

Implementasi Analisis Diskriminan dan Regresi Logistik Multinomial pada *Dataset Student Dropout and Academic Success*

ILLONA ANINDYA, DAFFA AHMAD PANGREKSA, RIZQI AQILAH CAHAYANI YUNIARTO, ULFA
SITI NURAINI*

Program Studi Sarjana Sains Data, Universitas Negeri Surabaya, Indonesia

Abstrak

Penelitian ini membandingkan *Linear Discriminant Analysis* (LDA) dan Regresi Logistik Multinomial (RLM) dalam mengklasifikasikan hasil akademik 4.424 mahasiswa (*Enrolled*, *Graduate*, *Dropout*) dari sebuah perguruan tinggi di Portugal. LDA diterapkan menggunakan enam prediktor numerik dan terbukti signifikan secara statistik berdasarkan uji Wilks' Lambda ($\Lambda = 0,6213$, $p < 0,0001$), namun hanya mencapai akurasi 65,03% dengan *recall Graduate* tertinggi (94,43%) dan gagal sepenuhnya mengklasifikasikan kelas *Enrolled*. RLM diterapkan menggunakan sepuluh prediktor (enam numerik dan empat biner) dengan model yang signifikan secara statistik ($G^2 = 2546,486$, $p \approx 0$) serta kecocokan model yang baik (R^2 McFadden = 0,2822), mencapai akurasi 69,39% dan *recall Dropout* yang jauh lebih baik (69,81% berbanding 55,67%). Status pembayaran UKT teridentifikasi sebagai prediktor paling berpengaruh, menurunkan probabilitas *Dropout* sebesar 41,29 pp dan meningkatkan probabilitas *Graduate* sebesar 34,80 pp. RLM mengungguli LDA pada hampir seluruh metrik evaluasi, terutama karena kemampuannya mengintegrasikan prediktor biner finansial yang tidak dapat diakomodasi oleh LDA.

Kata Kunci *Analisis Diskriminan Linear; Regresi Logistik Multinomial; Prediksi Student Dropout; Klasifikasi Hasil Akademik; Evaluasi Akademik*

Abstract

This study compares Linear Discriminant Analysis (LDA) and Multinomial Logistic Regression (MLR) for classifying academic outcomes of 4,424 students (*Enrolled*, *Graduate*, *Dropout*) from a Portuguese higher education institution. LDA was applied using six numeric predictors and proved statistically significant based on Wilks' Lambda ($\Lambda = 0.6213$, $p < 0.0001$), yet achieved only 65.03% accuracy with the highest *Graduate* recall (94.43%) while failing entirely on the *Enrolled* class. MLR was applied using ten predictors (six numeric and four binary), yielding a statistically significant model ($G^2 = 2546.486$, $p \approx 0$) with good fit (R^2 McFadden = 0.2822), achieving 69.39% accuracy and substantially better *Dropout* recall (69.81% vs. 55.67%). Tuition payment status was identified as the strongest predictor, reducing *Dropout* probability by 41.29 pp and increasing *Graduate* probability by 34.80 pp. MLR outperformed LDA on nearly all evaluation metrics, primarily due to its ability to incorporate binary financial predictors that LDA could not accommodate.

Keywords *Linear Discriminant Analysis; Multinomial Logistic Regression; Student Dropout*

*Corresponding Author Email: ulfanuraini@unesa.ac.id.

Prediction; Academic Outcome Classification; Academic Evaluation

1 Pendahuluan

1.1 Latar Belakang

Student dropout atau putus studi telah menjadi permasalahan kritis di perguruan tinggi karena berdampak pada trajektori akademik individu maupun kinerja institusi secara keseluruhan. Identifikasi mahasiswa yang berisiko putus studi sejak dini sangat penting untuk menerapkan strategi intervensi yang tepat sasaran dan tepat waktu (Realinho et al., 2021). Berbagai faktor demografis, sosial-ekonomi, dan kinerja akademik telah terbukti berperan signifikan dalam menentukan apakah seorang mahasiswa akan putus studi, tetap terdaftar, atau berhasil lulus (Barile, 2025).

Penelitian terdahulu telah menerapkan berbagai pendekatan untuk memprediksi hasil akademik mahasiswa, misalnya melalui model campuran (*finite mixture models*) multinomial (Barile, 2025) maupun pendekatan klasifikasi pada *dataset* multikelas yang tidak seimbang (Contreras Montes et al., 2023). Namun, pendekatan tersebut umumnya diterapkan secara tunggal tanpa membandingkan langsung dengan metode diskriminan klasik, dan jarang disertai verifikasi asumsi statistik secara eksplisit maupun integrasi prediktor finansial biner, seperti status pembayaran uang kuliah dan beasiswa, yang secara teoretis penting namun tidak dapat diakomodasi oleh Analisis Diskriminan karena keterbatasan asumsi normalitas multivariat. Celah penelitian inilah yang menjadi dasar penelitian ini.

Analisis Diskriminan merupakan teknik multivariat yang membangun fungsi diskriminan linear optimal untuk memisahkan observasi ke dalam kelompok yang telah ditentukan (Mwavita, 2023). Regresi Logistik Multinomial memperluas regresi logistik biner untuk mengakomodasi tiga kategori keluaran atau lebih, sekaligus menyediakan interpretasi probabilistik keanggotaan kelompok (Contreras Montes et al., 2023). Kedua metode ini menawarkan interpretasi koefisien yang jelas dan teruji secara inferensial, sehingga sesuai untuk kebutuhan pengambilan keputusan institusional yang memerlukan transparansi.

Penelitian ini memberikan kontribusi dengan: (1) menerapkan dan memverifikasi secara eksplisit seluruh asumsi yang diperlukan untuk Analisis Diskriminan dan Regresi Logistik Multinomial, (2) membandingkan kinerja kedua metode secara langsung pada *dataset* yang sama, dan (3) mengevaluasi kontribusi tambahan prediktor biner finansial yang hanya dapat diakomodasi oleh Regresi Logistik Multinomial. Penelitian ini menerapkan kedua metode tersebut pada *dataset Predict Students' Dropout and Academic Success* dari UCI Machine Learning Repository, yang mencakup 4.424 catatan mahasiswa dengan 36 fitur dan tiga kelas keluaran: *Dropout*, *Enrolled*, dan *Graduate*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah:

1. Apakah Analisis Diskriminan mampu membedakan dan mengklasifikasikan mahasiswa secara signifikan ke dalam tiga kategori hasil akademik (*Dropout*, *Enrolled*, dan *Graduate*) berdasarkan prediktor numerik?
2. Apakah model Regresi Logistik Multinomial signifikan dalam memodelkan hasil akademik mahasiswa (*Dropout*, *Enrolled*, dan *Graduate*), serta variabel prediktor apa saja yang berpengaruh signifikan terhadap probabilitas setiap kategori keluaran?

3. Bagaimana perbandingan kinerja klasifikasi antara Analisis Diskriminan dan Regresi Logistik Multinomial pada *dataset student dropout*?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Menerapkan Analisis Diskriminan dalam mengklasifikasikan mahasiswa ke dalam tiga kelompok hasil akademik dan mengevaluasi akurasi klasifikasinya.
2. Menerapkan Regresi Logistik Multinomial dalam memodelkan hasil akademik mahasiswa dan mengidentifikasi variabel prediktor yang signifikan.
3. Membandingkan kinerja kedua metode pada *dataset student dropout*.

2 Metodologi Penelitian

2.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah **Predict Students' Dropout and Academic Success**, yang tersedia secara publik di UCI Machine Learning Repository dan dapat diakses melalui tautan berikut: <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>.

Dataset ini terdiri dari 4.424 catatan mahasiswa yang dikumpulkan dari sebuah perguruan tinggi di Portugal. Setiap catatan mewakili satu mahasiswa unik sehingga seluruh observasi bersifat independen tanpa pengukuran berulang. Data mencakup mahasiswa yang terdaftar pada berbagai program sarjana, beserta informasi pada saat pendaftaran dan pada akhir tahun akademik. Fokus utama penelitian ini adalah mengklasifikasikan status akademik akhir mahasiswa (*Graduate*, *Dropout*, atau *Enrolled*) berdasarkan kinerja akademik, latar belakang demografis, dan indikator makroekonomi.

2.2 Variabel Penelitian

Variabel-variabel yang digunakan dalam penelitian ini disajikan dalam Tabel 1.

Table 1: Variabel Penelitian

Nama Variabel	Deskripsi Variabel	Peran
<i>target</i>	Status akademik: <i>Enrolled</i> (ref.), <i>Graduate</i> , <i>Dropout</i>	Variabel Dependen
<i>admission_grade</i>	Nilai penerimaan mahasiswa saat masuk	Prediktor Numerik
<i>cu1_grade</i>	Nilai semester 1	Prediktor Numerik
<i>cu2_grade</i>	Nilai semester 2	Prediktor Numerik
<i>log_age_enrollment</i>	Log usia saat pendaftaran	Prediktor Numerik
<i>log_unemployment_rate</i>	Log tingkat pengangguran	Prediktor Numerik
<i>gdp</i>	Indikator PDB	Prediktor Numerik
<i>gender</i>	Jenis kelamin: 0 = Perempuan, 1 = Laki-laki	Prediktor Biner
<i>scholarship_holder</i>	Status beasiswa: 0 = Tidak, 1 = Ya	Prediktor Biner
<i>debtor</i>	Status utang: 0 = Tidak ada utang, 1 = Memiliki utang	Prediktor Biner
<i>tuition_up_to_date</i>	Pembayaran UKT: 0 = Menunggak, 1 = Lunas	Prediktor Biner

Keterangan:

- Variabel *target* merupakan variabel respons dengan 3 level tidak terurut: *Enrolled* (kategori referensi), *Graduate*, dan *Dropout*.
- Prediktor numerik *admission_grade*, *cu1_grade*, dan *cu2_grade* mencerminkan kinerja akademik, sedangkan *log_age_enrollment* merefleksikan latar belakang demografis. *log_unemployment_rate* dan *gdp* merepresentasikan kondisi makroekonomi pada saat pendaftaran.
- Prediktor biner *gender*, *scholarship_holder*, *debtor*, dan *tuition_up_to_date* digunakan dalam Regresi Logistik Multinomial (RLM) dan dikecualikan dari Analisis Diskriminan Linear (LDA), karena asumsi normalitas multivariat pada LDA tidak dapat dipenuhi oleh variabel biner.

