

Prediksi Nilai Akhir Matematika Siswa Menggunakan Random Forest Berbasis AutoML dengan Seleksi Fitur Algoritma Genetika

Prediction of Student Mathematics Final Grades Using AutoML-Based Random Forest with Genetic Algorithm Feature Selection

Aprianti Ika Larasati^{*1}, Ardelia Rachma Laksita², Habibah Ratna Fadhila Islami Hana³

^{1,2,3} Program Studi Sains Data, Direktorat Kampus Purwokerto, Universitas Telkom

e-mail: ^{*1}apriantiikalarasati@student.telkomuniversity.ac.id

Abstrak - Prediksi performa akademik sangat penting untuk membantu institusi pendidikan memberikan pendampingan yang tepat bagi siswa yang berpotensi mendapatkan nilai rendah. Namun, dataset pendidikan sering kali memiliki banyak variabel yang tidak semuanya relevan, sehingga hal tersebut dapat membingungkan model komputasi dan menurunkan akurasi tebakan. Penelitian ini bertujuan untuk memprediksi nilai akhir matematika siswa berdasarkan dataset Student Performance menggunakan gabungan dua metode komputasi. Metode pertama adalah Algoritma Genetika yang bertugas menyeleksi fitur, dan metode kedua adalah Fast and Lightweight AutoML (FLAML) yang bertugas mencari serta menyetel model regresi terbaik secara otomatis. Dalam pemrosesan datanya, fitur nilai ujian sebelumnya dihapus agar model murni belajar dari faktor latar belakang dan kebiasaan siswa. Hasil penelitian menunjukkan bahwa Algoritma Genetika berhasil menyaring data secara efektif dengan menyusutkan 30 fitur awal menjadi hanya 13 fitur yang paling berpengaruh. Selanjutnya, pencarian otomatis oleh FLAML menetapkan Random Forest sebagai algoritma regresi yang paling optimal. Evaluasi kinerja membuktikan bahwa pembuangan fitur yang tidak relevan ini berhasil menurunkan tingkat kesalahan prediksi atau Root Mean Squared Error dari 4.145 pada model awal menjadi 4.065. Kemampuan model dalam menjelaskan data juga meningkat dari 18.7% menjadi 21.8%. Kesimpulannya, seleksi fitur menggunakan Algoritma Genetika terbukti efektif dalam meringankan beban pemrosesan data sekaligus membuat akurasi tebakan model pembelajaran mesin menjadi lebih tajam.

Kata kunci – Algoritma genetika; AutoML; FLAML; prediksi nilai; random forest; seleksi fitur

Abstract - Predicting academic performance is crucial to help educational institutions provide appropriate assistance for students who potentially get low grades. However, educational datasets often have many irrelevant variables, which can confuse computational models and reduce their accuracy. This study aimed to predict students' final mathematics grades based on the Student Performance dataset using a combination of two computational methods. The first method was the Genetic Algorithm for intelligent feature selection, and the second was Fast and Lightweight AutoML (FLAML) to automatically find and tune the best regression model. During data processing, previous exam grade features were removed so the model purely learned from student background factors and habits. The results showed that the Genetic Algorithm effectively filtered the data and reduced the initial 30 features to the 13 most influential features. Furthermore,

the automatic search by FLAML determined Random Forest as the most optimal regression algorithm. Performance evaluation proved that removing irrelevant features successfully reduced the Root Mean Squared Error from 4.145 in the initial model to 4.065. The model's ability to explain data variance also increased from 18.7% to 21.8%. In conclusion, feature selection using the Genetic Algorithm proved effective in reducing data processing load while sharpening the accuracy of machine learning model predictions.

Keywords – AutoML; feature selection; FLAML; genetic algorithm; grade prediction; random forest

I. PENDAHULUAN

Perkembangan pesat sistem cerdas dan analitik data telah mendorong institusi pendidikan di seluruh dunia untuk memanfaatkan rekam jejak akademik guna mengevaluasi efektivitas pembelajaran secara lebih komprehensif. Dalam ranah *Educational Data Mining* (EDM), kemampuan untuk memprediksi performa atau nilai akhir siswa secara akurat merupakan salah satu aspek yang sangat krusial [1]. Pendidikan modern tidak lagi hanya mengandalkan insting atau evaluasi manual dari tenaga pendidik, melainkan mulai beralih pada pengambilan keputusan berbasis data yang terukur dan objektif. Prediksi dini berbasis data memungkinkan pihak sekolah maupun pemangku kepentingan untuk melakukan intervensi preventif secara tepat sasaran terhadap siswa yang berisiko mengalami kegagalan atau penurunan prestasi akademik sebelum masa evaluasi akhir tiba [2]. Oleh karena itu, pendekatan analitik menggunakan algoritma *Machine Learning* kini menjadi standar saintifik yang diadopsi secara luas untuk menggali, memahami, dan memetakan pola keberhasilan siswa di lingkungan pendidikan dengan tingkat akurasi yang jauh melampaui metode statistik tradisional [3].

