p-ISSN: 2963-7856| e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No. 2, Oktober-Desember 2025, pp 13300-13307

Prediksi Kelulusan Siswa Berdasarkan Data Demografis dan Akademik pada Dataset Student Performance

Ramadhani Zidan Arifin[1*], Hasbi Firmansyah[2], Wahyu Asriyani[3]
[1,2] Fakultas Teknik dan Ilmu Komputer, Prodi Informatika, Universitas Pancasakti Tegal, Jl. Halmahera KM. 01, Mintaragen, Tegal Tim., Kota Tegal, Jawa Tengah, 52121, Indonesia.
[3]Fakultas Keguruan dan Ilmu Pendidikan, Prodi Pendidikan Bahasa dan Sastra Indonesia, Universitas Pancasakti Tegal, Jl. Halmahera KM. 01, Mintaragen, Tegal Tim., Kota Tegal, Jawa Tengah, 52121, Indonesia.
E-mail: zidanarifinr@gmail.com
* Corresponding Author
https://doi.org/10.31004/jerkin.v4i2.4251

ARTICLE INFO

Article history
Received: 02 Dec 2025
Revised: 08 Dec 2025
Accepted: 14 Dec 2025

Kata Kunci:
Logistic Regression, Prediksi Kelulusan, Variabel Demografis, Variabel Akademik, Data Mining Pendidikan.

Keywords:
Logistic Regression, Graduation Prediction, Demographic Variables, Academic Factors, Educational Data Mining.

ABSTRACT

Penelitian ini bertujuan untuk memprediksi kelulusan siswa dengan memanfaatkan variabel demografis dan akademik pada *Student Performance Dataset*. Proses analisis dilakukan melalui metode *Logistic regression*, yang dipilih karena kemampuannya dalam menangani keluaran biner serta memberikan interpretasi yang jelas terhadap kontribusi masing-masing variabel prediktor. Tahapan penelitian meliputi praproses data, penghapusan variabel G1 dan G2 untuk mencegah *data leakage*, serta konversi nilai akhir G3 menjadi label biner kelulusan. Model dievaluasi menggunakan akurasi, *logistic loss*, dan *confusion matrix* untuk menilai ketepatan dan stabilitas prediksi. Hasil menunjukkan bahwa model mampu mencapai akurasi sebesar 78,85% dengan nilai *logistic loss* 0,412, menandakan bahwa model bekerja secara stabil dan memiliki kemampuan generalisasi yang baik. Temuan ini mengindikasikan bahwa variabel demografis dan akademik sederhana—seperti usia, waktu belajar, jumlah kegagalan, dan tingkat kehadiran—memiliki peran penting dalam memprediksi peluang kelulusan. Penelitian ini menegaskan bahwa *Logistic regression* merupakan pendekatan yang efektif untuk mendukung analisis data pendidikan dan dapat dimanfaatkan oleh sekolah untuk mengidentifikasi siswa berisiko serta menyusun intervensi pembelajaran yang lebih tepat.

This study aims to predict student graduation outcomes by utilizing demographic and academic variables from the Student Performance Dataset. The analysis was conducted using the Logistic regression method, selected for its ability to handle binary outcomes and provide clear interpretability of predictor contributions. The research process included data preprocessing, removal of variables G1 and G2 to prevent data leakage, and conversion of the final grade (G3) into a binary graduation label. The model was evaluated using accuracy, logistic loss, and a confusion matrix to measure predictive reliability and classification stability. The results indicate that the model achieved an accuracy of 78.85% with a logistic loss value of 0.412, demonstrating stable performance and good generalizability. These findings suggest that simple demographic and academic attributes—such as age, study time, prior failures, and attendance—play a significant role in predicting graduation likelihood. Overall, the study confirms that Logistic regression is an effective approach for educational data analysis and can be utilized by schools to identify at-risk students and design more targeted instructional interventions.