2.3 Analisis Diskriminan

Linear Discriminant Analysis (LDA) adalah metode klasifikasi terbimbing yang membangun fungsi diskriminan, yaitu kombinasi linear dari prediktor, yang paling baik memisahkan kelompok yang telah ditentukan. Dengan $K = 3$ kelompok keluaran (*Enrolled*, *Graduate*, *Dropout*), LDA menghasilkan $K - 1 = 2$ fungsi diskriminan (Abubakar et al., 2024).

Setiap fungsi diskriminan berbentuk:

$$f_k(\mathbf{x}) = a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kp}x_p, \quad (1)$$

di mana a_{kj} adalah koefisien diskriminan dan x_j adalah variabel prediktor. Koefisien dipilih untuk memaksimalkan rasio variansi antarkelompok terhadap variansi dalam-kelompok. Secara formal, hal ini diekspresikan sebagai pemaksimalan $\mathbf{a}^\top \mathbf{B} \mathbf{a} / \mathbf{a}^\top \mathbf{W} \mathbf{a}$, di mana $\mathbf{B}$ adalah matriks *scatter* antarkelompok dan $\mathbf{W}$ adalah matriks *scatter* dalam-kelompok (Mwavita, 2023).

Observasi baru ditetapkan ke kelompok dengan probabilitas posterior tertinggi (Graf et al., 2023):

$$P(G = k | \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^K \pi_j f_j(\mathbf{x})}, \quad (2)$$

di mana π_k adalah probabilitas prior kelompok k dan $f_k(\mathbf{x})$ adalah densitas normal multivariat untuk kelompok k . Probabilitas prior diestimasi dari frekuensi kelas pada data latih. Kedua asumsi yang diperlukan untuk LDA, yaitu normalitas multivariat dan homogenitas matriks kovarians, telah diverifikasi pada Bagian 3.2.

2.4 Regresi Logistik Multinomial

Regresi Logistik Multinomial (RLM) adalah generalisasi dari regresi logistik biner untuk memodelkan variabel respons kategorik dengan 3 atau lebih kelas keluaran tidak terurut. Berbeda dengan LDA, RLM tidak menerapkan asumsi distribusional pada prediktor, sehingga dapat diterapkan pada *dataset* dengan prediktor kontinu dan biner yang bercampur (Barile, 2025). Dengan K kategori keluaran, RLM mengestimasi $K - 1$ persamaan log-*odds* relatif terhadap kategori referensi.

Dalam penelitian ini, *target* memiliki 3 kategori dengan *Enrolled* sebagai referensi:

$$\ln \frac{P(\text{Graduate})}{P(\text{Enrolled})} = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p, \quad (3)$$

$$\ln \frac{P(\text{Dropout})}{P(\text{Enrolled})} = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p. \quad (4)$$

Probabilitas prediksi untuk ketiga kategori diperoleh melalui transformasi *softmax*, yang menjamin nilainya non-negatif dan berjumlah satu (Contreras Montes et al., 2023):

$$\pi_{Enrolled} = \frac{1}{1 + e^{P_1} + e^{P_2}}, \quad \pi_{Graduate} = \frac{e^{P_1}}{1 + e^{P_1} + e^{P_2}}, \quad \pi_{Dropout} = \frac{e^{P_2}}{1 + e^{P_1} + e^{P_2}}, \quad (5)$$

di mana P_1 dan P_2 masing-masing merujuk pada sisi kanan persamaan (3) dan (4).

Parameter diestimasi melalui *Maximum Likelihood Estimation* (MLE) menggunakan algoritma iteratif Newton-Raphson, yang memaksimalkan *log-likelihood*:

$$\ell(\beta) = \sum_{i=1}^n \left[y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - \ln \left(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)} \right) \right]. \quad (6)$$

Residual Deviance ($-2 \log L$) dan *Akaike Information Criterion* (AIC, $-2 \log L + 2k$) digunakan sebagai ukuran kecocokan model, sedangkan R^2 McFadden didefinisikan sebagai:

$$R_{\text{McFadden}}^2 = 1 - \frac{\log L_{\text{full}}}{\log L_{\text{null}}}, \quad (7)$$

dengan nilai antara 0,2 dan 0,4 mengindikasikan kecocokan model yang baik (Barile, 2025).

Signifikansi model dievaluasi menggunakan dua uji:

1. **Uji simultan** menggunakan *Likelihood Ratio Test* (LRT), membandingkan model penuh dengan model nol hanya-intersep. Statistik $G^2 = -2 \ln(L_0/L_1)$ mengikuti distribusi chi-kuadrat dengan derajat bebas sama dengan jumlah parameter yang ditambahkan (Contreras Montes et al., 2023).
2. **Uji parsial** menggunakan statistik Wald $W^2 = \left(\hat{\beta}_j / \text{SE}(\hat{\beta}_j) \right)^2$ untuk mengevaluasi setiap prediktor secara individual, juga berdistribusi chi-kuadrat dengan 1 derajat bebas. Ambang batas signifikansi adalah $p < 0,05$.

Koefisien dilaporkan dalam 3 skala:

- **Log-odds** (β): Perubahan *log-odds* per satuan perubahan prediktor relatif terhadap referensi.
- **Relative Risk Ratios** ($\text{RRR} = e^{\beta}$): Perubahan multiplikatif terhadap *odds*.
- **Average Marginal Effects** (AME): Dihitung menggunakan paket `marginalEffects`, mengekspresikan rata-rata perubahan probabilitas prediksi di seluruh observasi, dan merupakan skala yang paling mudah diinterpretasikan secara langsung (Arel-Bundock et al., 2024).

2.5 Evaluasi dan Perbandingan Visual

Evaluasi dilakukan menggunakan tiga pendekatan komplementer: matriks konfusi dan APER untuk kedua metode, serta metrik per kelas (Presisi, *Recall*, *F1-Score*) terutama untuk RLM.

2.5.1 Matriks Konfusi dan APER

Matriks konfusi merupakan alat fundamental untuk mengevaluasi kinerja klasifikasi (Sathyanarayanan and Tantri, 2024). APER (*Apparent Error Rate*) didefinisikan sebagai fraksi observasi yang salah diklasifikasikan pada data latih:

$$\text{APER} = \frac{\text{Jumlah observasi yang salah diklasifikasikan}}{\text{Total observasi}}. \quad (8)$$

Nilai APER yang lebih rendah menunjukkan kinerja yang lebih baik. APER dihitung untuk DA dan RLM untuk memungkinkan perbandingan langsung.

2.5.2 Presisi, *Recall*, dan F1-Score

Selain akurasi keseluruhan, metrik per kelas memberikan evaluasi yang lebih mendalam pada *dataset* tidak seimbang, di mana model mungkin mencapai akurasi tinggi hanya dengan memprediksi kelas mayoritas. *Precision* mengukur proporsi prediksi positif yang benar, *Recall* mengukur proporsi positif aktual yang teridentifikasi dengan benar, dan F1-Score merepresentasikan keseimbangan keduanya (Sathyanarayanan and Tantri, 2024).

2.5.3 Perbandingan Visual

Dua jenis visualisasi digunakan: *heatmap* matriks konfusi untuk setiap metode guna menampilkan klasifikasi yang benar dan salah pada ketiga kelas keluaran, serta diagram batang yang menyajikan akurasi keseluruhan, APER, dan *recall* per kelas secara berdampingan untuk perbandingan langsung antara DA dan RLM.

3 Hasil dan Pembahasan

3.1 Karakteristik Data

Dataset terdiri dari 4.424 catatan mahasiswa dari sebuah perguruan tinggi di Portugal, dengan variabel respons *target* yang mengklasifikasikan setiap mahasiswa ke dalam salah satu dari tiga hasil akademik: *Enrolled*, *Graduate*, atau *Dropout*.

Table 2: Distribusi Variabel *Target*

Kategori	Jumlah	Proporsi
<i>Enrolled</i>	794	0,179
<i>Graduate</i>	2.209	0,499
<i>Dropout</i>	1.421	0,321

Berdasarkan Tabel 2, *dataset* bersifat tidak seimbang antarkelas, dengan *Graduate* sebagai kelompok terbesar (49,9%) dan *Enrolled* sebagai kelompok terkecil (17,95%). Hal ini merupakan pertimbangan penting karena pengklasifikasi cenderung memihak kelas mayoritas. Statistik deskriptif antarkelompok *target* menunjukkan bahwa *Graduate* secara konsisten mencapai nilai semester tertinggi (*cu1_grade* = 12,64; *cu2_grade* = 12,69), sementara *Dropout* menunjukkan nilai terendah (*cu1_grade* = 7,25; *cu2_grade* = 5,90), dengan kelompok *Enrolled* berada di posisi menengah. Mahasiswa *Dropout* juga cenderung lebih tua saat mendaftar (*log_age_enrollment* = 3,21), sedangkan variabel makroekonomi menunjukkan perbedaan yang sangat kecil antarkelompok.

