ANALISIS PERFORMA ALGORITMA KLASIFIKASI MULTILABEL PADA MECHANISM OF ACTION DATASET

Ryan Fadhilah Faizal Hakim[1] Catherine Vanya Pangemanan[2]

1.2. Universitas Koperasi Indonesia
Kawasan Pendidikan Tinggi Jl. Raya Jatinangor
No.KM. 20, RW.5, Cibeusi, Kec. Jatinangor,
Kabupaten Sumedang, Jawa Barat 45363
Email : -

1. PENDAHULUAN

Peta Konektivitas MoA, sebuah proyek dalam Broad Institute di MIT dan Harvard, Laboratorium Ilmu Inovasi di Harvard (LISH), dan Perpustakaan Tanda Tangan Seluler Berbasis Jaringan Terintegrasi Dana Umum NIH (LINCS), mempersembahkan dataset Mechanism of Action ini dengan tujuan untuk memajukan pengembangan obat melalui perbaikan algoritma prediksi Mekanisme Aksi (MoA) itu sendiri.

Project MoA (Mechanism of Action) merujuk pada inisiatif atau penelitian yang bertujuan untuk memahami mekanisme biologis atau kimiawi di balik cara suatu senyawa, obat, atau terapi bekerja di dalam tubuh. Secara spesifik, MoA menggambarkan langkahlangkah molekuler dan jalur yang terlibat dalam interaksi antara obat atau terapi dengan target biologisnya, seperti reseptor, enzim, atau protein tertentu. Contoh penelitian terkait MoA adalah misalnya mengenai obat, vaksin dan terapi. Secara keseluruhan, proyek MoA merupakan bagian penting dari riset biomedis dan farmasi yang berfokus pada identifikasi dan pemahaman cara suatu senyawa atau terapi bekerja dalam tubuh untuk mencapai hasil yang diinginkan.

Dataset Mechanism of Action (MoA) dari obatobatan menghadirkan tantangan yang signifikan dalam klasifikasi berbasis multilabel sebab dataset ini terdiri atas 23.000 baris data dan memiliki lebih dari 1.400 fitur, yang mencakup berbagai atribut molekuler dan karakteristik farmakologis, serta memerlukan pendekatan komputasi untuk pengolahan yang efektif. Tantangan utama dari dataset ini adalah kompleksitas data dan kebutuhan klasifikasi multilabel, di mana satu obat dapat memiliki beberapa mekanisme kerja (modulasi) yang tumpangtindih. Hal ini membuat pendekatan tradisional (pengamatan manual saat membaca data yang sangat banyak sebelum akhirnya memutuskan keadaan suatu faktor apakah menghasilkan keputusan pengklasifikasian MoA atau bukan, yang tepat) dalam prediksi menjadi kurang efektif, sehingga diperlukan algoritma machine learning untuk menangani tingkat kompleksitas yang tinggi dan meminimalisasi risiko bias dalam hasil prediksi (Liang et al., 2023; Brown & Zhang, 2022).

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma LightGBM efektif dalam menangani tantangan klasifikasi multilabel pada dataset besar dengan kompleksitas tinggi. Sebagai contoh, pada studi oleh Li et al. (2021), LightGBM berhasil mengklasifikasikan dataset multilabel dengan akurasi ratarata 92.3%, precision 89.5%, recall 87.4%, dan F1score 88.9%, lebih unggul dibandingkan metode lain seperti Random Forest dan XGBoost (Li et al., 2021). Hasil ini menunjukkan bahwa LightGBM memiliki keunggulan dalam memproses data berdimensi tinggi dan multilabel, menjadikannya pilihan yang ideal untuk dataset MoA ini.

Dalam studi lainnya, Zhang et al. (2022) menemukan bahwa LightGBM memberikan hasil akurasi dan efisiensi komputasi yang lebih tinggi dalam kasus klasifikasi multilabel dibandingkan Logistic Regression dan Random Forest, dengan akurasi sebesar 94.1% dan F1score 90.2% (Zhang et al., 2022). Penelitian ini juga menyebutkan bahwa penggunaan LightGBM mengurangi risiko overfitting pada dataset multilabel berukuran besar karena kemampuannya mengelola variabilitas data yang tinggi, yang penting untuk klasifikasi yang akurat dan generalis.

Namun, walaupun LightGBM telah menjadi algoritma paling direkomendasikan untuk kasus klasifikasi multilabel seperti ini, kami tetap akan mencoba algoritma klasifikasi lainnya. Maka dari itu penelitian ini mengusulkan kajian komparatif terhadap performa beberapa algoritma machine learning terkemuka—Logistic Regression, Random Forest, XGBoost, dan LightGBM—untuk klasifikasi multilabel pada dataset MoA. Kami berhipotesis bahwa LightGBM akan memberikan hasil paling unggul dibandingkan algoritma lainnya, tetapi penelitian ini akan memvalidasi asumsi tersebut melalui perbandingan akurasi, precision, recall, F1score, dan confusion matrix untuk setiap algoritma yang diuji.

2. TINJAUAN PUSTAKA

Li et al. (2021) meneliti efektivitas algoritma LightGBM dalam klasifikasi multilabel pada dataset yang memiliki kompleksitas tinggi. Penelitian ini menunjukkan bahwa LightGBM mampu memberikan akurasi 92.3%, precision 89.5%, recall 87.4%, dan F1score 88.9%. Hasil ini mengindikasikan bahwa LightGBM unggul dalam menangani dataset besar dengan banyak label, serta mengurangi risiko overfitting, menjadikannya algoritma yang sesuai untuk aplikasi multilabel yang kompleks (Li et al., 2021).