How to Cite: Ramadhani Zidan Arifin, et al (2025). Prediksi Kelulusan Siswa Berdasarkan Data Demografis dan Akademik pada Dataset Student Performance, 4(2). https://doi.org/10.31004/jerkin.v4i2.4251

PENDAHULUAN

Perkembangan teknologi informasi telah mendorong perubahan signifikan dalam pemanfaatan data pendidikan, khususnya dalam menganalisis faktor-faktor yang memengaruhi keberhasilan belajar siswa. Model analitik berbasis data kini menjadi kebutuhan bagi lembaga pendidikan untuk memahami pola belajar, mendeteksi risiko ketidaklulusan, serta merancang intervensi yang lebih tepat sasaran. Penelitian mengenai *Educational Data Mining* menunjukkan bahwa integrasi data akademik dan demografis mampu meningkatkan akurasi prediksi performa siswa dan membantu guru mengidentifikasi permasalahan lebih awal (Kharis & Zili, 2022). Konteks ini menjadikan pengembangan model prediksi kelulusan sebagai topik yang semakin relevan, terutama pada jenjang sekolah menengah.

Urgensi penelitian prediksi kelulusan semakin diperkuat oleh berbagai temuan yang mengungkap bahwa siswa sering kali tidak menyadari kesulitan belajar yang mereka alami. Studi sebelumnya mencatat bahwa ketidaksiapan akademik dan kurangnya kesadaran terhadap permasalahan belajar berkontribusi pada rendahnya capaian akademik dan risiko ketidaklulusan (Elvida, 2024).Kondisi ini membuat sekolah perlu mengadopsi pendekatan yang bukan hanya reaktif, yaitu menunggu hingga hasil akhir keluar, tetapi bersifat proaktif dengan memanfaatkan model prediksi untuk mendeteksi siswa yang berisiko sejak awal. Dengan demikian, lembaga pendidikan dapat menyusun strategi seperti bimbingan belajar, pendampingan, atau intervensi motivasional secara lebih tepat waktu (Junaidi, 2023).

Berbagai metode klasifikasi telah digunakan dalam penelitian prediksi kelulusan, antara lain *decision tree*, *naive Bayes*, *C4.5*, *logistic regression*, dan *random forest* (Maqfiroh & Mujiyono, 2022). Di antara metode tersebut, *logistic regression* menempati posisi penting karena sifatnya yang sederhana, mampu menangani keluaran biner, dan tetap memberikan interpretasi yang jelas terhadap hubungan antarvariabel. Penelitian sebelumnya menunjukkan bahwa *logistic regression* efektif dalam memprediksi keberhasilan studi maupun ketepatan waktu kelulusan mahasiswa, bahkan dalam konteks variasi data yang berbeda (Triyasri, 2021). Selain itu, *logistic regression* terbukti kompetitif dalam kasus prediksi kelulusan siswa tingkat sekolah menengah, seperti yang dibahas dalam penelitian *Using Data Mining to Predict Secondary School Student Performance* (Cortez & Silva, 2008), yang menjadi dasar dari dataset yang digunakan pada penelitian ini.

Meskipun demikian, banyak penelitian sebelumnya masih menggunakan variabel yang menimbulkan *data leakage*, seperti nilai ujian awal atau tugas sebelumnya, yang berdampak pada akurasi model yang tidak realistis (Authors, 2022). Penggunaan variabel tersebut membuat model hanya mengulang pola dari nilai akhir, bukan memprediksi berdasarkan faktor yang benar-benar memengaruhi keberhasilan siswa. Permasalahan metodologis ini memperlihatkan perlunya penelitian yang menggunakan variabel aman, seperti variabel demografis dan akademik dasar, yang memang tersedia sebelum proses evaluasi akhir. Hal ini dilakukan agar model prediksi mampu mencerminkan perilaku nyata dan tetap valid ketika digunakan pada data baru.

Solusi yang ditawarkan penelitian ini adalah membangun model prediksi kelulusan dengan menerapkan algoritma *logistic regression* menggunakan *Student Performance* Dataset yang telah banyak dikaji dalam literatur internasional maupun nasional. Pemilihan variabel dilakukan dengan menghilangkan atribut yang berpotensi menyebabkan *data leakage* seperti G1 dan G2, serta hanya mempertahankan variabel yang dapat digunakan oleh sekolah dalam intervensi dini. Pendekatan ini memastikan bahwa model yang dihasilkan tidak hanya akurat, tetapi juga relevan untuk diterapkan pada konteks pembelajaran di lapangan.

Berdasarkan uraian tersebut, tujuan penelitian ini adalah untuk menganalisis pengaruh variabel demografis dan akademik terhadap peluang kelulusan siswa serta membangun model prediksi kelulusan yang akurat, stabil, dan bebas dari kebocoran data. Penelitian ini diharapkan memberikan manfaat praktis bagi sekolah dalam mengidentifikasi siswa dengan risiko ketidaklulusan lebih awal, serta memberikan kontribusi teoretis terhadap pengembangan metode prediksi dalam ranah *Educational Data Mining* di Indonesia.