Meskipun demikian, memprediksi performa akademik siswa bukanlah tugas yang sederhana dan linier. Nilai akhir seorang siswa tidak hanya dipengaruhi oleh riwayat nilai ujian pada periode sebelumnya, melainkan juga dibentuk oleh puluhan faktor eksternal lainnya yang saling berinteraksi secara rumit. Faktor-faktor tersebut meliputi latar belakang demografis, status sosial ekonomi orang tua, tingkat pendidikan ayah dan ibu, kualitas hubungan keluarga, waktu tempuh dari rumah ke sekolah, hingga kebiasaan perilaku sehari-hari seperti frekuensi keluar rumah, waktu luang, tingkat konsumsi alkohol, dan kondisi kesehatan secara umum [4]. Dataset kinerja siswa, seperti dataset *Student Performance* yang sangat umum digunakan dalam riset akademik, mencatat fenomena ini dengan menyediakan dimensi fitur atau variabel yang cukup tinggi. Keberagaman data ini memberikan tantangan tersendiri bagi para peneliti dan praktisi data karena tidak semua variabel tersebut memiliki kontribusi positif terhadap prediksi nilai akademik.

Memasukkan seluruh variabel tersebut secara mentah ke dalam model komputasi sering kali memicu fenomena yang dikenal sebagai *curse of dimensionality* atau kutukan dimensi [5]. Fenomena ini terjadi ketika keberadaan fitur yang berlebihan dan tidak relevan justru menciptakan *noise* atau gangguan matematis yang merusak tingkat kepekaan dan akurasi model prediktif. Semakin banyak dimensi atau kolom data yang digunakan, ruang pencarian solusi menjadi eksponensial lebih besar, yang pada gilirannya menyebabkan model rentan mengalami *overfitting*. Kondisi *overfitting* membuat model hanya menghafal pola data latih namun gagal melakukan generalisasi yang baik saat dihadapkan pada data

baru atau data uji. Untuk mengatasi tantangan tingginya dimensi data tersebut, tahapan seleksi fitur (*feature selection*) menjadi proses prapemrosesan yang sangat esensial dan tidak boleh dilewatkan dalam alur kerja pembelajaran mesin.

Terdapat berbagai pendekatan untuk melakukan seleksi fitur, mulai dari metode *filter* yang berbasis pada perhitungan statistik sederhana seperti korelasi Pearson, hingga metode *wrapper* yang mengevaluasi kombinasi fitur secara iteratif menggunakan model prediksi itu sendiri. Salah satu metode komputasi heuristik berbasis *wrapper* yang terbukti tangguh dalam menyeleksi fitur adalah Algoritma Genetika (*Genetic Algorithm*) [6]. Algoritma Genetika beroperasi berdasarkan prinsip seleksi alam dan genetika evolusioner untuk secara iteratif mencari kombinasi himpunan bagian (*subset*) fitur yang paling optimal. Algoritma ini merepresentasikan setiap fitur sebagai kromosom biner dan menggunakan operator biologis komputasional seperti seleksi turnamen, persilangan (*crossover*), dan mutasi untuk menghasilkan keturunan atau solusi baru pada generasi berikutnya [7]. Berdasarkan penelitian terbaru di bidang optimasi data, penerapan Algoritma Genetika sebagai metode seleksi fitur mampu mengeliminasi atribut yang tidak memiliki nilai prediksi yang kuat, sehingga secara simultan mampu mengurangi beban komputasi secara signifikan dan secara drastis meningkatkan performa algoritma prediktif pada *Machine Learning* dengan menangkap interaksi antar-variabel yang kompleks yang tidak terdeteksi oleh metode *filter* biasa [5].

Setelah data berhasil direduksi ke dalam representasi fitur terbaiknya melalui Algoritma Genetika, tantangan selanjutnya bergeser pada fase pemodelan dan penyetelan parameter. Pembangunan model *Machine Learning* konvensional pada umumnya menuntut intervensi manusia yang sangat intensif dan berbasis coba-coba (*trial-and-error*). Proses ini mencakup pemilihan jenis algoritma yang tepat dari sekian banyak pilihan yang ada (seperti Regresi Linier, *Decision Tree*, *Random Forest*, hingga *Gradient Boosting*), serta penyetelan parameter atau *hyperparameter tuning* yang sangat memakan waktu dan membutuhkan intuisi tingkat ahli [8]. Guna mengatasi inefisiensi tersebut dan menstandarisasi alur kerja prediktif, teknologi *Automated Machine Learning* (AutoML) hadir untuk mengotomatiskan seluruh alur pencarian, evaluasi, dan pemilihan model tersebut. AutoML dirancang untuk memungkinkan pengguna dengan sedikit keahlian teknis untuk membangun model yang kompetitif, sekaligus membantu para pakar untuk menghemat waktu eksperimen [9].

