LAPORAN PROJECT ANALISIS MULTIVARIAT
SEMESTER GENAP 2024/2025

IDENTITAS PROYEK	
Judul	Perbandingan Regresi Logistik dan Analisis Diskriminan Linear dalam Memprediksi kelulusan mahasiswa
Topik	2
Identitas Penyusun	1. Shinta Usaila Farachin (23031554160) 2. Ikhrima Atusifah (23031554181) 3. Arina Tri Yuni Wahyuning Tiyas (23031554203)
Kelas	2023A

1. PENDAHULUAN

Pendahuluan penelitian 250 - 1000 kata..

1.1. Latar Belakang (min. 250 kata)

Kelulusan bagi mahasiswa merupakan salah satu aspek penting kelulusan dalam menyelesaikan sistem pendidikan tinggi. Kegagalan mahasiswa dalam menyelesaikan pembelajaran dapat memberikan hal negatif bagi mahasiswa baik secara individu maupun bagi universitas secara keseluruhan, hal ini termasuk menurunnya akreditasi dan efektivitas kampus [1]. Oleh karenanya, pengembangan sistem yang baik dapat mengidentifikasi potensi lulus atau tidaknya mahasiswa. Metode prediksi kelulusan dapat membantu suatu institusi untuk melakukan intervensi akademik atau administratif lebih awal sehingga dapat meminimalisir hal yang tidak diinginkan. Dua pendekatan statistik yang biasanya digunakan untuk membangun model prediksi kelulusan adalah regresi logistik dan analisis diskriminan. Regresi logistik cocok untuk memodelkan probabilitas dan variabel independen pada kategorikal biner, serta bersifat lebih fleksibel pada distribusi data [2]. Disisi lain, analisis diskriminan dapat bekerja optimal ketika asumsi normalitas dan homogenitas varians antar kelompok tercapai, dan dapat menghasilkan batas prediksi yang interpretatif [3]. Hal ini tentu saja menjadi topik menarik untuk dibuat perbandingan hasil keduanya.

Beberapa studi menunjukkan bahwa regresi logistik lebih memiliki keunggulan dalam kondisi data yang tidak berdistribusi normal, sedangkan analisis diskriminan lebih

unggul saat asumsi multivariat normal tercapai [4][5]. Oleh karena itu, cocok jika analisis diskriminan memberikan hasil lebih baik jika asumsi normalitasnya terpenuhi, tetapi dalam situasi lain, regresi logistik akan lebih sesuai [6]. Meskipun regresi logistik jauh lebih umum dan memiliki sejumlah sifat teoritis, analisis diskriminan harus menjadi pilihan yang lebih baik jika tahu bahwa populasi terdistribusi secara normal [7].

Pada konteks prediksi kelulusan mahasiswa, kedua pendekatan sudah diimplementasikan secara luas dengan hasil yang beragam tergantung pada struktur data, jumlah prediktor, dan kualitas variabel input. Oleh karena itu, penggunaan evaluasi secara komparatif dibutuhkan untuk mengetahui mode yang lebih akurat, stabil dan mudah diterapkan. Projek ini bertujuan untuk membandingkan hasil regresi logistik dan analisis diskriminan dalam prediksi kelulusan mahasiswa. Projek ini tidak hanya fokus pada akurasi prediksi tetapi juga pada interpretabilitas model dan kesesuaian asumsi terhadap suatu data.

1.2. Rumusan Masalah dan Tujuan

Rumusan masalah

1. Bagaimana performa model regresi logistik dalam prediksi kelulusan mahasiswa?
2. Bagaimana performa model analisis diskriminan dalam prediksi kelulusan mahasiswa?
3. Model mana yang memiliki tingkat lebih baik dalam prediksi kelulusan mahasiswa antara regresi logistik atau analisis diskriminan

Tujuan

2. Untuk mengevaluasi kinerja regresi logistik dalam prediksi kelulusan mahasiswa
3. Untuk mengevaluasi kinerja analisis diskriminan dalam prediksi kelulusan mahasiswa
4. Untuk membandingkan performa kedu model yaitu antara regresi logistik dan analisis diskriminan dalam prediksi kelulusan mahasiswa

5. Metodologi

Metodologi atau cara untuk mencapai tujuan yang telah ditetapkan ditulis tidak melebihi 1000 kata. Bagian ini berisi metode pre-processing dan/atau metode post processing yang dilengkapi dengan diagram alir penelitian

yang menggambarkan apa yang sudah dilaksanakan dan yang akan dikerjakan selama waktu yang diusulkan. Format diagram alir dapat berupa file JPG/PNG. Metode penelitian harus dibuat secara utuh dengan penahapan yang jelas.