Seperti terlihat pada Gambar 1, *boxplot* mengonfirmasi bahwa *cu1_grade* dan *cu2_grade* menunjukkan pemisahan visual terkuat antara kelompok *Graduate* dan *Dropout*, *admission_grade* menunjukkan pemisahan moderat, sementara *gdp* dan *log_unemployment_rate* hampir tidak menunjukkan pemisahan antarkelompok.

Gambar 2 mengungkapkan bahwa mahasiswa dengan pembayaran UKT yang lunas dan penerima beasiswa lebih cenderung *Graduate*, sementara mahasiswa dengan status *debtor* menunjukkan proporsi hasil *Dropout* yang lebih tinggi. Hal ini mengindikasikan bahwa faktor finansial memainkan peran penting dalam hasil akademik mahasiswa.

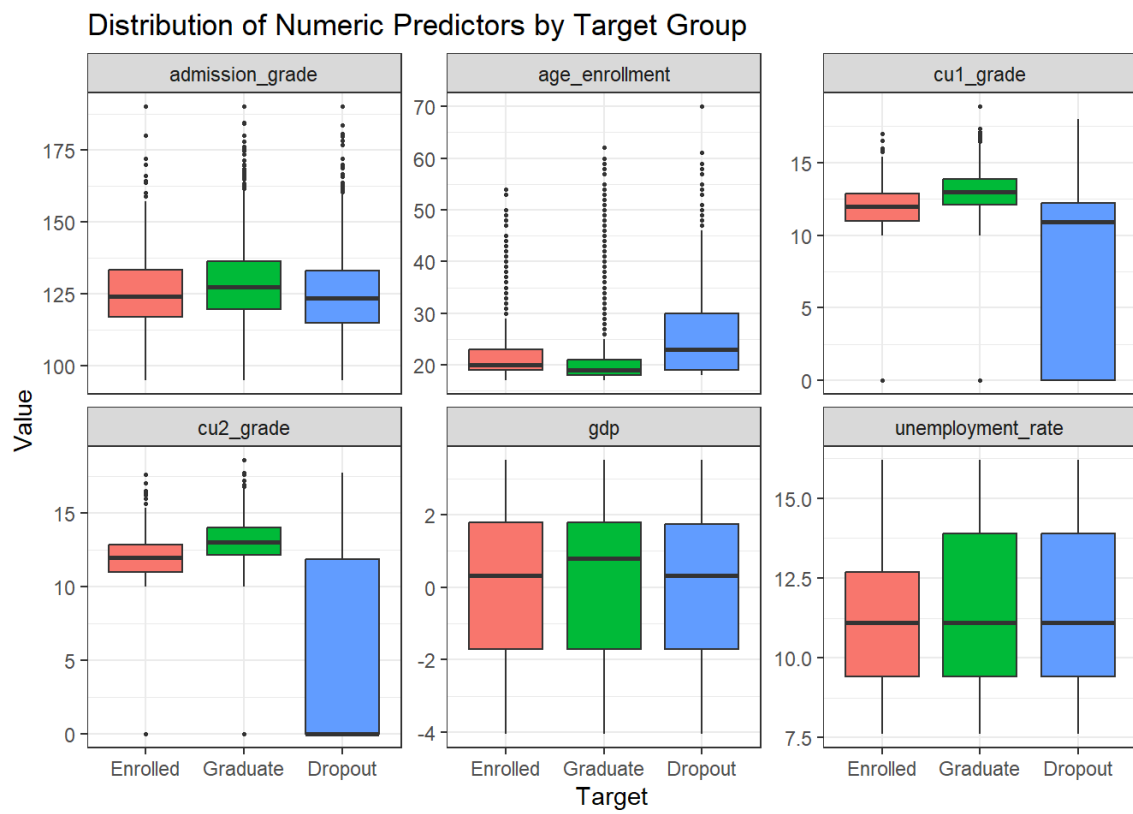


Figure 1: *Boxplot* Prediktor Numerik berdasarkan Kelompok *Target*

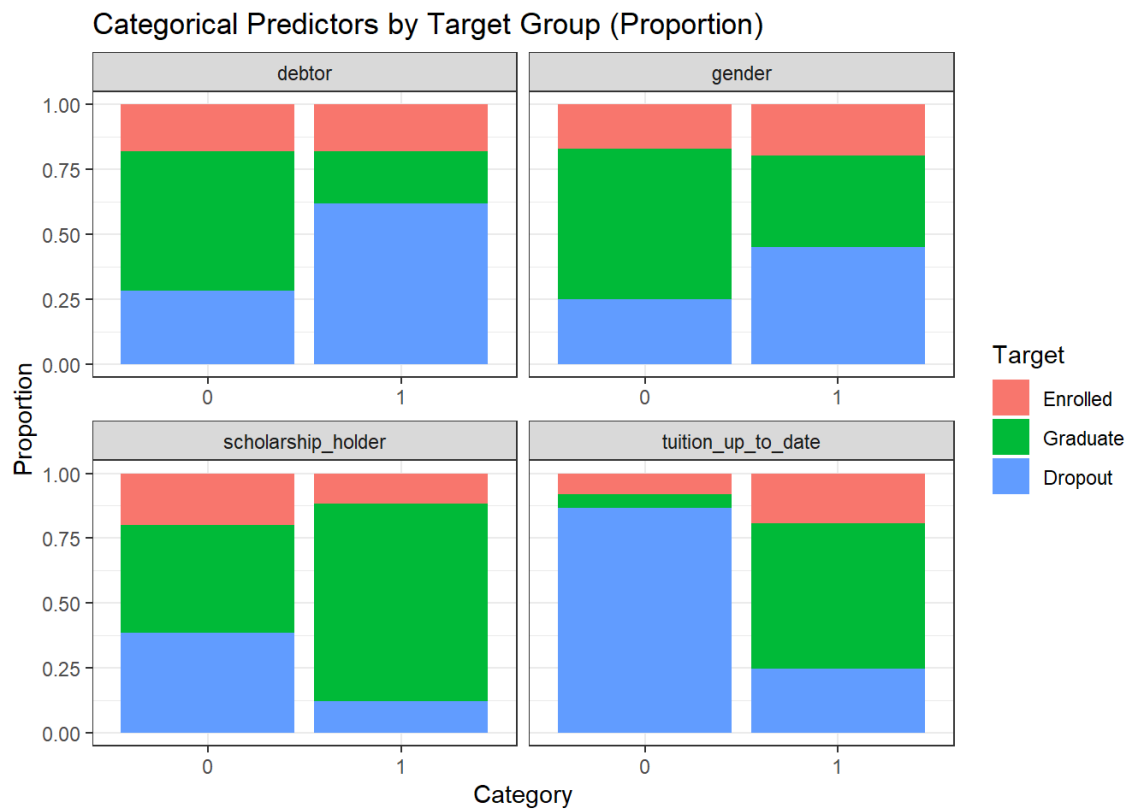


Figure 2: Diagram Batang Bertumpuk Prediktor Kategorik berdasarkan Kelompok *Target* (Proporsi)

3.2 Uji Asumsi

Sebelum menerapkan DA dan RLM, seluruh asumsi yang diperlukan telah diuji. Asumsi DA mencakup normalitas multivariat dan homogenitas matriks kovarians, sedangkan asumsi RLM mencakup linearitas log-*odds*, independensi observasi, tidak adanya multikolinearitas, dan tidak adanya *outlier* ekstrem.

3.2.1 Asumsi DA

Normalitas Multivariat. LDA mensyaratkan bahwa prediktor numerik secara bersama-sama mengikuti distribusi normal multivariat, yang dievaluasi menggunakan uji Mardia dan plot Q-Q chi-kuadrat dari jarak Mahalanobis.

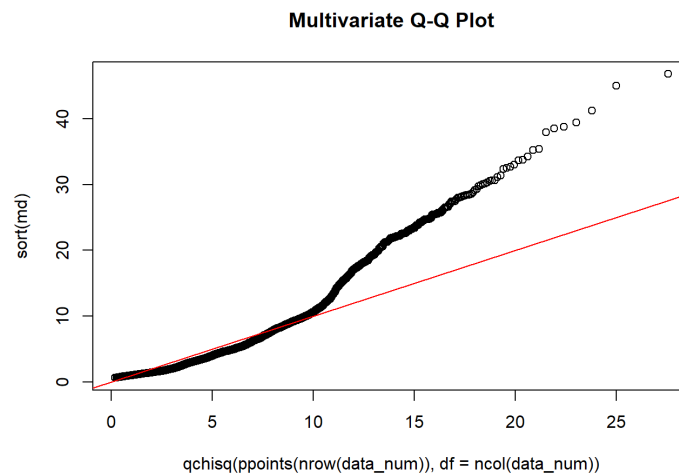


Figure 3: Plot Q-Q Chi-Kuadrat Multivariat

Dengan $n = 4.424$, uji formal memiliki daya yang sangat tinggi dan dapat menandai penyimpangan kecil sebagai signifikan, sehingga plot Q-Q (Gambar 3) berfungsi sebagai diagnostik utama. Karena titik-titik yang diplot mengikuti garis referensi diagonal secara cukup baik, Asumsi normalitas multivariat terpenuhi.

Homogenitas Matriks Kovarians. LDA mengasumsikan bahwa matriks kovarians prediktor numerik sama di ketiga kelompok *target*. Hal ini dievaluasi menggunakan uji Box's M.