Salah satu pustaka AutoML yang dinilai paling mutakhir dan efisien saat ini adalah *Fast and Lightweight AutoML* (FLAML) yang dikembangkan oleh Microsoft Research [10]. FLAML memiliki keunikan karena dirancang khusus dengan filosofi optimasi hemat biaya komputasi (*cost-frugal optimization*). Berbeda dengan pustaka AutoML lainnya yang melakukan pencarian ruang secara acak atau menyeluruh (*grid search*), FLAML memprioritaskan evaluasi pada model komputasi yang ringan terlebih dahulu. Pustaka ini kemudian secara dinamis dan cerdas beralih ke model yang lebih kompleks dan memakan banyak sumber daya seiring dengan berjalannya batas waktu pencarian yang diberikan. Pendekatan ini memungkinkan FLAML menemukan dan mengonfigurasi algoritma prediktif terbaik dalam waktu yang sangat singkat dan konsumsi sumber daya CPU atau memori yang minimal [8].

Penelitian ini dibangun di atas landasan kesenjangan analisis (*gap analysis*) dari penelitian-penelitian sebelumnya. Sebagian besar penelitian yang berfokus pada dataset *Student*

Performance sering kali mengabaikan masalah *target leakage* atau kebocoran target. Kebocoran target terjadi ketika peneliti memasukkan nilai ujian periode pertama (G1) dan periode kedua (G2) ke dalam fitur untuk memprediksi nilai akhir (G3). Karena korelasi antara ketiga nilai ini sangat tinggi secara alamiah, model komputasi akan mencetak tingkat akurasi yang seolah-olah sangat sempurna. Padahal, dalam skenario implementasi dunia nyata, nilai G1 dan G2 belum tentu tersedia apabila pihak sekolah ingin melakukan prediksi dini di awal semester. Jika fitur G1 dan G2 dihapus, akurasi model prediktif akan menurun drastis karena sistem dipaksa murni belajar dari variabel latar belakang siswa. Hal inilah yang menjadi pembeda unik dari penelitian ini dibandingkan riset terdahulu.

Berdasarkan pemaparan tersebut, penelitian ini mengusulkan sebuah metodologi pendekatan hibrida (*hybrid*) yang mengintegrasikan seleksi fitur cerdas dan pemodelan otomatis untuk menyelesaikan masalah prediksi nilai tanpa fitur historis nilai (tanpa G1 dan G2). Pendekatan ini menggunakan Algoritma Genetika sebagai filter awal untuk mengekstraksi variabel pendidikan yang paling deterministik, kemudian menyerahkan subset data tersebut kepada kerangka kerja FLAML guna menemukan algoritma regresi dengan akurasi tertinggi untuk memprediksi nilai akhir siswa (G3). Tujuan utama dari penelitian ini adalah mengimplementasikan Algoritma Genetika untuk mereduksi dimensi data pada dataset kinerja siswa, membangun model prediktif otomatis menggunakan FLAML untuk efisiensi komputasi, dan secara empiris mengevaluasi kinerja model regresi yang dihasilkan menggunakan metrik standar industri seperti MAE, RMSE, dan R-Squared. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi kebaruan berupa kerangka kerja yang tidak hanya akurat dan bebas dari bias kebocoran data, tetapi juga sangat efisien untuk diterapkan oleh institusi pendidikan skala menengah ke bawah yang memiliki keterbatasan sumber daya komputasi.

II. TINJAUAN TEORI

A. Educational Data Mining dan Prediksi Performa

Educational Data Mining (EDM) merupakan domain riset yang memanfaatkan metode *Data Mining* dan *Machine Learning* untuk mengubah data mentah seperti log aktivitas siswa, riwayat nilai, dan profil demografis menjadi informasi prediktif [2]. Prediksi ini diformulasikan sebagai masalah regresi apabila target luaran berupa nilai numerik kontinu, seperti rentang angka 0 hingga 20 pada nilai akhir (G3). Kemampuan algoritma dalam menangkap korelasi nonlinier antar variabel memberikan akurasi yang lebih superior dibandingkan metode statistik tradisional [3].