5.1. Eksplorasi Dataset

Pemahaman dataset yang dimiliki

Dataset ini berasal dari sebuah lembaga pendidikan yang menyediakan informasi tentang sekitar 2.866 mahasiswa. Data ini mencangkup berbagai variabel yang terkait dengan prestasi mahasiswa, kinerja akademik, dan kondisi sosial dan ekonomi. Ada 61 kolom dalam kumpulan data ini, yang mencakup fitur-fitur seperti jenis kelamin, usia masuk, status pernikahan, kewarganegaraan, kegiatan akademik terakhir, status akademik, fakultas, nilai ujian masuk, jenis pembiayaan, dan apakah seorang mahasiswa menerima beasiswa. Kolom tujuan dalam kumpulan data ini adalah "students dropout," yang menunjukkan apakah mahasiswa tersebut masih aktif, telah putus sekolah, atau telah lulus. Kumpulan data ini memberikan informasi berharga untuk analisis prediktif keberhasilan program gelar master berdasarkan berbagai faktor demografi, akademik, dan ekonomi. Berikut link datasetnya https://www.scidb.cn/en/detail?dataSetId=6769bc48b5f14ab7b2409173ab2dd032

5.2. Langkah Penelitian

Dalam melakukan Perbandingan Regresi Logistik dan Analisis Diskriminan Linear dalam Memprediksi kelulusan mahasiswa, diperlukan beberapa tahapan sistematis untuk mendapat hasil yang baik dan efisien. Berdasarkan studi Fortune dkk [3] serta berbagai referensi lain, berikut langkah-langkah yang dilakukan:

1. Pengumpulan data dan pemahaman dataset

   Tahap pertama dalam projek ini yaitu mengumpulkan data yang akan dianalisis. Dalam projek ini data didapat dari mendelay dengan judul student droput. Dataset ini akan digunakan untuk membangun model prediksi kelulusan mahasiswa.

2. Data Preprocessing

   Tahapan preprocessing merupakan salah satu tahapan yang penting untuk projek ini, tahapan ini penting untuk meminimalisir kesalahan atau noise yang dapat mengganggu akurasi model [8]. Selain itu, kualitas preprocessing akan mempengaruhi performa model [9]Pada tahapan ini data akan dilakukan penanganan missing value dan duplikat, konversi target data. Pada

penanganan missing value dan duplikat tidak ada missing value dan duplikat baris, untuk konversi target pada dataset semua data telah berupa numeric sehingga tidak dilakukan konversi target.

3. Statistika Deskriptif

   Pada tahap ini data akan dilakukan ukuran pemusatan (mean, median, modus), melihat ukuran penyebaran (range, standar deviasi), serta distribusi data. Kemudian pada proses ini juga kami melihat korelasi data untuk melihat variabel atau fitur utama yang bisa kita pilih untuk melakukan regresi logistik dan analisis diskriminan.

4. Pemisahan data latih dan uji

   Dataset ini nanti akan dibagi menjadi 2 bagian yaitu 70% data train dan 30% data test. Pembagian ini cukup penting untuk memastikan validitas model [10].

5. Model Regresi Logistik

   Pada tahapan ini data akan diprediksi kan menjadi biner untuk nanti digunakan regresi logistik. Pemilihan regresi logistik dikarenakan telah terbukti efektif dalam memodelkan dropout mahasiswa.

6. Model analisis diskriminan

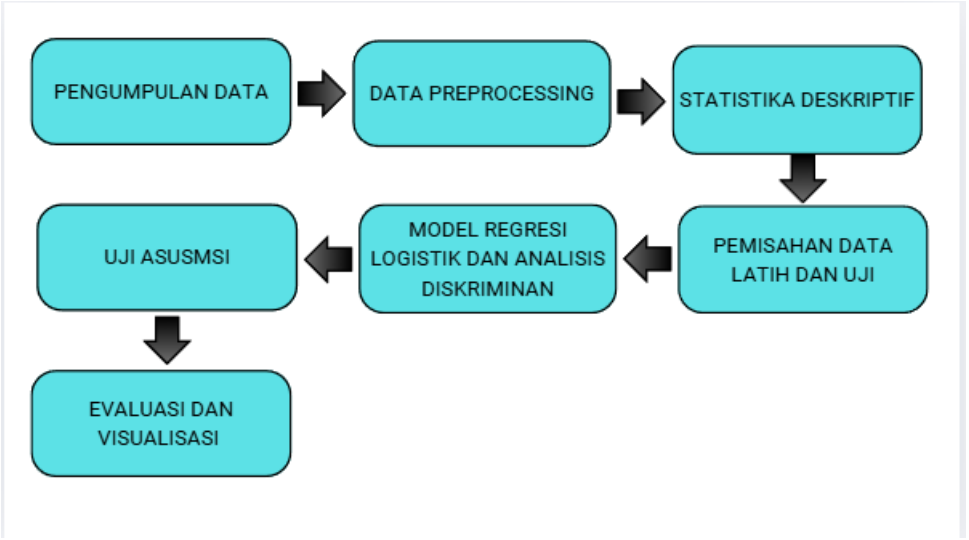
   Model analisis diskriminan dibangun menggunakan seluruh kelas target. Model ini sering digunakan dalam prediksi pendidikan karena kemampuannya dalam menangani banyak kelas.

4. Uji asumsi

   Uji asumsi dilakukan dari dua model yaitu regresi logistik dan analisis diskriminan, uji ini akan memuat uji multikolinearitas, normalitas residual serta uji Hosmer-Lemeshow test untuk regresi logistik dan multivariat normal, normalitas univariat uji box's m serta, Korelasi antar variabel prediktor untuk analisis diskriminan.

8. Evaluasi dan visualisasi

   Pada tahap ini nantinya semua model akan dievaluasi menggunakan confusion matrix, yang dapat memberikan insight pada akurasi dan kesalahan prediksi . Kemudian akan ditambahkan visualisasi untuk membantu hasil interpretasi model.