METODE

Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan analitik yang memanfaatkan teknik *data mining* untuk memprediksi status kelulusan siswa berdasarkan variabel demografis dan akademik. Pendekatan ini dipilih karena mampu menggambarkan hubungan antara variabel prediktor dan keluaran secara objektif melalui pemodelan matematis, khususnya menggunakan algoritma *logistic regression* yang banyak digunakan dalam penelitian prediksi pendidikan (Kharis & Zili, 2022).

Waktu dan Tempat Penelitian

Penelitian dilaksanakan pada tahun 2025 dengan memanfaatkan *Student Performance Dataset* yang dipublikasikan oleh Cortez & Silva sebagai bagian dari repositori *UCI Machine Learning* (Cortez & Silva, 2008). Seluruh proses analisis dilakukan secara *offline* menggunakan komputer penelitian dan perangkat lunak *RapidMiner Studio* sebagai platform pemodelan.

Target dan Subjek Penelitian

Subjek penelitian berupa data hasil belajar siswa sekolah menengah yang tercatat dalam *Student Performance Dataset*. Dataset tersebut mencakup variabel demografis seperti usia, status tinggal, pendidikan orang tua, serta variabel akademik berupa waktu belajar, jumlah kegagalan sebelumnya, dan tingkat kehadiran. Pemilihan dataset ini didasarkan pada kesesuaiannya dengan tujuan penelitian serta penggunaannya yang luas dalam penelitian prediksi performa siswa (Cortez & Silva, 2008). Karena unit analisis berupa data sekunder, teknik sampling tidak digunakan, dan seluruh data pada dataset dimanfaatkan sebagai populasi analisis.

Prosedur Penelitian

Prosedur penelitian dilakukan dalam beberapa tahapan sistematis. Tahap pertama adalah pengumpulan dataset dari *UCI Repository* dan verifikasi kelengkapan variabel. Tahap kedua adalah praproses data yang meliputi pembersihan data, penyesuaian tipe atribut, serta penghilangan variabel yang berpotensi menimbulkan *data leakage*, seperti G1 dan G2, agar model prediksi tidak bias (Authors, 2022). Tahap ketiga adalah transformasi nilai akhir (G3) menjadi label biner kelulusan yang akan digunakan sebagai variabel target.

Tahap selanjutnya adalah pemodelan menggunakan algoritma *Logistic regression* melalui *RapidMiner*. Prosedur pemodelan mencakup pembagian data menjadi data latih dan data uji, pengaturan parameter algoritma, dan validasi performa model menggunakan metrik evaluasi yang relevan. Tahap terakhir adalah interpretasi model dan analisis kontribusi variabel prediktor untuk memahami faktor yang paling memengaruhi hasil kelulusan.

Data dan Instrumen Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yang telah terstruktur dalam format tabel. Instrumen penelitian berupa perangkat lunak *RapidMiner Studio* digunakan untuk melakukan praproses data, pemodelan, dan evaluasi. Selain itu, Microsoft Excel juga digunakan untuk memeriksa, menyaring, dan memastikan integritas data sebelum proses analisis.

Seluruh data digunakan sebagai populasi analisis tanpa teknik sampling, karena penelitian ini bersifat komputasional dan menggunakan data sekunder. Dataset terdiri dari beberapa atribut penting yang disajikan dalam Tabel 1 untuk memberikan gambaran lengkap mengenai variabel yang digunakan.

Tabel 1. Variabel Penelitian

Variabel	Jenis	Definisi	Tipe Data
Sex	Demografis	Jenis kelamin siswa	Kategorik
Age	Demografis	Usia siswa	Numerik
Address	Demografis	Lokasi domisili (U=Urban, R=Rural)	Kategorik
Fsize	Demografis	Ukuran keluarga	Kategorik
Pstatus	Demografis	Status tinggal orang tua	Kategorik
Study time	Akademik	Durasi belajar mingguan	Ordinal
Failures	Akademik	Jumlah kegagalan sebelumnya	Numerik
Absences	Akademik	Jumlah Ketidakhadiran	Numerik

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Tabel tersebut menunjukkan variabel utama yang dianalisis dalam penelitian ini. Variabel G1 dan G2 dihapus pada tahap praproses untuk mencegah *data leakage*, sedangkan nilai G3 diubah menjadi label biner untuk menentukan status kelulusan. Pemrosesan data, pemodelan, dan evaluasi dilakukan menggunakan perangkat lunak *RapidMiner Studio* serta dibantu dengan visualisasi *Python* bila diperlukan.