Table 3: Hasil Uji Box's M

Chi-Kuadrat	db	<i>p</i> -value
4674,479	42	$< 2,2 \times 10^{-16}$

Hasil uji Box's M (Tabel 3) menghasilkan Chi-Kuadrat = 4674,479 ($db = 42$, $p < 2,2 \times 10^{-16}$), sehingga H_0 ditolak. *Heatmap* korelasi (Gambar 4) mengonfirmasi adanya perbedaan struktur kovarians antarkelompok. Asumsi homogenitas tidak terpenuhi. Namun, penelitian ini tetap melanjutkan LDA karena uji Box's M sangat sensitif pada sampel besar, sehingga penolakan

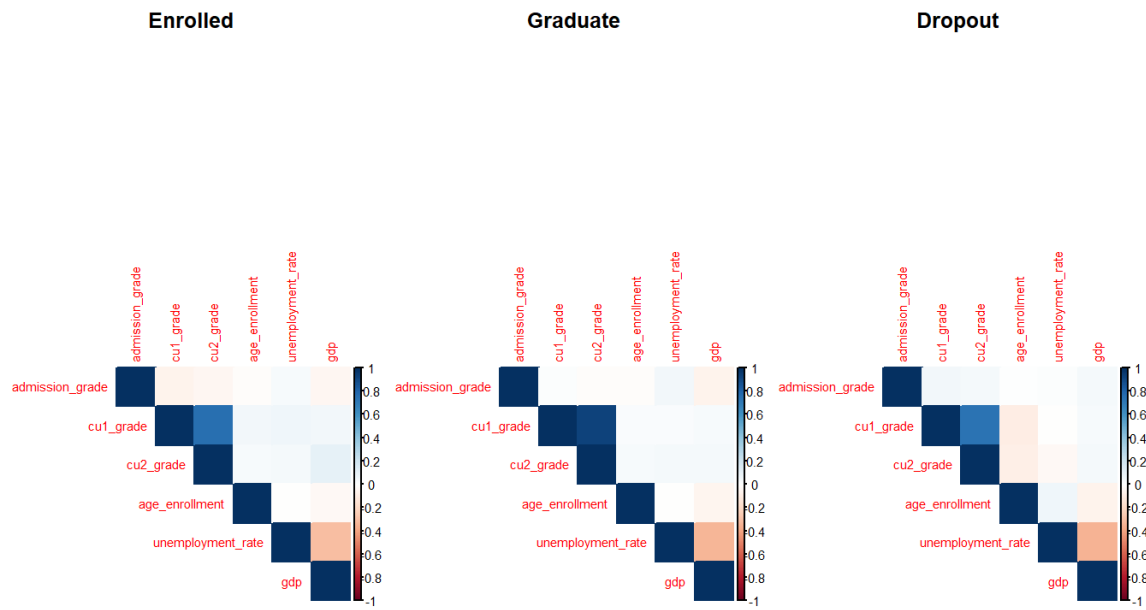


Figure 4: *Heatmap* Korelasi untuk Setiap Kelompok *Target* (*Enrolled*, *Graduate*, *Dropout*)

tersebut belum tentu bermakna secara praktis, dan LDA cukup *robust* terhadap pelanggaran ini ketika tujuan utamanya adalah klasifikasi. Keterbatasan ini diakui secara eksplisit.

3.2.2 Asumsi RLM

Linearitas Log-Odds. RLM mengasumsikan bahwa log-*odds* setiap keluaran relatif terhadap kategori referensi berhubungan secara linear dengan setiap prediktor kontinu. Hal ini dievaluasi menggunakan uji Box-Tidwell, yang menambahkan suku interaksi berbentuk $x \cdot \log(x)$ pada regresi logistik biner. Interaksi yang tidak signifikan ($p > 0,05$) mendukung asumsi linearitas.

Uji Box-Tidwell awalnya menandai *age_enrollment* dan *unemployment_rate* sebagai non-linear ($p < 0,05$), sehingga mendorong transformasi logaritmik pada kedua variabel tersebut. Setelah transformasi, non-linearitas tersisa pada *cu1_grade*, *cu2_grade*, dan *log_age_enrollment* diakui, namun RLM umumnya *robust* terhadap penyimpangan moderat pada sampel besar. Asumsi linearitas log-*odds* terpenuhi secara wajar.

Independensi Observasi. Setiap baris merepresentasikan satu mahasiswa unik tanpa pengukuran berulang maupun struktur bertingkat. Asumsi independensi observasi terpenuhi.

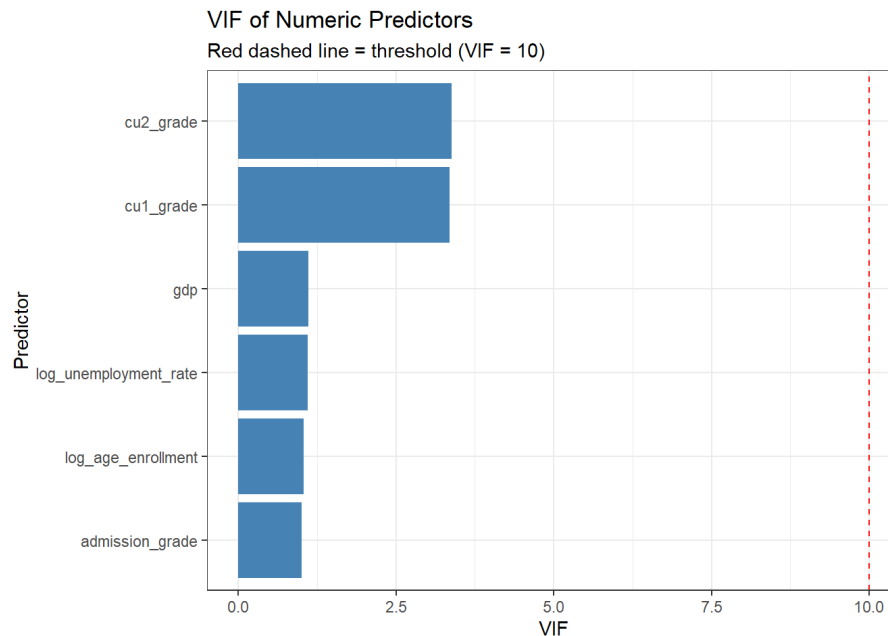


Figure 5: Diagram Batang VIF dengan Garis Referensi Merah Putus-Putus pada $VIF = 10$

Tidak Ada Multikolinearitas. Seluruh prediktor numerik memiliki nilai *Variance Inflation Factor* (VIF) jauh di bawah 10 (Gambar 5), mengindikasikan tidak adanya multikolinearitas yang parah di antara prediktor. Asumsi tidak ada multikolinearitas terpenuhi.

Tidak Ada Outlier. *Outlier* dideteksi melalui skor Z ($|Z| > 3$) dan ditangani menggunakan Winsorisasi pada persentil ke-1 dan ke-99.

Table 4: Jumlah *Outlier* per Variabel Setelah Winsorisasi

Variabel	Outliers_Z3	Pct (%)
<i>admission_grade</i>	0	0
<i>cu1_grade</i>	0	0
<i>cu2_grade</i>	0	0
<i>log_age_enrollment</i>	59	1,33
<i>log_unemployment_rate</i>	0	0
<i>gdp</i>	0	0

Berdasarkan Tabel 4, setelah penanganan, sebagian besar prediktor tidak menunjukkan *outlier*, dengan *log_age_enrollment* masih menyisakan 59 observasi (1,33%) akibat distribusinya yang condong kanan bahkan setelah transformasi log. Karena proporsi ini sangat kecil dan seluruh 4.424 data tetap dipertahankan, asumsi tidak ada *outlier* terpenuhi secara wajar.

3.3 Analisis Diskriminan

3.3.1 Data yang Digunakan

Analisis Diskriminan hanya menggunakan enam prediktor numerik dari *data_final*. Prediktor biner (*gender*, *scholarship_holder*, *debtor*, *tuition_up_to_date*) dikecualikan karena asumsi normalitas multivariat LDA tidak dapat dipenuhi oleh indikator biner. *Dataset* untuk DA terdiri dari 4.424 observasi dengan enam prediktor berikut: *admission_grade*, *cu1_grade*, *cu2_grade*, *log_age_enrollment*, *log_unemployment_rate*, dan *gdp*.

3.3.2 Ringkasan LDA

Parameter LDA diestimasi menggunakan matriks kovarians dalam-kelompok gabungan (*pooled*) dan vektor rata-rata kelompok. Koefisien diskriminan diperoleh dengan menyelesaikan masalah nilai eigen umum (*generalized eigenvalue problem*). Model diterapkan menggunakan probabilitas prior proporsional terhadap frekuensi kelas pada data latih.

Table 5: Probabilitas Prior dan Rata-Rata Kelompok pada LDA

Kelompok	Prior	<i>admission_grade</i>	<i>cu1_grade</i>	<i>cu2_grade</i>	<i>log_age_enrollment</i>	<i>log_unemployment_rate</i>	<i>gdp</i>
<i>Enrolled</i>	0,1795	125,4659	11,1221	11,1127	3,0779	2,3952	0,0533
<i>Graduate</i>	0,4993	128,7317	12,6350	12,6888	3,0476	2,4289	0,0818
<i>Dropout</i>	0,3212	124,9359	7,2514	5,8958	3,2120	2,4235	-0,1509

Tabel 5 dan Tabel 6 menunjukkan keluaran LDA. Probabilitas prior mencerminkan distribusi kelas: *Graduate* (49,94%), *Dropout* (32,12%), dan *Enrolled* (17,95%). Rata-rata kelompok mengungkapkan pemisahan yang jelas, khususnya untuk prediktor terkait nilai. Mahasiswa *Dropout* menunjukkan *cu1_grade* (7,25) dan *cu2_grade* (5,90) yang jauh lebih rendah dibandingkan *Graduate* (masing-masing 12,64 dan 12,69), sementara kelompok *Enrolled* berada di posisi menengah (11,12 untuk keduanya).