3. Hasil dan Analisis

3.1 Analisis Statistika Deskriptif

Dari analisis statistik deskriptif untuk 2.866 observasi mahasiswa, kita mendapatkan beberapa wawasan penting mengenai aspek demografis, akademik, dan finansial mereka. Rata-rata umur saat memulai kuliah adalah sekitar 18,74 tahun, sementara usia di semester saat ini adalah sekitar 19,43 tahun. Variabel seperti start_page menunjukkan skewness negatif yang tinggi (-2,78) dan kurtosis yang sangat tinggi (47,90), menandakan adanya beberapa outlier, kemungkinan mahasiswa yang jauh lebih muda atau jauh lebih tua dibandingkan mayoritas. Sebagian besar mahasiswa berstatus lajang (marital status = 6, median), berasal dari kewarganegaraan lokal (citizenship = 1), dan terdaftar pada semester 202020–202120. Ada dominasi gender yang cukup seimbang (rata-rata gender = 1,51, di mana nilai 1 dan 2 mungkin menunjukkan laki-laki dan perempuan), tetapi sebagian besar adalah gender 2 (karena median = 2).

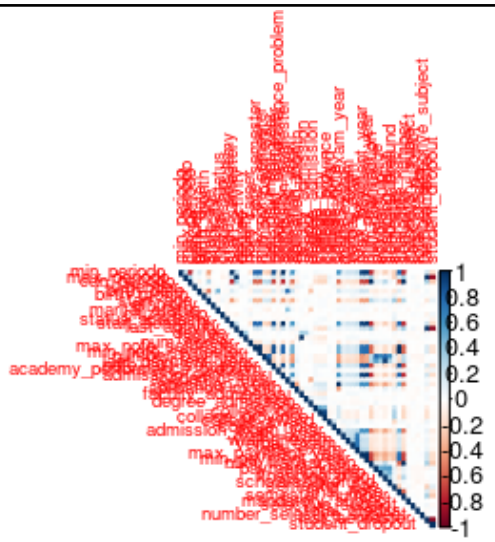
Secara akademik, nilai semester terakhir (note last semester) menunjukkan rata-rata yang tinggi (354,14) namun memiliki simpangan baku yang besar (476,88), serta menunjukkan skewness positif (0,61) yang menunjukkan bahwa sebagian besar nilai berada di bawah rata-rata, dengan beberapa nilai yang sangat tinggi. Fenomena yang sama terlihat pada final note, max note semester, dan min note semester yang memiliki nilai tertinggi 999 ada kemungkinan data ini memuat nilai placeholder atau ketidaknormalan. Distribusi academy performance problem sangat mencolok, dengan

rata-rata hanya 7,45 tetapi skewness sangat tinggi (11,55) dan kurtosis ekstrem (131,39), menunjukkan bahwa sebagian besar mahasiswa tidak menghadapi masalah akademik, tetapi terdapat beberapa kasus yang ekstrem.

Dari segi finansial, sebagian besar mahasiswa memiliki jumlah pembayaran (number payment) yang rendah (median = 0), serta nilai pembayaran (payment value) yang sangat bervariasi (min = 1, max = 119.615), dengan distribusi yang sangat miring ke kanan (skewness = 10,47). Ini menunjukkan bahwa hanya sedikit mahasiswa yang membayar dalam jumlah besar, mungkin disebabkan oleh beasiswa (rata-rata scholarship = 0,01) atau rencana pembiayaan lainnya. Distribusi ujian masuk (matematika, menulis, verbal) terlihat cukup normal, dengan rata-rata nilai sekitar 700 dari maksimum 1000. Akan tetapi, total nilai ujian (total exam) menunjukkan distribusi yang hampir normal (skewness = 0,10; kurtosis = -0,23), yang mengindikasikan bahwa seleksi awal mungkin cukup seragam.

Sebagai hasil akhirnya, variabel yang berhubungan dengan akademik seperti mandatory subject, elective subject, dan total credits memperlihatkan nilai rata-rata yang wajar, dan distribusinya cukup simetris. Namun, variabel seperti school, degree admission, dan college award memiliki nilai maksimum yang tidak normal (999), yang mungkin menandakan pengkodean untuk tidak diketahui atau kosong. Secara umum, analisis ini mengindikasikan bahwa dataset ini relatif bersih dan kaya informasi, namun terdapat beberapa variabel yang memiliki outlier ekstrim atau potensi kesalahan dalam entri data (seperti nilai 999).

Pada korelasi matrik didapatkan keterkaitan di antara berbagai variabel numerik dalam kumpulan data, di mana sebagian besar variabel menunjukkan hubungan yang lemah satu sama lain, terlihat dari warna-warna yang tidak mencolok. Hanya sejumlah kecil pasangan variabel yang menunjukkan korelasi yang sedang hingga kuat, baik dalam arah positif maupun negatif. Pola ini menunjukkan bahwa sebagian besar fitur memiliki sifat independen secara linier, sehingga dapat digunakan bersamaan dalam model prediksi tanpa kekhawatiran mengenai multikolinearitas yang berlebihan. Visualisasi ini sangat berguna dalam proses pemilihan fitur dan analisis lebih lanjut mengenai struktur data.