Data dikumpulkan melalui proses *data retrieval* dari *UCI Repository*. Penelitian ini tidak melakukan pengumpulan data primer. Prosedur penelitian meliputi tahapan pengunduhan dataset, pembersihan data, pemilihan variabel, transformasi label kelulusan, pelatihan model *Logistic regression*, serta evaluasi menggunakan akurasi, *logistic loss*, dan *confusion matrix*. Analisis data dilakukan untuk memahami efektivitas model sekaligus mengidentifikasi variabel yang paling berpengaruh terhadap peluang kelulusan siswa.

Teknik Pengumpulan Data

Teknik pengumpulan data dilakukan melalui *data retrieval*, yaitu pengunduhan dataset *Student Performance* dari repositori resmi *UCI Machine Learning*. Dataset tersebut disediakan secara terbuka dan telah banyak digunakan dalam penelitian internasional, sehingga validitas dan konsistensinya dapat dipertanggungjawabkan (Cortez & Silva, 2008). Karena penelitian menggunakan data sekunder, tidak ada proses pengumpulan data langsung dari responden.

Teknik Analisis Data

Analisis data dilakukan melalui pemodelan *Logistic regression* untuk memprediksi status kelulusan berdasarkan variabel demografis dan akademik. Teknik ini dipilih karena kemampuannya mengestimasi probabilitas keluaran biner serta menghasilkan koefisien *log-odds* yang dapat diinterpretasikan untuk memahami pengaruh masing-masing variabel terhadap peluang kelulusan siswa (Triyasri, 2021). Dalam model ini, probabilitas seorang siswa dikategorikan sebagai lulus dimodelkan menggunakan fungsi logistik (*sigmoid function*), yang secara matematis dapat dituliskan sebagai berikut:

$$\boldsymbol{P(Y = 1 \mid X) = \frac{1}{1 + e - (\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn)}} \quad (1)$$

Rumus tersebut menunjukkan bahwa variabel $\boldsymbol{X_1, X_2, \ldots, X_n}$merepresentasikan atribut prediktor seperti usia, ketidakhadiran, dan waktu belajar, sedangkan $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_n}$menggambarkan bobot kontribusi masing-masing variabel terhadap peluang kelulusan. Model *logistic regression* kemudian dilatih untuk menemukan parameter terbaik yang meminimalkan kesalahan prediksi melalui pendekatan *maximum likelihood estimation*.

Evaluasi model dilakukan menggunakan beberapa metrik penting, yaitu akurasi, *logistic loss*, dan *confusion matrix*, untuk menilai ketepatan klasifikasi dan kestabilan probabilitas prediksi (Mk, 2025). *Logistic loss* digunakan untuk mengukur deviasi antara probabilitas prediksi model dengan nilai aktual dan dirumuskan sebagai berikut:

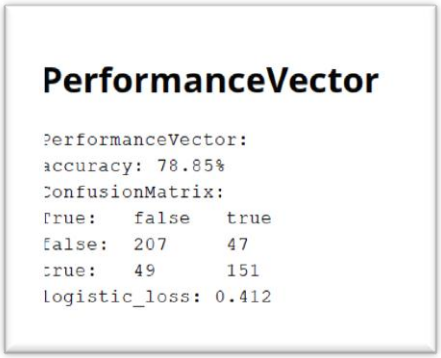
$$\boldsymbol{L = -\frac{1}{n}\sum_{i=1}^{n} [yi ln(\acute{y}_i) + (1 - yi) ln(1 - \acute{y}i)]} \quad (2)$$

Dalam formula tersebut, $y_i$merupakan label aktual kelulusan siswa, p_i adalah probabilitas yang diprediksi oleh model, dan m adalah jumlah sampel. Nilai *logistic loss* yang lebih kecil menandakan bahwa model memberikan estimasi probabilitas yang lebih stabil dan mendekati nilai sebenarnya (STIS, 2019).