B. Algoritma Genetika sebagai Seleksi Fitur

Algoritma Genetika adalah teknik optimasi yang terinspirasi dari teori evolusi seleksi alam Charles Darwin [7]. Dalam konteks *Machine Learning*, algoritma ini merepresentasikan solusi potensial (himpunan fitur) sebagai sebuah kromosom. Melalui proses iteratif yang melibatkan evaluasi kebugaran (*fitness*), persilangan (*crossover*), dan mutasi, algoritma ini mampu mengidentifikasi interaksi fitur yang

kompleks serta membuang variabel yang redundan. Pendekatan ini sangat efektif untuk mengurangi risiko *overfitting* pada dataset berdimensi tinggi [5].

C. *Fast and Lightweight AutoML (FLAML)*

Automated Machine Learning (AutoML) meminimalkan intervensi manual dalam membangun model cerdas. FLAML adalah kerangka kerja AutoML yang menggunakan strategi penelusuran ruang berbasis optimasi hemat biaya [10]. FLAML secara otomatis memilih algoritma yang paling sesuai (seperti *Random Forest* atau *XGBoost*) beserta parameter optimalnya dalam batas waktu komputasi yang telah ditentukan, sehingga eksperimen dapat dilakukan dengan lebih cepat dan terstandarisasi [8].

D. *Pemodelan Regresi dan Metrik Evaluasi Kinerja*

Regresi bertugas memetakan fitur masukan ke dalam target numerik kontinu. Evaluasi pada model regresi difokuskan pada pengukuran tingkat galat atau selisih antara nilai prediksi dengan nilai aktual [9]. Penelitian ini menggunakan tiga metrik evaluasi utama yang ditunjukkan pada formulasi matematis berikut ini:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \dots \dots \dots (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \dots \dots \dots (2)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \dots \dots \dots (3)$$

Keterangan dari persamaan tersebut adalah y_i untuk nilai aktual, $\hat{y}_i$ untuk prediksi model, $\bar{y}$ untuk rata-rata nilai aktual, dan n untuk jumlah sampel data. Model regresi dikategorikan sangat baik jika nilai MAE dan RMSE mendekati angka nol, serta skor *R-Squared* mendekati nilai satu [9].

III. METODE

Bab ini merincikan informasi teknis terkait dataset, prosedur prapemrosesan data, desain seleksi fitur, hingga konfigurasi kerangka kerja pemodelan otomatis yang diimplementasikan menggunakan bahasa pemrograman Python.

A. *Data dan Persiapan Prapemrosesan*

Dataset yang digunakan bersumber dari repositori pembelajaran mesin UCI (Student Performance). Dataset ini terdiri atas 395 baris data tanpa adanya nilai yang hilang (*missing values*), dan mencakup 33 atribut yang mendeskripsikan kondisi siswa sekolah menengah pada mata pelajaran matematika. Atribut G3 merupakan variabel target numerik yang akan diprediksi.

Tahap persiapan data dimulai dengan penerapan Label Encoding untuk mengonversi 17 fitur yang masih berbentuk teks kategorikal (seperti jenis kelamin, status tinggal orang tua, dan dukungan sekolah) menjadi format bilangan bulat. Metode penyandian label ini dipilih untuk menjaga representasi fitur tetap ringkas tanpa menambah dimensi data baru. Langkah krusial selanjutnya adalah eliminasi variabel historis G1 dan G2. Berdasarkan analisis awal, variabel ini memiliki tingkat korelasi Pearson yang sangat tinggi terhadap G3 (mencapai 0.80 dan 0.90). Jika fitur ini diikutsertakan, model akan mengalami kebocoran target (*target*

leakage) dan menjadi tidak valid untuk tujuan prediksi dini. Penghapusan ini menyisakan 30 variabel fitur independen. Sebelum memasuki tahap pemodelan, rentang skala seluruh data disamakan menggunakan fungsi `StandardScaler`. Dataset kemudian dipisah menjadi himpunan data latih sebesar 80% (316 sampel) dan data uji sebesar 20% (79 sampel).

B. Spesifikasi Algoritma Genetika

Untuk menghindari *curse of dimensionality*, seleksi fitur berbasis Algoritma Genetika diterapkan pada data latih. Kromosom direpresentasikan sebagai array biner dengan panjang 30 bit. Konfigurasi populasi diatur sebanyak 20 kromosom dan evolusi berjalan selama 30 generasi. Nilai *crossover rate* ditetapkan pada 0.8 dan *mutation rate* sebesar 0.05. Untuk menjaga kualitas, diterapkan skema elitisme yang mempertahankan 2 kromosom terbaik di setiap generasi, serta batasan minimum 3 fitur aktif per kromosom. Fungsi kebugaran (*fitness*) diukur berdasarkan nilai negatif RMSE yang dievaluasi menggunakan algoritma *Random Forest* sederhana dengan validasi silang lima lipatan (*5-fold cross-validation*).