Berdasarkan Tabel 7, LD1 menjelaskan 98,63% dari total variansi antarkelompok, menunjukkan bahwa hampir seluruh informasi diskriminatif terkonsentrasi pada satu dimensi dominan yang didorong oleh nilai semester dan usia pendaftaran. LD2 hanya menyumbang 1,37%.

Table 6: Koefisien Fungsi Diskriminan Linear (LDA)

Prediktor	LD1	LD2
<i>admission_grade</i>	−0,0009	−0,0483
<i>cu1_grade</i>	−0,0052	−0,1489
<i>cu2_grade</i>	−0,2125	0,1299
<i>log_age_enrollment</i>	1,3688	−0,6422
<i>log_unemployment_rate</i>	−0,0009	−2,5244
<i>gdp</i>	0,0026	−0,0710

Table 7: Proporsi Variansi Antar-Kelompok yang Dijelaskan

	<i>Eigenvalue</i>	Proporsi	Kumulatif
LD1	1318,0448	0,9863	0,9863
LD2	18,3591	0,0137	1,0000

3.3.3 Uji Signifikansi (Wilks' Lambda)

Wilks' Lambda merupakan statistik uji multivariat keseluruhan untuk analisis diskriminan. Statistik ini mengukur proporsi total variansi dalam ruang prediktor yang tidak dijelaskan oleh perbedaan kelompok. Nilai mendekati 0 mengindikasikan pemisahan kelompok yang lebih kuat, sedangkan nilai 1 menunjukkan tidak adanya pemisahan sama sekali.

Hipotesis yang diuji adalah: H_0 : vektor rata-rata kelompok sama di ketiga kategori (tidak ada fungsi diskriminan yang signifikan); H_1 : setidaknya satu vektor rata-rata kelompok berbeda secara signifikan dari yang lain.

H_0 ditolak jika aproksimasi chi-kuadrat yang dihitung melebihi nilai kritis:

$$\chi^2 = - \left[(n - 1) - \frac{p + K}{2} \right] \ln \Lambda, \quad (9)$$

di mana n adalah jumlah observasi, p adalah jumlah prediktor, K adalah jumlah kelompok, dan Λ adalah nilai Wilks' Lambda.

Table 8: Hasil Uji Wilks' Lambda

	Df	Wilks' Λ	Approx. F	num Df / den Df	<i>p</i> -value
<i>target</i>	2	0,6213	197,7400	12 / 8832	$< 2,2 \times 10^{-16}$

Tabel 8 menunjukkan Wilks' Lambda $\approx 0,6213$, mengindikasikan bahwa sekitar 37,87% variansi total dalam ruang prediktor dapat diatribusikan pada perbedaan kelompok (dihitung sebagai $1 - \Lambda = 0,3787$). Aproksimasi F yang terkait sangat signifikan ($p < 0,0001$), sehingga H_0 ditolak. Hal ini mengonfirmasi bahwa tiga kelompok mahasiswa (*Enrolled*, *Graduate*, *Dropout*) dapat dibedakan secara statistik menggunakan enam prediktor numerik. Model analisis diskriminan secara keseluruhan signifikan.

3.3.4 Hasil Klasifikasi LDA

Hasil klasifikasi dan metrik per kelas LDA dibahas secara komprehensif pada Bagian 3.5. Model LDA mencapai akurasi pelatihan *apparent* sebesar 65,03%, dengan akurasi LOO-CV yang sangat

serupa, mengindikasikan *overfitting* yang minimal.

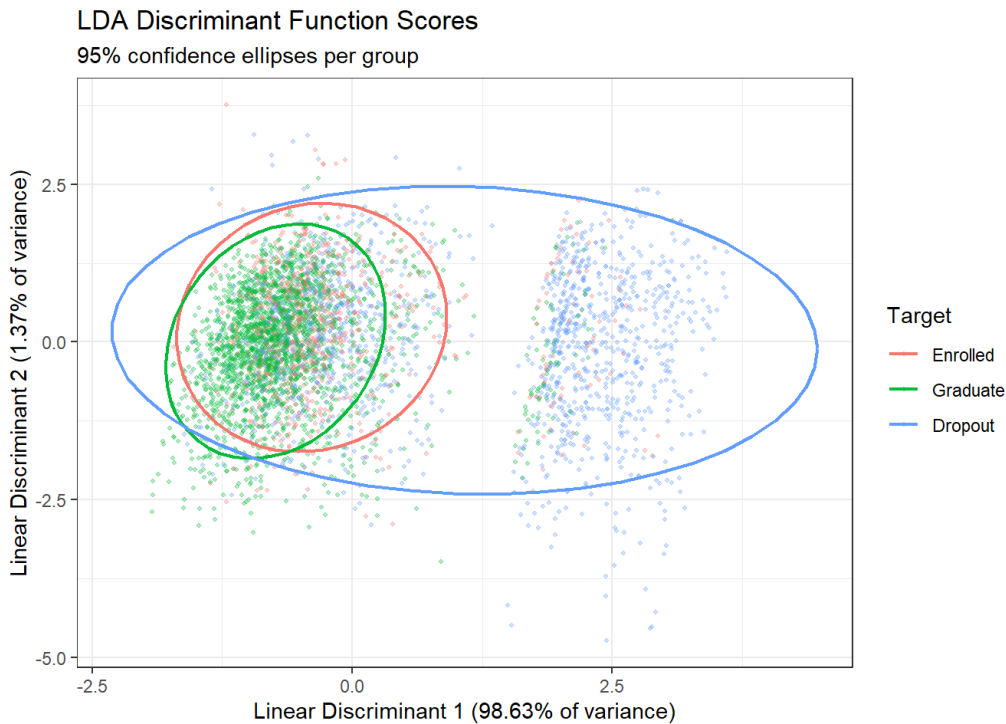


Figure 6: Scatter Plot Skor Fungsi Diskriminan LDA (LD1 vs. LD2)

Sepanjang LD1 yang menjelaskan 98,63% variansi antarkelompok (Gambar 6), mahasiswa *Graduate* berkelompok di nilai positif dan mahasiswa *Dropout* di nilai negatif, sementara mahasiswa *Enrolled* menempati wilayah menengah yang tumpang-tindih, konsisten dengan *recall* mendekati nol yang diamati untuk kelas tersebut.

3.3.5 Simpulan LDA

LDA secara global signifikan (Wilks' Lambda = 0,6213, $p < 0,0001$), mengonfirmasi bahwa tiga kelompok mahasiswa dapat dibedakan secara statistik menggunakan enam prediktor numerik. Kinerja akademik, khususnya nilai semester, merupakan pendorong utama pemisahan kelompok, sementara variabel makroekonomi berkontribusi relatif kecil. Model mencapai akurasi keseluruhan 65,03%, dengan kinerja terbaik pada kelas *Graduate* dan terburuk pada kelas *Enrolled*.

3.4 Regresi Logistik Multinomial

3.4.1 Data yang Digunakan

Dataset yang digunakan adalah `data_final` ($n = 4.424$), mencakup: 6 prediktor numerik (`admission_grade`, `cu1_grade`, `cu2_grade`, `log_age_enrollment`, `log_unemployment_rate`, `gdp`) dan 4 prediktor biner (`gender`, `scholarship_holder`, `debtor`, `tuition_up_to_date`). Distribusi *target*: *Enrolled* = 794, *Graduate* = 2.209, *Dropout* = 1.421, dengan *Enrolled* sebagai kategori referensi.

3.4.2 Estimasi Parameter

Diestimasi melalui MLE menggunakan algoritma Newton-Raphson. Model konvergen setelah 30 iterasi, dengan *log-likelihood* menurun dari $-4860,26$ ke nilai akhir $-3238,59$, mengonfirmasi konvergensi yang stabil. Metrik kecocokan model ($k = 22$):

$$\text{Residual Deviance} = -2 \times (-3238,59) = 6477,18$$

$$\text{AIC} = 6477,18 + 2(22) = 6521,18$$

3.4.3 Ringkasan Model

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Graduate	(Intercept)	-5.1796	0.9047	-5.7254	0.0000	-6.9528	-3.4065
Graduate	admission_grade	0.0184	0.0033	5.6243	0.0000	0.0120	0.0248
Graduate	cu1_grade	0.0564	0.0212	2.6639	0.0077	0.0149	0.0979
Graduate	cu2_grade	0.0958	0.0212	4.5103	0.0000	0.0542	0.1374
Graduate	log_age_enrollment	-0.0384	0.1880	-0.2044	0.8381	-0.4069	0.3301
Graduate	log_unemployment_rate	0.4315	0.1922	2.2451	0.0248	0.0548	0.8082
Graduate	gdp	0.0236	0.0204	1.1590	0.2464	-0.0163	0.0635
Graduate	gender1	-0.4621	0.0931	-4.9631	0.0000	-0.6446	-0.2796
Graduate	scholarship_holder1	1.0053	0.1097	9.1620	0.0000	0.7903	1.2204
Graduate	debtor1	-0.8320	0.1648	-5.0473	0.0000	-1.1551	-0.5089
Graduate	tuition_up_to_date1	1.0854	0.2597	4.1786	0.0000	0.5763	1.5945
Dropout	(Intercept)	-1.7790	0.9604	-1.8524	0.0640	-3.6614	0.1033
Dropout	admission_grade	-0.0032	0.0036	-0.8893	0.3738	-0.0104	0.0039
Dropout	cu1_grade	0.0300	0.0190	1.5804	0.1140	-0.0072	0.0671
Dropout	cu2_grade	-0.1867	0.0183	-10.1801	0.0000	-0.2226	-0.1508
Dropout	log_age_enrollment	1.4521	0.1967	7.3822	0.0000	1.0666	1.8377
Dropout	log_unemployment_rate	0.4768	0.2212	2.1557	0.0311	0.0433	0.9103
Dropout	gdp	-0.0034	0.0238	-0.1424	0.8868	-0.0500	0.0432
Dropout	gender1	0.1570	0.1042	1.5066	0.1319	-0.0472	0.3612
Dropout	scholarship_holder1	-0.3064	0.1492	-2.0535	0.0400	-0.5989	-0.0140
Dropout	debtor1	0.0571	0.1598	0.3577	0.7206	-0.2560	0.3703
Dropout	tuition_up_to_date1	-1.8975	0.1882	-10.0821	0.0000	-2.2664	-1.5286