Kemudian pada deskripsi analisis kami hanya menggunakan 6 fitur utama yaitu final note, semester number, total credits, math exam, verbal exam, payment value dengan alasan Enam fitur dipilih karena mewakili aspek penting yang mempengaruhi drop out mahasiswa, yaitu kinerja akademik (final note, math exam, verbal exam), progres studi (semester number, total credits), dan kondisi keuangan (payment value). Jumlah fitur dibatasi agar model tetap sederhana, menghindari overfitting, dan memenuhi asumsi analisis diskriminan linear (LDA), seperti normalitas multivariat dan homogenitas kovarians. Selain itu, dengan sedikit fitur, interpretasi dan visualisasi hasil menjadi lebih mudah dan informatif. Kemudian untuk penanganan missing value kami menggunakan median karena lebih fleksibel dalam menangani banyak data.

3.2 Uji Asumsi (jika ada)

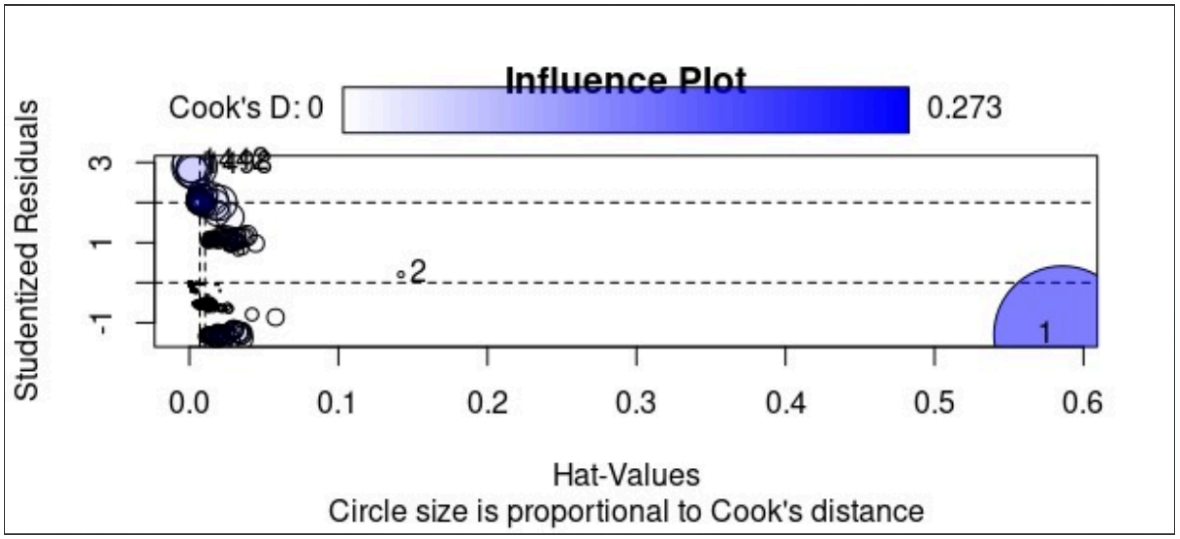
REGRESI LOGISTIK

Uji multikolinearitas

Berdasarkan hasil asumsi model regresi logistik yang telah dibuat, ditemukan beberapa hal penting yang menunjukkan bahwa model tersebut memiliki kinerja yang cukup baik. Pertama, analisis Variance Inflation Factor (VIF) menunjukkan bahwa semua variabel prediktor memiliki nilai VIF di bawah 1,1. Ini menunjukkan bahwa tidak terdapat masalah multikolinearitas yang berarti di antara variabel-variabel prediktor, sehingga semua variabel tersebut dapat digunakan dengan aman dalam model regresi logistik.

Influence plot

Selanjutnya, pada influence plot,sejumlah observasi (seperti observasi ke-1, 2, 1442, dan 1498) menunjukkan adanya pengaruh yang lebih besar terhadap model, berdasarkan nilai leverage atau Cook's Distance yang cenderung tinggi. Walau begitu, nilai Cook's Distance itu masih jauh di bawah batas kritis umum (yaitu 1), sehingga tidak terdapat indikasi bahwa model secara keseluruhan terpengaruh oleh pengamatan-pengamatan ekstrem tersebut.