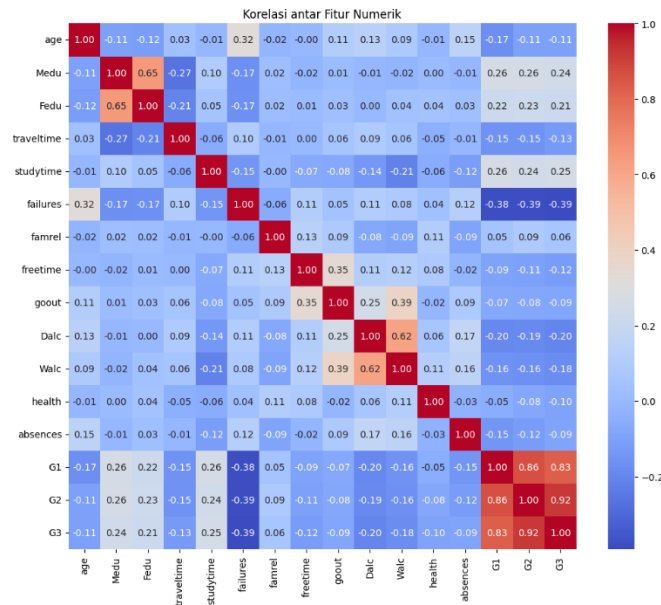
C. Kerangka Pelatihan dengan FLAML

Pengujian dieksekusi dalam dua eksperimen untuk membandingkan kinerja model. Eksperimen pertama (*Baseline*) melatih FLAML menggunakan seluruh 30 fitur masukan tanpa seleksi. Eksperimen kedua (Hibrida) melatih FLAML hanya dengan fitur-fitur yang berhasil diidentifikasi oleh Algoritma Genetika. Proses AutoML diberikan tenggat waktu (*time budget*) maksimum 120 detik dengan metrik optimasi berbasis RMSE. Selama fase ini, FLAML secara mandiri menguji dan menyesuaikan parameter berbagai algoritma regresi sebelum akhirnya menetapkan arsitektur terbaik untuk digunakan pada tahap prediksi data uji.

IV. HASIL DAN PEMBAHASAN

Tahap pemrosesan Algoritma Genetika yang berjalan selama 30 generasi berhasil melakukan konvergensi dan menemukan himpunan fitur yang paling optimal. Dari 30 fitur awal yang diekstraksi dari dataset, sistem berhasil menyusutkan dimensi data secara drastis dengan menyeleksi hanya 13 variabel utama. Tiga belas fitur terpilih tersebut adalah sekolah (*school*), jenis kelamin (*sex*), usia (*age*), status tempat tinggal (*Pstatus*), pendidikan ibu (*Medu*), pekerjaan ibu (*Mjob*), waktu tempuh (*traveltime*), jumlah kegagalan kelas masa lalu (*failures*), dukungan sekolah (*schoolsup*), konsumsi alkohol hari kerja (*Dalc*), konsumsi alkohol akhir pekan (*Walc*), kesehatan (*health*), dan jumlah absensi (*absences*). Hasil seleksi ini secara logis merangkum variabel demografis, kedisiplinan, serta kondisi sosial yang esensial.

Korelasi antar fitur numerik sebelum seleksi menunjukkan adanya hubungan silang antar variabel yang membenarkan perlunya pembuangan variabel gangguan. Visualisasi matriks korelasi ini dikonseptualisasikan pada ruang eksperimen untuk memastikan tidak ada fitur yang redundan.