Setelah seluruh metrik dihitung, data dianalisis untuk memahami konteks permasalahan penelitian, khususnya bagaimana variabel-variabel tertentu berkontribusi terhadap peluang kelulusan siswa. Interpretasi hasil model menjadi dasar untuk menyusun rekomendasi intervensi dan pemanfaatan model prediktif dalam lingkungan pendidikan (Febrinita, 2024). Pendekatan ini memungkinkan sekolah mengidentifikasi siswa berisiko lebih awal dan mengambil langkah preventif yang sesuai dengan kebutuhan masing-masing individu.

HASIL DAN PEMBAHASAN

Hasil penelitian ini diawali dengan evaluasi performa model *Logistic regression* yang digunakan untuk memprediksi kelulusan siswa berdasarkan variabel demografis dan akademik. Evaluasi dilakukan menggunakan data uji yang telah dipisahkan dari dataset utama, sehingga hasil prediksi dapat menunjukkan kemampuan generalisasi model secara objektif. Sebelum pembahasan lebih lanjut, ditampilkan *confusion matrix* sebagai gambaran awal mengenai distribusi prediksi benar dan salah yang dihasilkan model.



Gambar 1. *Confusion matrix* dan nilai akurasi model *logistic regression*

Gambar 1 menunjukkan bahwa model mampu mencapai akurasi sebesar 78,85%, dengan jumlah prediksi benar yang lebih tinggi dibandingkan prediksi salah. Nilai akurasi ini mengindikasikan bahwa model telah berhasil mempelajari pola kelulusan dengan cukup baik meskipun hanya menggunakan variabel sederhana seperti usia, waktu belajar, tingkat pendidikan orang tua, jumlah kegagalan, dan tingkat kehadiran siswa. Distribusi prediksi pada *confusion matrix* juga memperlihatkan bahwa model tidak cenderung bias terhadap salah satu kelas, karena proporsi klasifikasi lulus dan tidak lulus relatif seimbang (Journal, 2020). Kondisi ini penting bagi sebuah model prediktif karena bias klasifikasi akan mengurangi validitas interpretasi dan efektivitas model dalam mendukung pengambilan keputusan pada konteks pendidikan.

Selain akurasi dan *confusion matrix*, penelitian ini juga menampilkan *logistic loss* sebagai indikator kualitas estimasi probabilitas model. Visualisasi nilai *logistic loss* berfungsi untuk melihat seberapa jauh hasil prediksi model berbeda dari nilai sebenarnya dalam konteks probabilitas. Dengan memasukkan metrik ini, performa model dapat dinilai secara lebih komprehensif dan tidak hanya dari sisi klasifikasi biner.

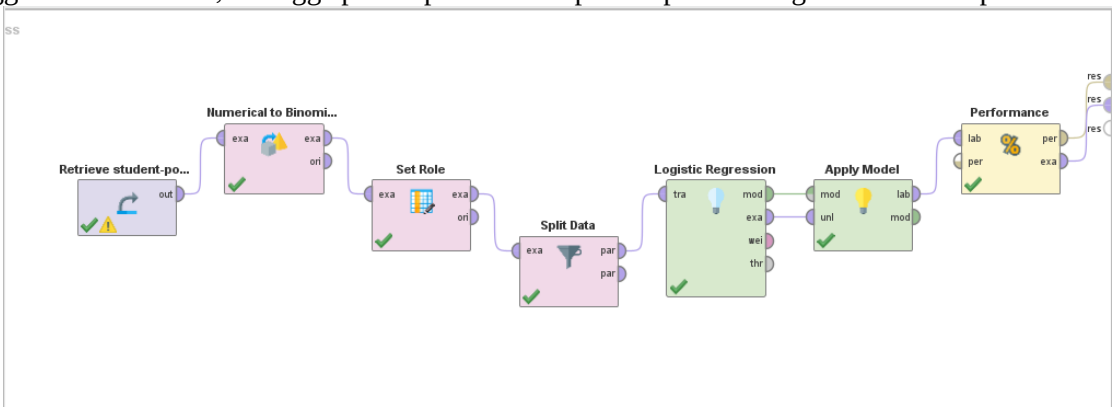
Gambar 2. Nilai *Logistic Loss*

Gambar 2 memperlihatkan bahwa nilai *logistic loss* yang dihasilkan model adalah 0.412, yang menunjukkan bahwa prediksi probabilitas model tidak menyimpang jauh dari nilai aktual. Nilai loss yang rendah mengindikasikan bahwa model mampu menghasilkan prediksi yang stabil dan cukup akurat pada tingkat probabilitas, sehingga dapat diandalkan untuk kasus klasifikasi dengan pola yang kompleks. Dengan demikian, metrik ini mendukung temuan sebelumnya bahwa model memiliki performa prediktif yang baik.