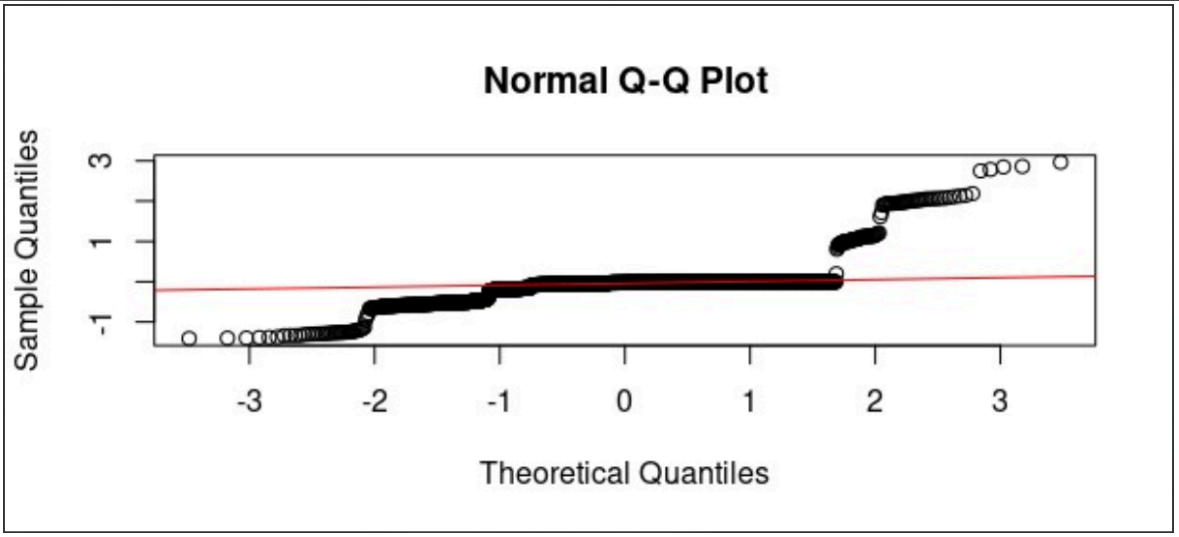


Uji Hosmer-Lemeshow

Dalam hal uji kesesuaian model (goodness-of-fit) menerapkan uji Hosmer-Lemeshow, diperoleh nilai p-value sebesar 0.9982. Angka ini sangat tinggi dan melebihi batas 0.05, yang menunjukkan bahwa tidak ada perbedaan signifikan antara model prediksi dan data yang diamati. Dengan kata lain, model regresi logistik yang diterapkan menunjukkan kesesuaian yang sangat baik dengan data, atau dengan kata lain, model ini secara statistik sesuai.

QQ plot residual deviance

Untuk penilaian residual, dilakukan representasi melalui QQ Plot terhadap deviance residual. Visualisasi ini umumnya dimanfaatkan untuk mengevaluasi normalitas residual, walaupun pada regresi logistik, asumsi normalitas residual tidak sekrusial pada regresi linear. Dengan demikian, adanya penyimpangan kecil dalam distribusi residual tidak menjadi isu yang signifikan.



ANALISIS DISKRIMINAN

Uji normalitas univariat shapiro-wilk

Sementara itu, dalam analisis Linear Discriminant Analysis (LDA) ,Berdasarkan hasil pengujian normalitas univariat Shapiro-Wilk, dapat disimpulkan bahwa mayoritas variabel prediktor tidak berdistribusi normal dalam setiap kelompok kelas target. Nilai P yang sangat rendah ($p < 0.05$) pada hampir semua kombinasi variabel dan kelompok mengindikasikan penolakan hipotesis nol normalitas. Hanya ada dua kombinasi variabel dan kelas yang menunjukkan kemungkinan normalitas: total\credits dalam kelompok dropout ($p = 0.0559$) dan math\exam dalam kelompok dropout ($p = 0.2022$). Ini mengindikasikan bahwa asumsi normalitas univariat pada dataset pelatihan umumnya tidak terpenuhi.

Uji normalitas multivariat Mardia

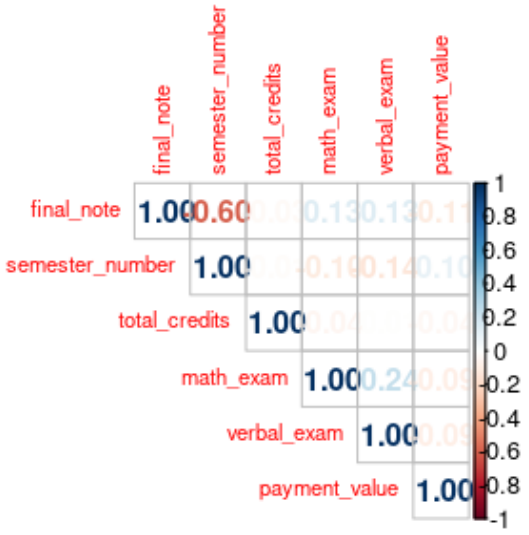
Selanjutnya, hasil dari uji Mardia yang diterapkan untuk mengevaluasi normalitas multivariat mengindikasikan bahwa data tidak memenuhi asumsi normalitas multivariat, baik dari segi skewness maupun kurtosis, dengan p-value = 0 pada kedua komponen. Oleh karena itu, model analisis diskriminan (seperti LDA) yang bergantung pada asumsi normalitas multivariat mungkin tidak ideal jika digunakan tanpa transformasi atau modifikasi metode.

Uji box's m

Selanjutnya, uji Box's M memperlihatkan nilai p yang sangat rendah (p < 2.2e-16), yang menunjukkan bahwa asumsi kesamaan matriks kovarians di antara kelompok tidak terpenuhi. Ini juga merupakan pertimbangan penting karena Linear Discriminant Analysis (LDA) mengasumsikan bahwa kelompok-kelompok memiliki kovarians yang serupa. Pelanggaran terhadap asumsi ini bisa membuat model LDA menjadi kurang tepat atau terpengaruh.

Korelasi antar variabel prediktor

Dari sudut pandang korelasi di antara variabel prediktor, memperlihatkan bahwa sebagian besar variabel dalam data memiliki hubungan yang lemah atau hampir tidak ada korelasi antara satu sama lain, kecuali untuk final note dan semester number yang menunjukkan korelasi positif sedang sekitar 0,60. Ini menunjukkan bahwa semakin banyak semester yang dijalani, ada kecenderungan untuk mendapatkan nilai akhir yang lebih tinggi. Di sisi lain, variabel lainnya seperti math exam, verbal exam, total credits, dan payment value tidak menunjukkan hubungan linier yang berarti, sehingga diperlukan analisis tambahan untuk mengidentifikasi pola lain di luar korelasi linier.