Figure 7: Tabel Koefisien RLM (Log-Odds) dengan Interval Kepercayaan 95%

Gambar 7 menunjukkan bahwa:

- Untuk keluaran *Graduate*, prediktor positif yang signifikan meliputi *admission_grade*, *cu1_grade*, *cu2_grade*, *log_unemployment_rate*, *scholarship_holder*, dan *tuition_up_to_date*, sementara status *gender* (laki-laki) dan *debtor* secara signifikan mengurangi log-odds kelulusan.
- Untuk keluaran *Dropout*, *cu2_grade* dan *tuition_up_to_date* merupakan prediktor negatif terkuat, sedangkan *log_age_enrollment* merupakan prediktor positif terkuat.

3.4.4 Uji Signifikansi Parameter

Uji Simultan (Likelihood Ratio Test).

- $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$
- H_1 : Setidaknya satu $\beta_j \neq 0$
- Tolak H_0 jika $G^2 > \chi^2_{\alpha,p}$

Table 9: Uji Signifikansi Parameter Simultan

G^2	db	p -value	Keputusan
2546,486	20	≈ 0	Tolak H_0

Berdasarkan Tabel 9, setidaknya satu prediktor berkontribusi secara signifikan dalam menjelaskan status akademik mahasiswa. Model secara keseluruhan signifikan secara statistik.

Uji Parsial (Wald Test).

- $H_0: \beta_j = 0$
- $H_1: \beta_j \neq 0$
- Tolak H_0 jika $W^2 > \chi^2_{\alpha,1}$

Hasil Uji Wald parsial (Gambar 8) mengungkapkan:

- Prediktor signifikan untuk *Graduate* vs. *Enrolled* ($p < 0,05$): *admission_grade* ($W^2 = 31,63$), *cu1_grade* ($W^2 = 7,10$), *cu2_grade* ($W^2 = 20,34$), *log_unemployment_rate* ($W^2 = 5,04$), *gender* ($W^2 = 24,63$), *scholarship_holder* ($W^2 = 83,94$), *debtor* ($W^2 = 25,47$), dan *tuition_up_to_date* ($W^2 = 17,46$).
- Prediktor tidak signifikan untuk *Graduate* vs. *Enrolled*: *log_age_enrollment* dan *gdp*.
- Prediktor signifikan untuk *Dropout* vs. *Enrolled* ($p < 0,05$): *cu2_grade* ($W^2 = 103,63$), *log_age_enrollment* ($W^2 = 54,50$), *log_unemployment_rate* ($W^2 = 4,65$), *scholarship_holder* ($W^2 = 4,22$), dan *tuition_up_to_date* ($W^2 = 101,65$).
- Prediktor tidak signifikan untuk *Dropout* vs. *Enrolled*: *admission_grade*, *cu1_grade*, *gdp*, *gender*, dan *debtor*.

3.4.5 Kecocokan Model

Table 10: Ringkasan Kecocokan Model

Metrik	Nilai
Log-Likelihood	-3238,59
Residual Deviance	6477,18
AIC	6521,18
R^2 McFadden	0,2822

Berdasarkan Tabel 10, R^2 McFadden = 0,2822 berada dalam rentang 0,2–0,4, mengindikasikan kecocokan model yang baik. Model menjelaskan 28,22% variasi dalam log-likelihood relatif terhadap model nol, mengonfirmasi daya prediksi yang wajar dari kombinasi prediktor akademik dan finansial.

Category	Predictor	Beta	SE	z	W^2	p-value	Significant
Graduate	(Intercept)	-5.1796	0.9047	-5.7254	32.7800	0.0000	Yes (*)
Graduate	admission_grade	0.0184	0.0033	5.6243	31.6332	0.0000	Yes (*)
Graduate	cu1_grade	0.0564	0.0212	2.6639	7.0963	0.0077	Yes (*)
Graduate	cu2_grade	0.0958	0.0212	4.5103	20.3425	0.0000	Yes (*)
Graduate	log_age_enrollment	-0.0384	0.1880	-0.2044	0.0418	0.8381	No
Graduate	log_unemployment_rate	0.4315	0.1922	2.2451	5.0407	0.0248	Yes (*)
Graduate	gdp	0.0236	0.0204	1.1590	1.3434	0.2464	No
Graduate	gender1	-0.4621	0.0931	-4.9631	24.6322	0.0000	Yes (*)
Graduate	scholarship_holder1	1.0053	0.1097	9.1620	83.9414	0.0000	Yes (*)
Graduate	debtor1	-0.8320	0.1648	-5.0473	25.4747	0.0000	Yes (*)
Graduate	tuition_up_to_date1	1.0854	0.2597	4.1786	17.4610	0.0000	Yes (*)
Dropout	(Intercept)	-1.7790	0.9604	-1.8524	3.4314	0.0640	No
Dropout	admission_grade	-0.0032	0.0036	-0.8893	0.7909	0.3738	No
Dropout	cu1_grade	0.0300	0.0190	1.5804	2.4975	0.1140	No
Dropout	cu2_grade	-0.1867	0.0183	-10.1801	103.6347	0.0000	Yes (*)
Dropout	log_age_enrollment	1.4521	0.1967	7.3822	54.4975	0.0000	Yes (*)
Dropout	log_unemployment_rate	0.4768	0.2212	2.1557	4.6469	0.0311	Yes (*)
Dropout	gdp	-0.0034	0.0238	-0.1424	0.0203	0.8868	No
Dropout	gender1	0.1570	0.1042	1.5066	2.2699	0.1319	No
Dropout	scholarship_holder1	-0.3064	0.1492	-2.0535	4.2169	0.0400	Yes (*)
Dropout	debtor1	0.0571	0.1598	0.3577	0.1279	0.7206	No
Dropout	tuition_up_to_date1	-1.8975	0.1882	-10.0821	101.6495	0.0000	Yes (*)

Figure 8: Uji Wald Parsial per Prediktor

Category	Predictor	Log-Odds (β)	SE	p-value	CI 2.5%	CI 97.5%
Graduate	(Intercept)	-5.1796	0.9047	0.0000	-6.9528	-3.4065
Graduate	admission_grade	0.0184	0.0033	0.0000	0.0120	0.0248
Graduate	cu1_grade	0.0564	0.0212	0.0077	0.0149	0.0979
Graduate	cu2_grade	0.0958	0.0212	0.0000	0.0542	0.1374
Graduate	log_age_enrollment	-0.0384	0.1880	0.8381	-0.4069	0.3301
Graduate	log_unemployment_rate	0.4315	0.1922	0.0248	0.0548	0.8082
Graduate	gdp	0.0236	0.0204	0.2464	-0.0163	0.0635
Graduate	gender1	-0.4621	0.0931	0.0000	-0.6446	-0.2796
Graduate	scholarship_holder1	1.0053	0.1097	0.0000	0.7903	1.2204
Graduate	debtor1	-0.8320	0.1648	0.0000	-1.1551	-0.5089
Graduate	tuition_up_to_date1	1.0854	0.2597	0.0000	0.5763	1.5945
Dropout	(Intercept)	-1.7790	0.9604	0.0640	-3.6614	0.1033
Dropout	admission_grade	-0.0032	0.0036	0.3738	-0.0104	0.0039
Dropout	cu1_grade	0.0300	0.0190	0.1140	-0.0072	0.0671
Dropout	cu2_grade	-0.1867	0.0183	0.0000	-0.2226	-0.1508
Dropout	log_age_enrollment	1.4521	0.1967	0.0000	1.0666	1.8377
Dropout	log_unemployment_rate	0.4768	0.2212	0.0311	0.0433	0.9103
Dropout	gdp	-0.0034	0.0238	0.8868	-0.0500	0.0432
Dropout	gender1	0.1570	0.1042	0.1319	-0.0472	0.3612
Dropout	scholarship_holder1	-0.3064	0.1492	0.0400	-0.5989	-0.0140
Dropout	debtor1	0.0571	0.1598	0.7206	-0.2560	0.3703
Dropout	tuition_up_to_date1	-1.8975	0.1882	0.0000	-2.2664	-1.5286

Figure 9: Koefisien Log-Odds Relatif

3.4.6 Interpretasi Model

Log-Odds Relatif. Berdasarkan Gambar 9:

- Untuk *Graduate*, efek positif terkuat berasal dari *tuition_up_to_date* ($\beta = 1,085$) dan *scholarship_holder* ($\beta = 1,005$), sementara *debtor* ($\beta = -0,832$) dan *gender* laki-laki ($\beta = -0,462$) mengurangi log-odds kelulusan.
- Untuk *Dropout*, *tuition_up_to_date* ($\beta = -1,898$) memiliki efek negatif terbesar, sementara *log_age_enrollment* ($\beta = 1,452$) memiliki efek positif terbesar, yang artinya, mahasiswa yang lebih tua dengan UKT menunggak berisiko paling tinggi untuk *dropout*.