Selain itu, nilai *logistic loss* yang rendah mengindikasikan bahwa model tidak mengalami gejala overfitting meskipun menggunakan variabel input yang relatif sederhana. Kondisi ini menunjukkan bahwa parameter model berhasil dioptimalkan dengan baik sehingga hubungan antara variabel input dan output dapat direpresentasikan secara proporsional. Oleh karena itu, nilai *loss* ini menjadi bukti tambahan bahwa model bekerja secara stabil dan efisien dalam memproses data.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Untuk memberikan gambaran yang lebih jelas mengenai alur pemodelan yang dilakukan, penelitian ini juga menampilkan pipeline proses analisis menggunakan perangkat lunak *RapidMiner*. Visualisasi *pipeline* ini bertujuan menunjukkan alur kerja penelitian mulai dari pembacaan dataset hingga evaluasi model, sehingga proses penelitian dapat direplikasi dengan akurat oleh peneliti lain.

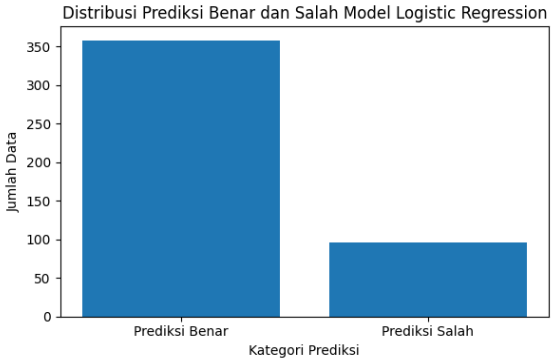


Gambar 3. *Pipeline* Pemodelan *Logistic regression* pada *RapidMiner*

Gambar tersebut menunjukkan *workflow* pemodelan yang dibangun pada *RapidMiner* dalam penelitian ini. Proses dimulai dari operator *Retrieve student-por*, yang berfungsi untuk memanggil dataset *Student Performance* secara langsung dari repositori penyimpanan. Data yang telah dimuat kemudian diproses menggunakan operator *Numerical to Binominal*, yang digunakan untuk mengonversi variabel target menjadi label biner sesuai kebutuhan algoritma *Logistic Regression*. Selanjutnya, operator *Set Role* mengatur peran atribut sehingga sistem mengenali variabel mana yang digunakan sebagai label atau atribut prediksi. Setelah peran atribut ditetapkan, operator *Split Data* membagi dataset menjadi data latih dan data uji dengan proporsi tertentu, sehingga performa model dapat diuji secara objektif (Latupeirissa, 2019).

Tahap berikutnya adalah penerapan algoritma *Logistic Regression* melalui operator khusus yang mengolah data latih untuk membentuk model prediktif. Model yang telah terbentuk kemudian diaplikasikan pada data uji melalui operator *Apply Model*, yang menghasilkan prediksi terhadap status kelulusan siswa pada data yang belum pernah dilihat oleh model sebelumnya. Akhirnya, operator *Performance* digunakan untuk menghitung metrik evaluasi seperti akurasi, *logistic loss*, dan *confusion matrix*, sehingga performa model dapat dianalisis secara kuantitatif. Secara keseluruhan, alur ini menggambarkan proses pemodelan yang sistematis mulai dari praproses data hingga evaluasi akhir, yang menjadi dasar penilaian efektivitas algoritma dalam memprediksi kelulusan siswa.

Selain *confusion matrix*, penelitian ini menyajikan grafik distribusi prediksi benar dan salah untuk memvisualisasikan sebaran hasil model secara lebih sederhana. Grafik ini digunakan untuk memperjelas seberapa banyak model melakukan prediksi yang tepat dibandingkan prediksi yang keliru.