Kesimpulan

Berdasarkan hasil pengujian asumsi tersebut, regresi logistik terlihat lebih cocok dan dapat diandalkan untuk memodelkan data ini dibandingkan LDA, mengingat asumsi

yang dipenuhi lebih komprehensif dan kuat. Pelanggaran terhadap asumsi normalitas dan homogenitas kovarians dalam LDA dapat mengakibatkan penurunan kinerja model atau interpretasi yang tidak akurat. Sebagai hasilnya, untuk analisis klasifikasi pada set data ini, regresi logistik menjadi pilihan yang lebih sesuai. Jika ingin terus menggunakan LDA, mungkin perlu mempertimbangkan transformasi data atau metode lain yang lebih toleran terhadap pelanggaran asumsi.

3.3 Analisis Hasil dan Pembahasan

Hasil jelaskan hasil koefisien dan signifikansi masing-masing variabel yang didapatkan

Regresi logistic

Hasil regresi logistik menunjukkan bahwa variabel yang paling berpengaruh secara signifikan terhadap kemungkinan mahasiswa mengalami putus studi adalah semester_number. Koefisien tersebut memiliki nilai negatif dan signifikan pada tingkat kepercayaan yang tinggi ($p < 0,01$), yang menunjukkan bahwa semakin banyak semester yang dijalani oleh mahasiswa, semakin rendah kemungkinan mereka untuk dropout. Secara logis, hal ini wajar karena mahasiswa di semester akhir biasanya sudah terlalu jauh untuk kembali, dan memiliki komitmen yang lebih tinggi untuk menyelesaikan pendidikan. Variabel lain yang turut berkontribusi meskipun tidak sekuat semester_number ialah final_note. Koefisiennya pun negatif, menunjukkan bahwa semakin tinggi nilai akhir mahasiswa, semakin rendah kemungkinan mereka untuk meninggalkan studi. Akan tetapi, tingkat signifikansinya tidak setinggi semester_number, sehingga dampaknya bersifat marginal. Sementara itu, variabel lain seperti total_credits, math_exam, verbal_exam, dan payment_value tidak menunjukkan pengaruh signifikan secara statistik terhadap kemungkinan terjadinya dropout. Ini mungkin disebabkan oleh ketidakselarasan skala, distribusi yang tidak merata, atau hubungan yang lemah dengan variabel target. Secara keseluruhan, model regresi logistik menunjukkan signifikansi indikator keberlanjutan akademik seperti semester aktif dan kinerja akademik terhadap risiko putus sekolah, serta merekomendasikan agar intervensi akademik ditujukan kepada mahasiswa di semester awal dengan kinerja rendah.

```
##
## Call:
## glm(formula = formula_log, family = "binomial", data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.635e+00  9.641e-01   2.733  0.00627 **
## final_note      -8.430e-03  4.518e-03  -1.866  0.06204 .
## semester_number -2.175e+00  2.075e-01 -10.485  < 2e-16 ***
## total_credits   -1.406e-03  2.374e-03  -0.592  0.55366
## math_exam        8.743e-04  1.002e-03   0.872  0.38307
## verbal_exam     -1.144e-03  1.027e-03  -1.114  0.26516
## payment_value    4.967e-05  2.407e-05   2.063  0.03908 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 746.82  on 2005  degrees of freedom
## Residual deviance: 381.63  on 1999  degrees of freedom
## AIC: 395.63
##
## Number of Fisher Scoring iterations: 11
```

Analisis Diskriminan

Dalam model Linear Discriminant Analysis (LDA), hasil yang didapatkan menunjukkan bahwa variabel yang paling berpengaruh terhadap pemisahan antara mahasiswa yang berhenti kuliah dan yang tidak adalah semester_number. Variabel ini memiliki koefisien terbesar dalam fungsi diskriminan, menunjukkan bahwa nilai semester dapat memisahkan kedua kelompok dengan efektif. Dengan kata lain, distribusi semester_number sangat bervariasi antara mahasiswa yang tetap dan mereka yang mengundurkan diri, dan menjadi dimensi kunci yang digunakan LDA untuk memproyeksikan data ke dalam ruang klasifikasi satu dimensi. Variabel final_note juga berperan dalam fungsi diskriminan, tetapi dengan proporsi yang lebih sedikit. Ini mengindikasikan bahwa meskipun prestasi akademik masih penting dalam membedakan kelompok, dampaknya tidak sebesarnya waktu belajar. Variabel lain seperti total_credits, math_exam, verbal_exam, dan payment_value memberikan kontribusi yang sangat minim terhadap fungsi diskriminan, yang menunjukkan bahwa nilai rata-ratanya tidak banyak berbeda di antara kelompok. Karena LDA tidak mengevaluasi signifikansi secara statistik seperti pada regresi logistik, penafsiran

berfokus pada daya pisah antar kelompok berdasar kombinasi linear variabel. Hasil LDA ini sejalan dengan model regresi logistik, dan menegaskan bahwa durasi studi serta kinerja akademik merupakan dua faktor utama dalam menentukan profil risiko dropout.

