC), dan jumlah trombosit (PLT).

Atribut target dalam dataset ini adalah Class, yang menunjukkan kategori jenis anemia yang dialami oleh pasien. Kelas yang terdapat dalam dataset meliputi *No Anemia*, *Mild Anemia*, *Moderate Anemia*, dan *Severe Anemia*. Data tersebut kemudian digunakan untuk melatih dan menguji model klasifikasi, dengan tujuan menghasilkan sistem yang mampu mengidentifikasi tipe anemia berdasarkan hasil pemeriksaan laboratorium secara akurat dan terukur.

3.3 Tahapan Penelitian

Tahapan penelitian ini dirancang secara sistematis agar proses klasifikasi jenis anemia dapat dilakukan secara optimal dan menghasilkan model yang akurat serta mudah diinterpretasikan. Adapun langkah-langkah penelitian yang dilakukan meliputi:

1. Pengumpulan Data

   Dataset yang digunakan berasal dari sumber publik berjudul *Anemia Types Classification* di platform Kaggle, yang berisi data hasil pemeriksaan darah lengkap (*Complete Blood Count/CBC*) beserta kategori jenis anemia.
2. Pra-pemrosesan (data cleaning, encoding, normalisasi)

   Tahap ini meliputi proses pembersihan data dengan menghapus nilai kosong atau data duplikat yang dapat memengaruhi performa model. Selanjutnya dilakukan normalisasi untuk menyamakan skala antar variabel numerik agar model tidak bias terhadap fitur dengan nilai besar. Jika ditemukan ketidakseimbangan jumlah data antar kelas anemia, maka penyesuaian dilakukan dengan menerapkan parameter *class_weight* pada algoritma CatBoost, sehingga model dapat memberikan bobot lebih besar pada kelas minoritas tanpa perlu melakukan penambahan data sintetis. Pendekatan ini membantu menjaga proporsi data asli sekaligus meningkatkan keadilan model dalam mempelajari seluruh kategori anemia.
3. Pembagian Data (Train-Test Split)

   Setelah data siap, dataset dibagi menjadi dua bagian, yaitu 80% sebagai data latih dan 20% sebagai data uji. Pembagian ini bertujuan untuk memastikan bahwa performa model dapat diukur secara objektif terhadap data yang belum pernah dilihat sebelumnya.

4. Penerapan algoritma CatBoost

   Pada tahap ini dilakukan pembangunan model klasifikasi menggunakan algoritma CatBoost, yang merupakan salah satu algoritma *gradient boosting* modern dengan keunggulan dalam menangani data tabular dan mencegah *overfitting*. Proses pelatihan model disertai dengan tuning parameter untuk memperoleh kombinasi terbaik yang menghasilkan akurasi maksimal.
5. Interpretasi Model dengan SHAP

   Tahap akhir melibatkan penggunaan metode **SHAP (Shapley Additive Explanations)** untuk menjelaskan kontribusi masing-masing fitur terhadap hasil prediksi model. Dengan demikian, hasil klasifikasi tidak hanya memberikan nilai akurasi, tetapi juga penjelasan yang transparan mengenai faktor-faktor laboratorium yang paling memengaruhi diagnosis anemia.

3.4 Evaluasi Hasil

Evaluasi performa model dilakukan dengan menggunakan Confusion Matrix serta lima metrik utama, yaitu accuracy, precision, recall, F1-score, dan AUC (Area Under the Curve).

Metrik accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data uji, memberikan gambaran umum mengenai tingkat keakuratan model. Precision mengindikasikan seberapa tepat model dalam memprediksi kelas positif atau kategori tertentu tanpa banyak kesalahan klasifikasi. Recall menunjukkan sejauh mana model mampu mengenali seluruh sampel dalam suatu kelas, terutama penting dalam konteks diagnosis anemia untuk meminimalkan kasus yang tidak terdeteksi. F1-score berfungsi sebagai ukuran keseimbangan antara precision dan recall, memberikan penilaian yang lebih adil ketika distribusi kelas tidak seimbang.

Selain itu, digunakan pula metrik AUC (Area Under the Receiver Operating Characteristic Curve) untuk menilai kemampuan model dalam membedakan antar kelas. Nilai

AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik dalam memisahkan jenis anemia berdasarkan parameter darah.