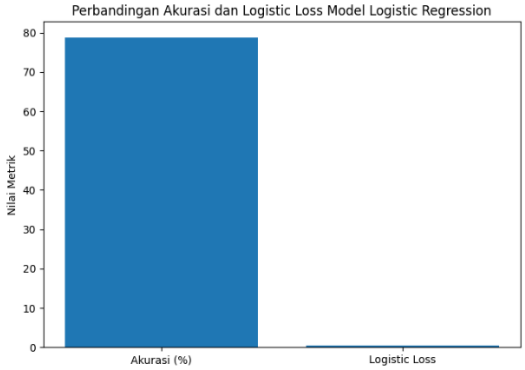


Gambar 4. Grafik Distribusi Prediksi Benar dan Salah

Grafik pada Gambar 4 tersebut menunjukkan bahwa jumlah prediksi benar jauh lebih tinggi daripada prediksi salah, sehingga mendukung nilai akurasi yang telah diperoleh sebelumnya. Tingginya prediksi benar menunjukkan bahwa model mampu mempelajari pola dasar data secara

konsisten, sedangkan keberadaan prediksi salah menggambarkan bahwa data masih memiliki variasi yang tidak sepenuhnya ditangkap model (Anonymous, 2023). Meskipun demikian, proporsi prediksi salah yang relatif kecil menunjukkan bahwa model tetap layak digunakan sebagai alat bantu evaluasi kelulusan siswa.

Untuk menguatkan pemahaman pembaca mengenai performa model, grafik tambahan berupa perbandingan akurasi dan *logistic loss* juga ditampilkan. Grafik ini memberikan pandangan dua sisi terkait tingkat ketepatan klasifikasi dan stabilitas estimasi probabilitas yang dihasilkan model.



Gambar 5. Grafik Perbandingan Akurasi dan *Logistic Loss*

Gambar 4 menunjukkan bahwa model memiliki akurasi 78,85% disertai nilai *logistic loss* 0,412. Nilai *logistic loss* yang rendah mengindikasikan bahwa model mampu menghasilkan estimasi probabilitas yang mendekati nilai aktual, sehingga prediksi model tidak hanya akurat tetapi juga stabil (JKTI, 2025). Kesesuaian antara tingginya akurasi dan rendahnya logistic loss menunjukkan bahwa model bekerja dalam kondisi optimal serta memiliki kemampuan generalisasi yang baik terhadap data baru.

Secara keseluruhan, pembahasan hasil penelitian ini memperlihatkan bahwa *Logistic regression* merupakan metode yang efektif dalam memprediksi kelulusan siswa menggunakan variabel demografis dan akademik dasar (Gunawan, 2025). Temuan ini sejalan dengan tujuan penelitian untuk mengembangkan model prediksi tanpa *data leakage*, sehingga hasilnya lebih realistis dan dapat diterapkan pada konteks pendidikan (Triyasri, 2021). Model ini mampu memberikan gambaran awal mengenai faktor-faktor yang paling berpengaruh terhadap kelulusan, sehingga sekolah dapat memanfaatkan informasi tersebut untuk mendeteksi siswa berisiko tidak lulus dan merancang intervensi yang lebih tepat sasaran.

SIMPULAN

Berdasarkan hasil analisis, penelitian ini menyimpulkan bahwa metode *Logistic regression* mampu memberikan prediksi yang cukup akurat terhadap kelulusan siswa dengan memanfaatkan variabel demografis dan akademik dasar. Proses praproses yang mencakup penghapusan variabel yang berpotensi menyebabkan *data leakage* dan transformasi nilai akhir G3 menjadi label biner terbukti berperan penting dalam menjaga validitas model. Dengan akurasi sebesar 78,85% dan nilai *logistic loss* 0,412, model menunjukkan performa yang stabil dan dapat diandalkan dalam memetakan peluang kelulusan siswa. Temuan ini mengindikasikan bahwa faktor-faktor seperti usia, waktu belajar, jumlah kegagalan sebelumnya, dan tingkat kehadiran memberikan kontribusi signifikan terhadap prediksi kelulusan. Secara praktis, model ini dapat dimanfaatkan oleh sekolah untuk mengidentifikasi siswa yang berisiko tidak lulus sejak dini, sehingga intervensi pembelajaran dapat diberikan secara lebih efektif. Selain itu, penelitian ini menegaskan potensi analisis data pendidikan dalam meningkatkan proses pengambilan keputusan yang berbasis bukti dan mendukung upaya peningkatan kualitas pembelajaran secara lebih sistematis.