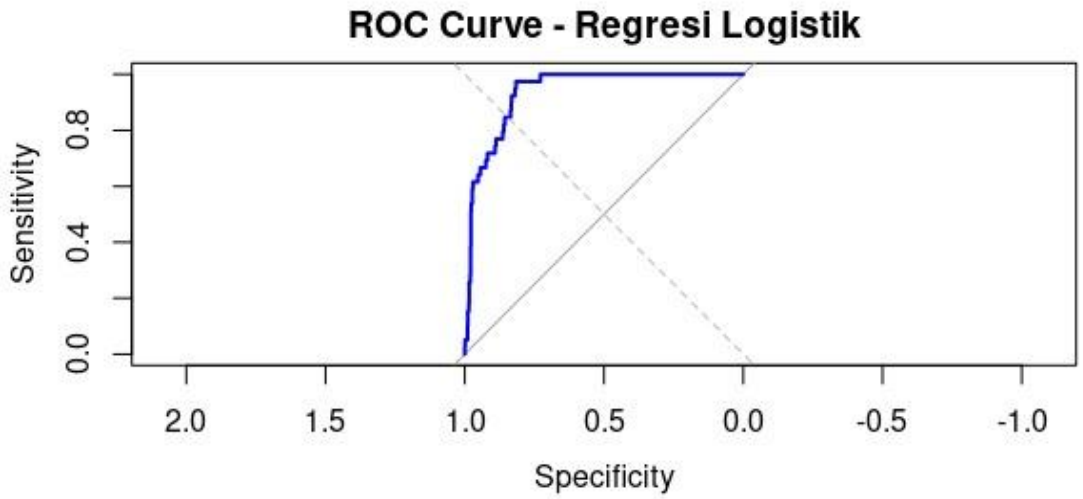
```
## Variabel: final_note, Group: 0 → p-value: 0.0000
## Variabel: final_note, Group: 1 → p-value: 0.0000
## Variabel: semester_number, Group: 0 → p-value: 0.0000
## Variabel: semester_number, Group: 1 → p-value: 0.0000
## Variabel: total_credits, Group: 0 → p-value: 0.0000
## Variabel: total_credits, Group: 1 → p-value: 0.0559
## Variabel: math_exam, Group: 0 → p-value: 0.0000
## Variabel: math_exam, Group: 1 → p-value: 0.2022
## Variabel: verbal_exam, Group: 0 → p-value: 0.0000
## Variabel: verbal_exam, Group: 1 → p-value: 0.0037
## Variabel: payment_value, Group: 0 → p-value: 0.0000
## Variabel: payment_value, Group: 1 → p-value: 0.0000
```

Hasil Evaluasi Model

Berdasarkan evaluasi yang dilakukan terhadap model logistic regression dan Linear Discriminant Analysis (LDA) pada data uji, keduanya menunjukkan akurasi yang tinggi, yakni 95,47% untuk logistic regression dan 94,88% untuk LDA. Walaupun akurasinya hampir sama, nilai Kappa mengindikasikan bahwa logistic regression (0,5175) memiliki kemampuan prediksi yang sedikit lebih unggul daripada LDA (0,4954), yang menunjukkan bahwa model ini lebih konsisten dalam membedakan dua kelas target. Dalam hal sensitivitas (kapasitas untuk mendeteksi kelas mayoritas yaitu ‘0’), keduanya sangat tinggi, dengan logistic regression mencapai 97,20% dan LDA mencapai 96,47%. Namun, untuk specificity (kemampuan dalam mendeteksi kelas minoritas yaitu ‘1’), LDA sedikit lebih baik (61,54%) dibandingkan dengan logistic regression (58,97%), meskipun keduanya masih tergolong rendah. Ini menunjukkan bahwa kedua model cukup efektif dalam mengidentifikasi mahasiswa yang tidak memiliki masalah, namun kurang optimal dalam menemukan mahasiswa yang memiliki masalah (target = 1).

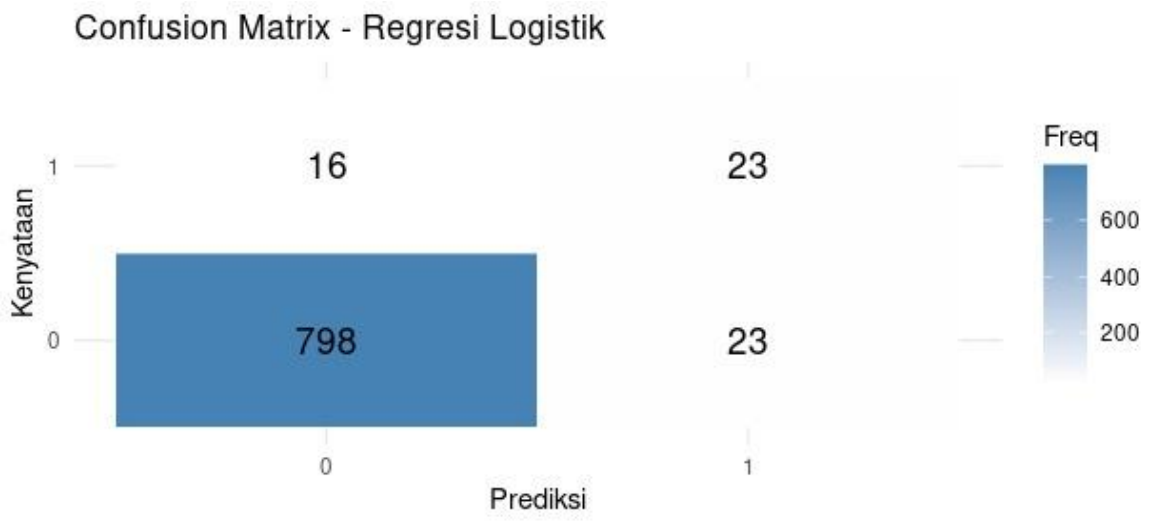
Nilai Balanced Accuracy, yang merupakan rata-rata dari sensitivitas dan spesifisitas, menunjukkan bahwa LDA memiliki kinerja yang sedikit lebih seimbang (79,00%) dibandingkan regresi logistik (78,09%), walaupun selisihnya kecil. Meskipun demikian, analisis McNemar mengindikasikan bahwa tidak terdapat perbedaan yang signifikan antara prediksi dari kedua model dan label yang sebenarnya, dengan p-value sebesar 0,3367 untuk logistic regression dan 0,05002 untuk LDA, yang mendekati batas signifikansi. Secara keseluruhan, kedua model berfungsi dengan baik dalam meramalkan target, namun memiliki batasan dalam mengenali kelas minoritas.