Penelitian Zhang et al. (2022) mendukung temuan tersebut dengan menyatakan bahwa LightGBM memberikan kinerja lebih tinggi dibandingkan dengan algoritma lain seperti Random Forest dan Logistic Regression dalam klasifikasi multilabel. Pada dataset dengan karakteristik serupa, mereka memperoleh akurasi 94.1% dan F1score 90.2% untuk LightGBM, menunjukkan keunggulan dalam hal akurasi dan efisiensi komputasi. Zhang et al. juga menyebutkan bahwa LightGBM lebih efektif dalam menangani fiturfitur berdimensi tinggi yang terdapat pada dataset MoA, menjadikannya algoritma yang diunggulkan untuk masalah klasifikasi yang serupa (Zhang et al., 2022).

Brown dan Zhang (2022) meneliti berbagai algoritma machine learning, termasuk Random Forest dan XGBoost, untuk mengatasi tantangan tumpangtindih kelas pada dataset multilabel. Meskipun kedua algoritma ini menunjukkan hasil yang cukup baik, penelitian ini menyebutkan bahwa LightGBM memiliki kinerja yang lebih konsisten, terutama pada metrik F1 score dan recall. Studi mereka mengungkapkan bahwa algoritma ensemble seperti Random Forest dan XGBoost sering kali memerlukan finetuning parameter yang lebih intensif untuk mencapai hasil optimal, dan tanpa optimasi ini, performa dapat lebih rendah dibandingkan dengan LightGBM (Brown & Zhang, 2022).

Liang et al. (2023) menyoroti pentingnya hyperparameter tuning dalam meningkatkan kinerja algoritma pada dataset multilabel. Dalam studi mereka, algoritma LightGBM yang telah dioptimalkan menunjukkan peningkatan dalam precision dan recall, sehingga mengurangi bias dan menghasilkan prediksi yang lebih akurat pada data yang kompleks. Penelitian ini menekankan bahwa hyperparameter tuning tidak hanya meningkatkan akurasi tetapi juga menambah generalisasi model untuk data baru, yang sangat penting dalam aplikasi multilabel seperti pada dataset MoA (Liang et al., 2023).

Berdasarkan hasil penelitian tersebut, algoritma LightGBM tampaknya memiliki keunggulan dalam menangani kasus klasifikasi multilabel dengan dataset besar dan kompleks. Namun, penelitian ini akan membandingkan performa beberapa algoritma terkemuka, yaitu Logistic Regression, Random Forest, XGBoost, dan LightGBM, untuk mengidentifikasi algoritma yang paling optimal dalam klasifikasi multilabel pada dataset MoA. Dengan menguji dan membandingkan berbagai metrik, seperti akurasi, precision, recall, dan F1score, kami berharap dapat mengevaluasi secara komprehensif keunggulan setiap algoritma dalam mengklasifikasikan dataset MoA ini agar penelitian ini berhasil mengembangkan model yang dapat memprediksi probabilitas dari target MoA relevan untuk sampel yang diberikan, dengan menggunakan fitur seperti ekspresi gen dan viabilitas sel, serta informasi tentang perlakuan senyawa kimia dan kontrol.

3. METODOLOGI PENELITIAN

Studi ini menggunakan dataset Mechanism of Action (MoA) Prediction yang diakuisisi dari Laboratory for Innovation Science at Harvard pada November 2024. Dataset ini terdiri dari 23.000 sampel dengan lebih dari 1.400 fitur, mencakup

data ekspresi gen, viabilitas sel, dan informasi perlakuan senyawa kimia. Target prediksi adalah respons MoA, di mana nilai biner 1 menandakan keberadaan MoA dan nilai 0 menandakan tidak ada MoA. Komponen dataset terdiri atas:

1. train_features.csv merupakan dataset utama untuk pelatihan model, berisi
   - Data ekspresi gen (g) dan viabilitas sel (c).
   - Jenis perlakuan (cp_type) yang mengidentifikasi apakah sampel diberi senyawa kimia atau kontrol.
   - Durasi perlakuan (cp_time) dan dosis (cp_dose).
2. train_targets_scored.csv berisi label target MoA biner yang akan diprediksi model untuk evaluasi kinerja.
3. train_drug.csv (opsional) menyediakan drug_id anonim yang terkait dengan senyawa kimia, dapat digunakan untuk analisis tambahan.

Proses penggunaan data yaitu dengan data pelatihan menggunakan train_features.csv sebagai fitur input dan train_targets_scored.csv sebagai label target untuk melatih model. Model yang dilatih kemudian diuji menggunakan test_features.csv untuk memprediksi probabilitas terjadinya setiap target MoA. Data tambahan yaitu train_targets_nonscored.csv: Berisi target MoA tambahan yang tidak digunakan dalam penilaian model dan kompetisi.

Algoritma yang akan diujikan adalah Logistic Regression, Random Forest, XGBoost, dan LightGBM dengan data cleaning, preprocessing, normalisasi atau standarisasi atau generalisasi dan hyperparameter tuning yang sedang dalam tahap perancangan.