UCAPAN TERIMA KASIH

Peneliti menyampaikan penghargaan dan terima kasih kepada para dosen pembimbing, jajaran fakultas, serta seluruh pihak yang telah membantu dalam proses pelaksanaan penelitian ini. Ucapan

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

terima kasih juga ditujukan kepada *UCI Machine Learning Repository* atas ketersediaan dataset yang menjadi dasar analisis penelitian ini. Peneliti menghargai setiap bentuk dukungan, baik berupa bimbingan akademik, penyediaan fasilitas, maupun diskusi ilmiah yang telah memperkaya proses penyusunan artikel ini.

REFERENSI

Anonymous. (2023). Prediksi Kategori Kelulusan Mahasiswa Menggunakan Metode Regresi Logistik Multinomial. *ResearchGate Preprint*. https://www.researchgate.net/publication/371084473_Prediksi_Kategori_Kelulusan_Mahasiswa_Menggunakan_Metode_Regresi_Logistik_Multinomial

Authors, U. P. (2022). Regresi Logistik Biner untuk Mengklasifikasikan Cara Belajar Mahasiswa. *SciLine Journal (Unup Purwokerto)*. https://journal.unupurwokerto.ac.id/index.php/sciline/article/download/182/209/

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *Proceedings of the International Conference on Educational Data Mining / Related Workshop*.

Elvida, N. (2024). Penerapan Data Mining untuk Prediksi Kelulusan Siswa. *BCE Attractive Journal*. https://attractivejournal.com/index.php/bce/article/view/1538

Febrinita, F. (2024). Faktor-Faktor yang Mempengaruhi Hasil Belajar Statistika: Analisis Empiris. *Kognitif Journal*. https://etdci.org/journal/kognitif/article/view/1588

Gunawan, P. H. (2025). Deteksi Tingkat Potensi Kelulusan Calon Mahasiswa Menggunakan Data Akademik Sekolah. *STMSI (Unisi) Journal*. https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/download/5331/1032

JKTI, U. (2025). Predicting Student Graduation Using *Logistic regression* and Adam Optimization. *Jurnal Teknologi Dan Informatika (JKTI)*. https://jurnal.unimus.ac.id/index.php/JKTI/article/viewFile/16189/pdf

Journal, U. K. (2020). Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik (Naive Bayes Study). *Jurnal Decode (Universitas Halu Oleo / UM Kendari)*. https://journal.umkendari.ac.id/decode/article/download/308/144/1882

Junaidi, S. (2023). Prediksi Kelulusan Tepat Waktu Mahasiswa. *EDikInformatika (Ejournal.Upgrisba.Ac.Id)*. https://ejournal.upgrisba.ac.id/index.php/eDikInformatika/article/view/7324

Kharis, S. A. A., & Zili, A. H. A. (2022). Learning Analytics dan Educational Data Mining pada Data Pendidikan. *Jurnal Universitas Terbuka / ResearchGate Preprint*. https://www.researchgate.net/publication/359631908_Learning_Analytics_dan_Educational_Data_Mining_pada_Data_Pendidikan

Latupeirissa, S. J. (2019). Pemodelan Lama Masa Studi Mahasiswa Menggunakan Regresi Logistik Ordinal. *Jurnal FMIPA (Garuda / Repository Nasional)*. https://download.garuda.kemdikbud.go.id/article.php?article=1442829&title=Pemodelan+Lama+Masa+Studi+Mahasiswa+Fmipa+Unpatti+Menggunakan+Regresi+Logistik+Ordinal+Dengan+Efek+Interaksi&Val=17683

Maqfiroh, & Mujiyono, S. (2022). Penerapan Klasifikasi Algoritma Data Mining C4.5 untuk Memprediksi Tingkat Kelulusan Siswa. *Attractive Journal*. https://attractivejournal.com/index.php/bce/article/view/1538

Mk, C. R. P. (2025). Data Mining Classification Model For Timeliness Of Student Graduation. *DE Journal (Undhari)*. https://ejournal.undhari.ac.id/index.php/de_journal/article/view/1349

STIS, P. (2019). Penerapan Metode Regresi Logistik Biner untuk Mengetahui Faktor Pengaruh. *Prosiding Seminar Nasional Statistik (STIS)*. https://prosiding.stis.ac.id/index.php/semnasoffstat/article/download/146/43/

Triyasri, N. (2021). Prediction of Academic Success Using *Logistic regression*. *Jurnal JUSTIN, Universitas Tanjungpura*. https://jurnal.untan.ac.id/index.php/justin/article/download/89731/75676607260