Category	Predictor	RRR (e^{β})	% Change in Odds	Direction	p-value	CI 2.5%	CI 97.5%
Graduate	(Intercept)	0.0056	99.4370	-	0.0000	0.0010	0.0332
Graduate	admission_grade	1.0185	1.8550	+	0.0000	1.0120	1.0251
Graduate	cu1_grade	1.0580	5.8014	+	0.0077	1.0150	1.1028
Graduate	cu2_grade	1.1005	10.0517	+	0.0000	1.0557	1.1473
Graduate	log_age_enrollment	0.9623	3.7694	-	0.8381	0.6657	1.3911
Graduate	log_unemployment_rate	1.5396	53.9598	+	0.0248	1.0563	2.2439
Graduate	gdp	1.0239	2.3882	+	0.2464	0.9838	1.0656
Graduate	gender1	0.6300	37.0047	-	0.0000	0.5249	0.7561
Graduate	scholarship_holder1	2.7328	173.2845	+	0.0000	2.2040	3.3886
Graduate	debtor1	0.4352	56.4824	-	0.0000	0.3150	0.6011
Graduate	tuition_up_to_date1	2.9606	196.0581	+	0.0000	1.7794	4.9258
Dropout	(Intercept)	0.1688	83.1201	-	0.0640	0.0257	1.1088
Dropout	admission_grade	0.9968	0.3226	-	0.3738	0.9897	1.0039
Dropout	cu1_grade	1.0304	3.0414	+	0.1140	0.9928	1.0694
Dropout	cu2_grade	0.8297	17.0308	-	0.0000	0.8004	0.8601
Dropout	log_age_enrollment	4.2722	327.2155	+	0.0000	2.9054	6.2818
Dropout	log_unemployment_rate	1.6109	61.0862	+	0.0311	1.0442	2.4850
Dropout	gdp	0.9966	0.3378	-	0.8868	0.9513	1.0441
Dropout	gender1	1.1700	16.9967	+	0.1319	0.9539	1.4350
Dropout	scholarship_holder1	0.7361	26.3934	-	0.0400	0.5494	0.9861
Dropout	debtor1	1.0588	5.8811	+	0.7206	0.7741	1.4481
Dropout	tuition_up_to_date1	0.1499	85.0058	-	0.0000	0.1037	0.2168

Figure 10: *Relative Risk Ratios* ($RRR = e^{\beta}$) dan Persentase Perubahan *Odds*

Relative Risk Ratios (RRR). Gambar 10 menunjukkan bahwa:

- Mahasiswa dengan UKT lunas memiliki peluang kelulusan hampir tiga kali lipat ($RRR = 2,96$) dan hanya 15% peluang *Dropout* ($RRR = 0,15$), merupakan efek terbesar dalam model.
- Penerima beasiswa memiliki peluang kelulusan 2,73 kali lebih tinggi dan peluang *Dropout* 26,4% lebih rendah.
- Mahasiswa *debtor* memiliki peluang kelulusan kurang dari separuh ($RRR = 0,44$), memotong peluang kelulusan sebesar 56,48%.
- Usia yang lebih tua saat pendaftaran melipatgandakan peluang *Dropout* lebih dari empat kali lipat ($RRR = 4,27$, peluang meningkat 327%).
- Mahasiswa laki-laki memiliki peluang kelulusan 37% lebih rendah dibandingkan perempuan.

Table 11: Contoh AME pada Probabilitas Prediksi (*admission_grade*)

Kelompok	Estimate	Std. Error	z	p -value	S	CI 2,5%	CI 97,5%
<i>Enrolled</i>	-0,0015	0,0004	-3,6900	< 0,001	12,1000	-0,0023	-0,0007
<i>Graduate</i>	0,0032	0,0004	7,2800	< 0,001	41,4000	0,0023	0,0040
<i>Dropout</i>	-0,0017	0,0004	-4,6200	< 0,001	18,0000	-0,0024	-0,0010

Average Marginal Effects (AME). Catatan: *term* = *admission_grade*, *type* = *probs*, *comparison* = +1.

Tabel 11 menunjukkan contoh AME untuk prediktor *admission_grade* pada ketiga kategori keluaran.

Berdasarkan hasil AME secara keseluruhan:

- Efek tunggal terbesar dalam model adalah *tuition_up_to_date* pada *Dropout* (AME = -0,4129), diikuti oleh efeknya pada *Graduate* (+0,3480).
- Di antara prediktor numerik, *cu2_grade* memiliki dampak terkuat (+0,0278 pada *Graduate*; -0,0276 pada *Dropout*), sementara *log_age_enrollment* memiliki efek terbesar pada *Dropout* (+0,2095) dan efek negatif substansial pada *Graduate* (-0,1402).
- Beasiswa meningkatkan kelulusan sebesar +18,29 pp dan mengurangi *Dropout* sebesar -10,35 pp.
- Status *debtor* mengurangi kelulusan sebesar -14,41 pp.
- Mahasiswa laki-laki memiliki probabilitas kelulusan -8,82 pp dan probabilitas *Dropout* +5,26 pp dibandingkan perempuan.
- *gdp* tidak memiliki efek marginal yang signifikan pada keluaran apa pun.

Catatan: pp = *predicted probability* (probabilitas prediksi).

3.4.7 Simpulan RLM

Secara umum, model RLM signifikan secara statistik dan mencapai kecocokan yang baik (R^2 McFadden = 0,2822). Stabilitas finansial, khususnya status pembayaran UKT, merupakan pendorong utama hasil akademik mahasiswa, diikuti oleh kinerja akademik (*cu2_grade*), usia saat pendaftaran, dan status beasiswa. Variabel makroekonomi memiliki efek marginal yang minimal. Mahasiswa laki-laki menghadapi probabilitas kelulusan yang lebih rendah dan probabilitas *dropout* yang lebih tinggi dibandingkan perempuan.

3.5 Evaluasi dan Perbandingan Visual

3.5.1 Matriks Konfusi dan APER

Matriks konfusi untuk kedua metode pada data latih disajikan di bawah ini.

Table 12: Matriks Konfusi Analisis Diskriminan

		Prediksi		
		<i>Enrolled</i>	<i>Graduate</i>	<i>Dropout</i>
Aktual	<i>Enrolled</i>	0	700	94
	<i>Graduate</i>	0	2.086	123
	<i>Dropout</i>	1	629	791

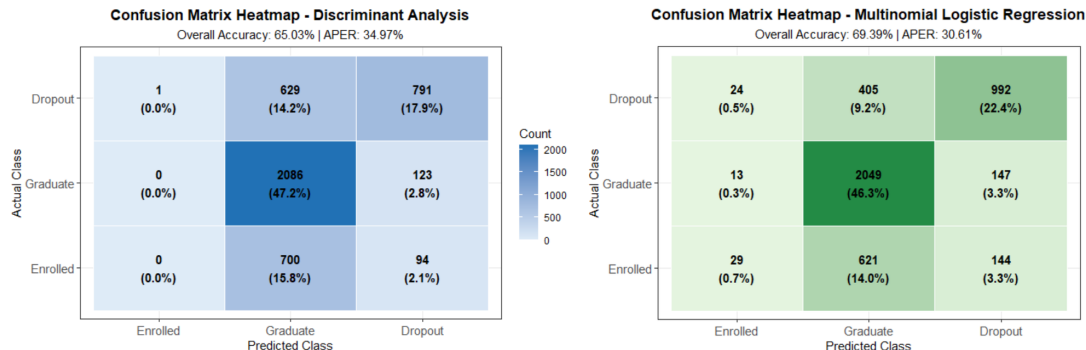
Table 13: Matriks Konfusi Regresi Logistik Multinomial

		Prediksi		
		<i>Enrolled</i>	<i>Graduate</i>	<i>Dropout</i>
Aktual	<i>Enrolled</i>	29	621	144
	<i>Graduate</i>	13	2.049	147
	<i>Dropout</i>	24	405	992

Perbandingan kedua matriks (Tabel 12 dan Tabel 13) menunjukkan bahwa RLM mengungguli DA dengan akurasi keseluruhan 69,39% (APER = 30,61%) dibanding 65,03% DA (APER = 34,97%), selisih 4,36 poin persentase. Perbedaan terkritik terletak pada kelas *Dropout*: RLM berhasil mengidentifikasi 992 dari 1.421 mahasiswa *Dropout* ($recall = 69,81\%$) dibandingkan hanya 791 pada DA ($recall = 55,67\%$), sebuah peningkatan 14,14 poin persentase. Untuk kelas *Enrolled*, DA gagal mengklasifikasikan satu pun dari 794 mahasiswa dengan benar ($recall = 0,00\%$), sementara RLM berhasil mendeteksi 29 mahasiswa secara benar ($recall = 3,65\%$).

Kelas *Graduate* merupakan satu-satunya kategori di mana DA memiliki keunggulan marginal dalam $recall$ (94,43% vs. 92,76%). Misklasifikasi *Dropout*→*Graduate* juga berkurang signifikan dalam RLM (405 mahasiswa, 9,2%) dibandingkan DA (629 mahasiswa, 14,2%), mengindikasikan kemampuan RLM yang lebih kuat dalam membedakan mahasiswa *Dropout* dari *Graduate*.