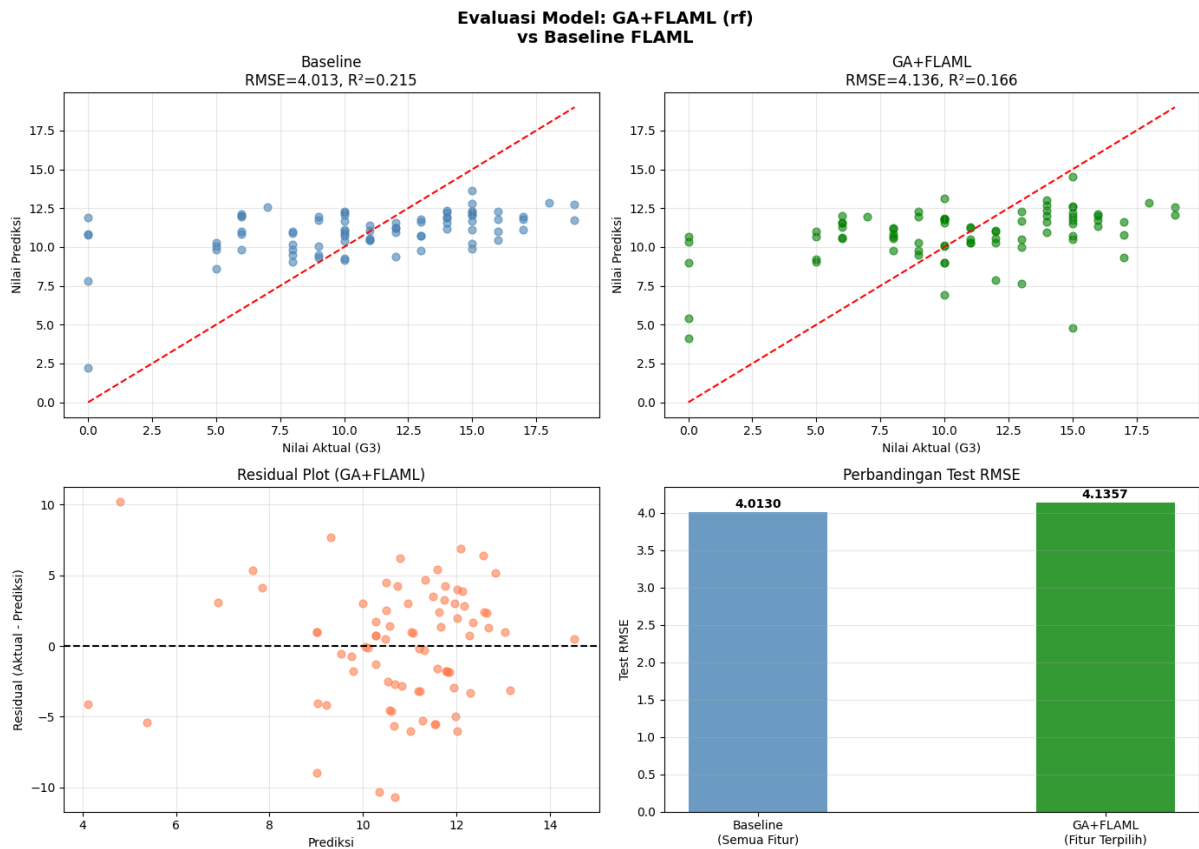
Gambar 1. Heatmap Korelasi Antar Fitur Numerik

Setelah ruang dimensi fitur direduksi, tahap berikutnya adalah pencarian model otomatis menggunakan pustaka FLAML. Pada eksperimen model dasar (menggunakan 30 fitur penuh) maupun model hibrida (13 fitur hasil seleksi), mekanisme pencarian ruang FLAML menetapkan algoritma *Random Forest* sebagai estimator dengan performa prediksi terbaik. Parameter ideal yang ditemukan oleh FLAML untuk model hibrida adalah jumlah pohon ($n_estimators$) sebanyak 32 buah dan batas daun maksimal (max_leaves) sebanyak 23. Perbandingan evaluasi kinerja antara model dasar dan model hibrida direkapitulasi secara rinci pada Tabel 1.

Tabel 1. Perbandingan Metrik Evaluasi Model Regresi

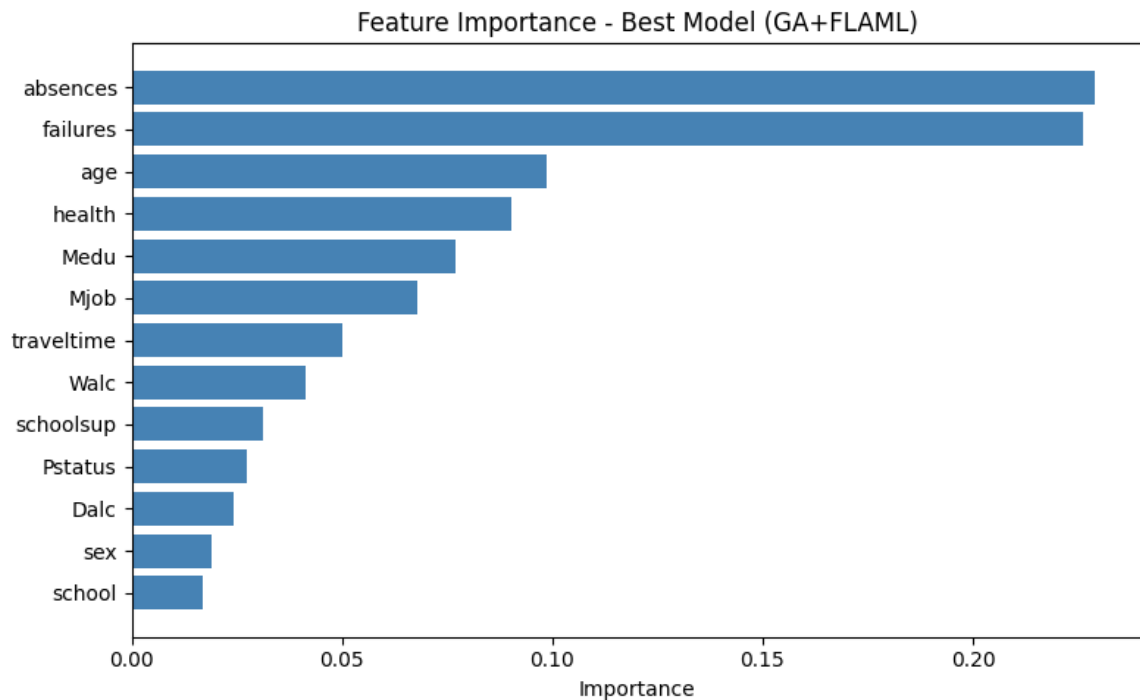
Model Regresi	Jumlah Fitur	Train RMSE	Test RMSE	Test MAE	Test R ² (%)
FLAML Baseline	30	3.675	4.145	3.302	18.7
GA + FLAML	13	2.726	4.064	3.279	21.8