Untuk hasil plot kurva ROC model regresi logistik dapat dilihat sebagai berikut



Plot Kurva ROC untuk model Regresi Logistik menunjukkan kinerja klasifikasi yang sangat baik, ditandai dengan kurva yang secara signifikan naik ke arah sudut kiri atas grafik. Ini menunjukkan bahwa model dapat dengan baik membedakan antara siswa yang mengalami dropout dan yang tidak dropout. Tingkat sensitivitas yang tinggi pada berbagai tingkat spesifisitas menunjukkan bahwa model efektif dalam mengidentifikasi kasus dropout. Secara keseluruhan, kurva ini menggambarkan bahwa model regresi logistik memiliki kemampuan prediksi yang kuat dengan nilai AUC yang mungkin mendekati 1.0.

Gambar confusion matrix



Confusion matrix menunjukkan bahwa model regresi logistik menunjukkan kinerja yang sangat baik dalam mengidentifikasi mahasiswa yang tidak keluar (798 benar dari total 821), tetapi kurang efisien dalam mengenali mahasiswa yang benar-benar keluar, dengan hanya 16 kasus dropout yang terdeteksi dengan benar dari 39 kasus sebenarnya. Kesalahan dalam prediksi terutama terjadi pada kelompok dropout (False Negative = 23), yang mungkin disebabkan oleh ketidakseimbangan dalam data. Meskipun demikian, model ini tetap efektif dalam mengidentifikasi pola mayoritas, tetapi memerlukan pendekatan tambahan untuk meningkatkan deteksi pada kelompok minoritas.

Hasil insight dari kedua model yaitu, regresi logistik serta LDA, menunjukkan performa prediksi yang kompetitif dan memuaskan pada dataset ini, dengan akurasi melebihi 94%. Regresi logistik lebih unggul dalam sensitivitas dan prediksi kelas dominan, sedangkan LDA sedikit lebih baik dalam hal spesifisitas. Dengan mempertimbangkan pelanggaran asumsi pada LDA sebelumnya, regresi logistik bisa jadi lebih konsisten dan stabil untuk situasi ini. Akan tetapi, selisih kinerja antara kedua model tidak signifikan, jadi pemilihan model dapat dipertimbangkan berdasarkan konteks penggunaan dan interpretasi yang diharapkan. Untuk memperbaiki kemampuan prediksi untuk kelas minoritas, mungkin diperlukan penelusuran metode lain seperti teknik sampling atau model klasifikasi yang lebih tangguh terhadap ketidakseimbangan kelas.

6. Daftar Pustaka

Sitasi disusun dan ditulis berdasarkan sistem nomor sesuai dengan urutan pengutipan, mengikuti format **Vancouver**. Hanya pustaka yang disitasi pada usulan penelitian yang dicantumkan dalam Daftar Pustaka. Pustaka yang disitasi maksimal 8 tahun terakhir sebanyak minimal 10 pustaka.

[1]. Lynn, N. D., & Emanuel, A. W. R. (2021, March). Using data mining techniques to predict students' performance. a review. In *IOP Conference series: materials science and engineering* (Vol. 1096, No. 1, p. 012083). IOP Publishing.

[2]. Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2.

[3]. Meka, F., & Nwaka, R. N. (2022). A Comparison of Logistic Regression and Discriminant Analysis in Predicting Student's Academic Outcome. *Article ID*.

[4]. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. *Frontiers in psychology*, *12*, 698490.

[5]. Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571.

[6]. Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, *1*(1), 143.

[7]. Hasan, M. N. (2019, December). A comparison of logistic regression and linear discriminant analysis in predicting of female students attrition from school in Bangladesh. In *2019 4th international conference on electrical information and communication technology (EICT)* (pp. 1-3). IEEE.

[8]. Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905-971.

[9]. Ho, I. M. K., Cheong, K. Y., & Weldon, A. (2021). Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques. *Plos one*, *16*(4), e0249423.

[10]. Mubarak, A. A., Cao, H., & Zhang, W. (2022). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, *30*(8), 1414-1433.

[11]. Link Rpubs https://rpubs.com/Arinatyas/1317027