3.5.2 Heatmap Matriks Konfusi

Figure 11: *Heatmap* Matriks Konfusi untuk DA dan RLM

Heatmap (Gambar 11) secara visual mengonfirmasi pola yang dijelaskan sebelumnya. *Heatmap* DA menunjukkan diagonal *Enrolled* yang sepenuhnya kosong dan sel *off-diagonal* *Dropout*→*Graduate* yang sangat gelap, sementara *heatmap* RLM menampilkan diagonal *Dropout* yang lebih gelap, misklasifikasi *off-diagonal* yang berkurang, dengan diagonal *Enrolled* kecil namun bukan nol yang muncul untuk pertama kalinya.

3.5.3 Metrik Per Kelas dan Perbandingan Keseluruhan

Tabel 14 mengonfirmasi keunggulan RLM di hampir seluruh metrik. Peningkatan paling signifikan terjadi pada kelas *Dropout* (F1-Score: 0,6618 → 0,7549) dan kelas *Enrolled* (F1-Score: 0,0000 → 0,0673), sementara kelas *Graduate* tetap kuat pada kedua metode (F1-Score: 0,8350 vs. 0,8426).

Table 14: Perbandingan Keseluruhan Analisis Diskriminan dan Regresi Logistik Multinomial

Metrik	DA	RLM
Akurasi Keseluruhan	65,03%	69,39%
APER	34,97%	30,61%
Presisi <i>Enrolled</i>	0,0000	0,4394
Presisi <i>Graduate</i>	0,7483	0,7717
Presisi <i>Dropout</i>	0,8896	0,8218
Recall <i>Enrolled</i>	0,0000	0,0365
Recall <i>Graduate</i>	0,9443	0,9276
Recall <i>Dropout</i>	0,5567	0,6981
F1-Score <i>Enrolled</i>	0,0000	0,0673
F1-Score <i>Graduate</i>	0,8350	0,8426
F1-Score <i>Dropout</i>	0,6618	0,7549

3.5.4 Perbandingan Kinerja Visual

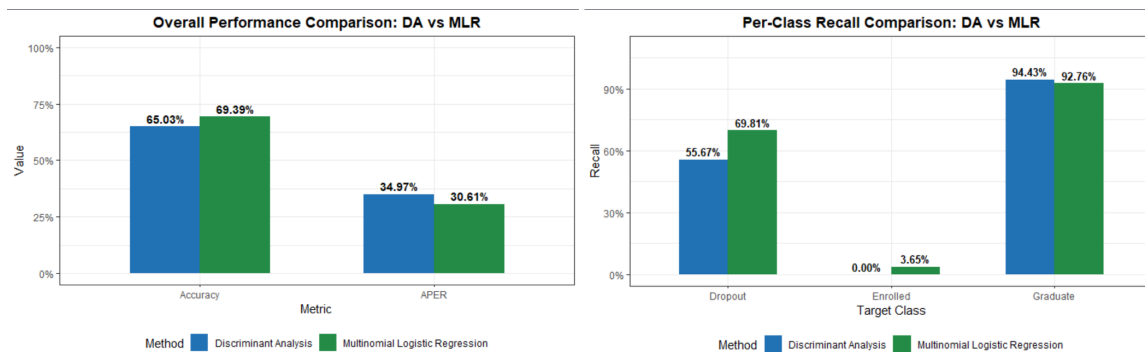


Figure 12: Perbandingan Kinerja Visual antara DA dan RLM

Gambar 12 mengonfirmasi bahwa RLM mengungguli DA dengan akurasi lebih tinggi (69,39% vs. 65,03%) dan APER lebih rendah (30,61% vs. 34,97%), mengindikasikan lebih sedikit kesalahan klasifikasi.

Perbandingan *recall* per kelas mengungkapkan bahwa peningkatan terbesar RLM terjadi pada kelas **Dropout**, di mana *recall* meningkat secara substansial dari 55,67% menjadi 69,81%. Peningkatan ini sangat penting karena mengidentifikasi mahasiswa yang berisiko *dropout* dengan benar merupakan tujuan utama penelitian dan memiliki implikasi praktis yang lebih besar untuk strategi intervensi dini. Untuk kelas **Enrolled**, kedua metode menunjukkan kinerja yang buruk, meskipun RLM menunjukkan sedikit peningkatan dari 0,00% menjadi 3,65%, mengindikasikan bahwa kategori ini tetap sulit diklasifikasikan secara akurat. Sementara itu, kelas **Graduate** mempertahankan *recall* tertinggi pada kedua metode, meskipun terdapat sedikit penurunan kinerja RLM dari 94,43% menjadi 92,76%.

Secara keseluruhan, perbandingan visual menunjukkan bahwa RLM mencapai kinerja klasifikasi yang lebih seimbang antar-kelas, khususnya dengan meningkatkan sensitivitas terhadap kelompok minoritas *Dropout* yang secara praktis penting.

4 Simpulan dan Rekomendasi

Berdasarkan hasil dan pembahasan yang telah diuraikan, simpulan penelitian ini disusun untuk menjawab ketiga rumusan masalah, sebagai berikut:

1. **Analisis Diskriminan mampu mengklasifikasikan mahasiswa secara signifikan ke dalam tiga kategori hasil akademik.** Uji Wilks' Lambda menunjukkan bahwa LDA signifikan secara statistik (Wilks' $\Lambda = 0,6213$, $p < 0,0001$), mengonfirmasi bahwa ketiga kelompok mahasiswa (*Enrolled*, *Graduate*, *Dropout*) dapat dibedakan menggunakan enam prediktor numerik, dengan nilai semester dan usia pendaftaran sebagai pendorong utama pemisahan kelompok. Meskipun demikian, signifikansi statistik ini tidak sepenuhnya tercermin pada kinerja klasifikasinya: LDA hanya mencapai akurasi 65,03% dan gagal sepenuhnya mengklasifikasikan kelompok *Enrolled* ($recall = 0,00\%$), sehingga LDA lebih baik dalam mendeteksi perbedaan kelompok secara statistik daripada dalam memisahkan seluruh kelas secara praktis.
2. **Regresi Logistik Multinomial mampu memodelkan probabilitas keanggotaan mahasiswa pada setiap kategori hasil akademik secara efektif serta mengidentifikasi prediktor yang signifikan.** Uji simultan (*Likelihood Ratio Test*) menunjukkan model RLM signifikan secara statistik ($G^2 = 2546,486$, $db = 20$, $p \approx 0$) dengan kecocokan model yang baik (R^2 McFadden = 0,2822). RLM berhasil mengidentifikasi beberapa prediktor yang signifikan memengaruhi probabilitas *Graduate* maupun *Dropout* relatif terhadap *Enrolled*, dengan status pembayaran UKT (*tuition_up_to_date*) sebagai prediktor paling berpengaruh, diikuti oleh usia pendaftaran, nilai semester kedua (*cu2_grade*), dan status beasiswa, sementara variabel makroekonomi (*gdp*) memiliki pengaruh minimal.
3. **Regresi Logistik Multinomial memiliki kinerja yang lebih baik dibandingkan Analisis Diskriminan dalam mengklasifikasikan hasil akademik mahasiswa.** RLM mengungguli LDA pada hampir seluruh metrik evaluasi, dengan akurasi keseluruhan lebih tinggi (69,39% vs. 65,03%), APER lebih rendah (30,61% vs. 34,97%), serta *recall* kelas *Dropout* yang jauh lebih baik (69,81% vs. 55,67%). Keunggulan ini terutama disebabkan oleh kemampuan RLM mengintegrasikan prediktor biner finansial (*gender*, *scholarship_holder*, *debtor*, *tuition_up_to_date*) yang tidak dapat disertakan dalam LDA akibat pelanggaran asumsi normalitas multivariat.

Berdasarkan simpulan tersebut, RLM direkomendasikan sebagai metode yang lebih sesuai untuk klasifikasi hasil akademik pada *dataset* ini. Institusi pendidikan tinggi disarankan untuk memantau status pembayaran UKT dan usia pendaftaran sebagai indikator risiko *dropout* dini guna memandu strategi intervensi yang tepat sasaran, serta mempertimbangkan penambahan prediktor lain atau metode klasifikasi berbasis *machine learning* pada penelitian selanjutnya untuk memperbaiki identifikasi kelompok *Enrolled* yang masih lemah pada kedua metode.

Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada Laboratorium Analisis Big Data yang telah memberikan dukungan, bantuan, dan kontribusi dalam pelaksanaan penelitian ini. Penulis juga berterima kasih kepada Program Studi S1 Sains Data, Universitas Negeri Surabaya atas fasilitas dan dukungan akademik yang diberikan selama proses penelitian.

References

- Abubakar, A. et al. (2024). A comprehensive review on discriminant analysis for addressing challenges of class-level limitations, small sample size, and robustness. *Processes*, 12(7):1382.
- Arel-Bundock, V., Greifer, N., and Heiss, A. (2024). How to interpret statistical models using marginaleffects. *Journal of Statistical Software*, 111(9):1–32.
- Barile, e. a. (2025). Predicting dropout from higher education with multinomial finite mixture models: Empirical evidence from Italy. *Quality & Quantity*.
- Contreras Montes, J. A. et al. (2023). Drop-out prediction in higher education using imbalanced multiclass dataset. *Journal for ReAttach Therapy and Developmental Diversities*.
- Graf, A. C. et al. (2023). Comparing linear discriminant analysis and supervised learning algorithms for binary classification: A method comparison study. *Biometrical Journal*, 65(3):e2200098.
- Mwavita, e. a. (2023). Application of discriminant analysis to predict students' performances in Mathematics. *European Journal of Statistics*.
- Realinho, V. et al. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository.
- Sathyanarayanan, S. and Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*.