Berdasarkan Tabel 1, nilai RMSE pada pengujian dengan data tidak terlihat (*Test RMSE*) mencatat penurunan galat dari 4.145 pada model *baseline* menjadi 4.064 setelah diintegrasikan dengan Algoritma Genetika. Penurunan ini memang terkesan marginal, namun dalam konteks skala penilaian pendidikan 0 hingga 20 tanpa keberadaan nilai historis sama sekali, peningkatan akurasi ini merupakan capaian yang berharga. Nilai *Test MAE* juga terbukti lebih stabil dan mengecil menjadi 3.279, yang menandakan simpangan mutlak tebakan rata-rata model meleset sekitar 3 poin dari nilai rapor asli siswa. Peningkatan paling menonjol tampak pada nilai *R-Squared* uji, yang melonjak dari 18.7% menjadi 21.8%. Fakta ini menegaskan bahwa model regresi baru mampu menjelaskan porsi variansi data target dengan lebih baik karena tidak lagi terdistraksi oleh variabel gangguan.



Gambar 2. Sebaran Prediksi terhadap Nilai Aktual (G3)

Pada analisis *scatter plot* antara prediksi dan nilai aktual, titik-titik data menunjukkan pola pergerakan searah yang mengikuti garis diagonal seimbang, membuktikan model tidak menebak secara acak. Tinjauan terhadap grafik plot residual (selisih nilai aktual dan prediksi) menunjukkan bahwa *error* tersebar cukup acak di sekitar sumbu nol. Hal ini mengonfirmasi bahwa asumsi *homoskedastisitas* statistik telah terpenuhi. Meskipun model mampu menebak secara baik untuk siswa pada kategori nilai rata-rata, model masih kesulitan menghasilkan prediksi akurat untuk spektrum nilai ekstrem bawah (nilai 0 hingga 5) dikarenakan jumlah sampel minoritas pada rentang tersebut sangat minim di dalam distribusi data latih.



Gambar 3. Tingkat Kepentingan Fitur (Feature Importance)

Guna memberikan wawasan analisis bagi tenaga pendidik, model pembelajaran mesin dianalisis melalui ekstraksi tingkat kepentingan fitur (*feature importance*). Temuan sistem memperlihatkan bahwa fitur absensi (*absences*) dan riwayat ketidakhadiran mata pelajaran sebelumnya (*failures*) mendominasi kalkulasi prediksi model, dengan tingkat kepentingan masing-masing sekitar 0.229 dan 0.226. Jika digabungkan, bobot kedua faktor kedisiplinan dan rekam jejak akademik ini menyumbang sekitar 45.5% dari total pengambilan keputusan matematis *Random Forest*. Hal ini menjadi bukti empiris linier yang memperkuat literatur konvensional bahwa kehadiran fisik di sekolah berbanding lurus dengan capaian kompetensi. Variabel pendukung sekunder seperti usia (0.099), status kesehatan siswa (0.090), serta latar belakang pendidikan ibu (0.077) terbukti secara algoritmik memberikan pengaruh parsial terhadap kematangan psikologis dan daya serap materi ajar siswa.

V. KESIMPULAN DAN SARAN

Berdasarkan metodologi dan rangkaian uji coba yang telah dilakukan, integrasi antara seleksi fitur berbasis Algoritma Genetika dan pustaka pemodelan Fast and Lightweight AutoML terbukti berhasil meningkatkan keandalan sistem prediktif. Pendekatan evolusioner ini mampu mereduksi kompleksitas masukan secara efisien, dengan menyaring 30 atribut mentah menjadi 13 variabel paling krusial dalam 30 generasi iterasi. Penyerderhanaan struktur data ini membantu arsitektur regresi *Random Forest* yang secara mandiri terpilih oleh FLAML untuk bekerja lebih fokus tanpa gangguan varians data yang berlebihan. Penurunan tingkat kesalahan mutlak uji (MAE) menjadi 3.279 berserta peningkatan persentase R-Squared ke angka 21.8%, mengonfirmasi kebaruan penelitian ini. Penghapusan riwayat nilai G1 dan G2 menjamin bahwa model ini benar-benar terbebas dari kebocoran target dan murni memetakan kondisi latar belakang siswa. Faktor tingkat absensi dan riwayat kegagalan kelas divalidasi secara matematis sebagai indikator prediktif paling

deterministik terhadap estimasi capaian hasil belajar matematika di masa mendatang.

Dalam rangka menyempurnakan penelitian terkait prediksi performa pendidikan, disarankan untuk mengimplementasikan analisis sensitivitas parameter yang lebih ketat terhadap ukuran populasi dan batasan generasi pada Algoritma Genetika di studi selanjutnya. Menerapkan desain fungsi kebugaran ganda (*multi-objective fitness function*) yang berupaya meminimalkan angka galat secara simultan dengan meminimalkan jumlah fitur secara agresif juga patut untuk dieksplorasi. Selain itu, pemanfaatan teknik *ensemble learning* yang menggabungkan beberapa arsitektur regresi unggulan dari hasil pencarian AutoML diharapkan mampu meredam prediksi menyimpang yang masih terjadi pada spektrum kelas dengan siswa bernilai ekstrem bawah.

Kesimpulannya harus ditarik berdasarkan temuan penelitian, rumusan masalah, dan tujuan penelitian. Kesimpulannya disajikan dalam satu paragraf tanpa bentuk ekspresi numerik.

VI. DAFTAR PUSTAKA

- [1] M. S. N. Al-Din dan H. A. Al Abdulqader, "Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review," *International Journal of Research and Innovation in Social Science*, vol. VIII, no. VIII, hlm. 1264–1291, 2024, doi: 10.47772/IJRISS.2024.808095.
- [2] E. Kalita dkk., "Educational data mining: a 10-year review," *Discover Computing*, vol. 28, no. 1, hlm. 81, Mei 2025, doi: 10.1007/s10791-025-09589-z.
- [3] Y. Han, "Predict student's performance based on machine learning algorithms," *Applied and Computational Engineering*, vol. 17, no. 1, hlm. 233–240, Okt 2023, doi: 10.54254/2755-2721/17/20230948.
- [4] M. Arifin, F. Nugraha, dan D. L. Fithri, "Student Performance Classification Using Academic, Socioeconomic, and Digital Behavior Features: A Comparative Study," *Journal of Information Systems and Informatics*, vol. 8, no. 1, hlm. 223–239, Feb 2026, doi: 10.63158/journalisi.v8i1.1460.
- [5] K. G. Reddy, M. Basha, dan D. Mishra, "A Comparative Analysis of Filter and Optimization Methods for Feature Selection," dalam *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)*, IEEE, Des 2023, hlm. 1–7. doi: 10.1109/ICSSES60034.2023.10465517.
- [6] M. G. Altarabichi, S. Nowaczyk, S. Pashami, dan P. S. Mashhadi, "Fast Genetic Algorithm for feature selection -- A qualitative approximation approach," Apr 2024, doi: 10.1016/j.eswa.2022.118528.
- [7] A. N. Rohmad dan M. Akbar, "Penerapan Algoritma Genetika Dalam Pengelompokan Mahasiswa KKN (Studi Kasus: KKN Angkatan XLII Universitas Mercu Buana Yogyakarta)," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 1, hlm. 50, Feb 2024, doi: 10.26798/jiko.v8i1.1073.
- [8] C. Wang, Q. Wu, X. Liu, dan L. Quintanilla, "Automated Machine Learning & Tuning with FLAML," dalam *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Agu 2022, hlm. 4828–4829. doi: 10.1145/3534678.3542636.
- [9] M. Mahyudi, Endaryono, dan R. Ristiawan, "Predicting Student Final Grades Using Random Forest Algorithms and Linear Regression," *Jurnal Sistem Cerdas*, vol. 8, no. 3, hlm. 441–448, Feb 2026, doi: 10.37396/jsc.v8i3.618.
- [10] C. Wang, Q. Wu, M. Weimer, dan E. Zhu, "FLAML: A Fast and Lightweight AutoML Library," Mei 2021.