LAPORAN UJIAN AKHIR SEMESTER
DATA MINING

Analisis Faktor-Faktor yang Mempengaruhi Kelulusan Mahasiswa Menggunakan Regresi Logistik

OLEH:

535220146 - Stefanus Anthony Harry

535220151 - Valeroy Putra Sientika

535220164 - Arethusa Rayhan Subrata

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS TARUMANAGARA
JUNI 2024

Daftar Isi

Daftar Tabel

Daftar Gambar

Bab I
Pendahuluan

1.1 Latar Belakang

Data mining adalah proses ekstraksi informasi berguna dari sejumlah besar data yang ada, dengan tujuan utama menemukan pola, hubungan, dan tren tersembunyi yang dapat digunakan dalam pengambilan keputusan [1]. Proses ini melibatkan penggunaan perangkat lunak yang didukung oleh perhitungan statistik, matematika, dan teknologi kecerdasan buatan atau *Artificial Intelligence* (*AI*). Dengan mengaplikasikan berbagai teknik dan algoritma seperti deteksi anomali, pengelompokan (*clustering*), aturan asosiasi, regresi, dan klasifikasi (*classification*), data mining mampu mengidentifikasi pola-pola yang signifikan, belum diketahui sebelumnya, dan memiliki implikasi yang penting.

Organisasi saat ini mengumpulkan data dalam jumlah besar dari berbagai sumber, termasuk transaksi bisnis, media sosial, sensor, dan lain-lain. Volume data yang sangat besar ini, jika tidak dianalisis dengan benar, hanya akan menjadi beban. Data mining memungkinkan pengolahan data mentah ini menjadi wawasan yang berharga, yang kemudian dapat digunakan untuk meningkatkan efisiensi operasional, memahami perilaku pelanggan, mendeteksi penipuan, dan berbagai aplikasi lainnya [2]. Misalnya, dalam bisnis, data mining dapat membantu mengidentifikasi tren pembelian pelanggan, yang pada gilirannya dapat digunakan untuk merancang strategi pemasaran yang lebih efektif.

Teknik-teknik data mining seperti pengelompokan, klasifikasi, regresi, dan deteksi anomali masing-masing memiliki peran khusus dalam analisis data. Pengelompokan (*clustering*) digunakan untuk mengelompokkan data berdasarkan kesamaan, sedangkan klasifikasi digunakan untuk mengklasifikasikan data ke dalam kategori yang telah ditentukan sebelumnya [3]. Regresi digunakan untuk memprediksi nilai atau tren masa depan berdasarkan data historis, sementara deteksi anomali digunakan untuk mengidentifikasi data yang tidak sesuai dengan pola umum atau data yang menyimpang [4]. Dengan teknik-teknik ini, data mining membantu mengubah data mentah menjadi informasi yang dapat diandalkan untuk mendukung pengambilan keputusan yang lebih baik dan lebih tepat waktu.

1.2 Uraian Teori

Regresi logistik adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen biner (dikotomi) dan satu atau lebih variabel independen [5]. Berbeda dengan regresi linear yang digunakan untuk memodelkan hubungan antara variabel dependen kontinu dan satu atau lebih variabel independen, regresi logistik memfokuskan pada probabilitas kejadian suatu peristiwa [6]. Fungsi logit, yang merupakan logaritma dari odds (peluang), memungkinkan kita untuk mengubah probabilitas kejadian menjadi bentuk linear yang dapat dimodelkan[6]. Regresi logistik memiliki beberapa asumsi penting, termasuk linearitas dalam logit, independensi observasi, tidak adanya multikolinearitas, dan ukuran sampel yang memadai untuk estimasi yang stabil. Parameter dalam regresi logistik diestimasi menggunakan metode Maximum Likelihood Estimation (MLE), sebagai berikut

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$
$$= ln\ (e^{\beta_0+\Sigma_j^p} = 1^{\beta_{jxj}})$$
$$= \beta_0 + \Sigma_j^p = 1^{\beta_{jxj}}) \quad (1.1)$$

Dalam mencocokkan model logistik, penting untuk memilih model yang memiliki fungsi penghubung dan variabel penjelas yang paling sesuai. Untuk menentukan kecocokan model yang optimal, uji statistik Goodness of Fit digunakan untuk membandingkan berbagai model. Salah satu uji yang sering digunakan untuk tujuan ini adalah uji Hosmer dan Lemeshow [6].

$$\hat{C} = \Sigma_k^g = 1 \frac{(O_k - n'_k\overline{\Pi}_k)^2}{n'_k\overline{\Pi}_k(1-\overline{\Pi}_k)} [6] \quad (1.2)$$

Analisis residual, kurva ROC (Receiver Operating Characteristic), dan Area Under the Curve (AUC) adalah alat diagnostic yang penting dalam menilai dan memvalidasi kinerja model regresi logistik [7]. ROC curve berfungsi sebagai gambaran kinerja model klasifikasi biner. Grafik ini dibuat dengan plot True Positive Rate (TPR) atau Sensitivity di sumbu y, dan False Positive Rate (FPR) atau (1 - Specificity) di sumbu x pada berbagai ambang batas (thershold) yang berbeda [7].

1. True Positive Rate (TPR) atau Sensitivity, yaitu Persentase data positif yang benar-benar diklasifikasikan sebagai positif.

$$TPR = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} [8] \quad (1.3)$$

2. False Positive Rate (FPR), yaitu Persentase data negatif yang salah diklasifikasikan sebagai positif.

$$FPR = \frac{False\ Positive\ (FP)}{False\ Positive\ (FP) + False\ Negative\ (FN)} [8] \quad (1.4)$$

AUC adalah luas di bawah kurva *ROC*. *AUC* memberikan satu nilai yang merangkum kinerja model dengan memanfaatkan area dibawaha kurfa *ROC [9]*. Nilai *AUC* berkisar antara 0 dan 1, semakin nilai tinggi maka performa model juga menjadi lebih baik[9].

1. $AUC = 0.5$: Model tidak lebih baik dari tebak-tebakan acak [10].
2. $AUC < 0.5$: Model lebih buruk daripada tebak-tebakan acak (menunjukkan model tersebut biasanya terbalik prediksinya) [10].
3. $AUC > 1$: Model lebih baik dari tebak-tebakan acak. AUC = 1.0 menunjukkan model sempurna[10].

Regresi logistik digunakan di berbagai jenis industri, seperti manufaktur untuk memperkirakan kemungkinan kegagalan mesin, pemasaran untuk memprediksi klik iklan online, dan keuangan untuk menganalisis transaksi terkait penipuan serta menilai risiko pengajuan pinjaman dan asuransi [11]. Implementasi metode ini dilakukan menggunakan *software Matlab*, sehingga mengolah dan menganalisis data dapat dilakukan dengan lebih mudah.

Distribusi binomial adalah distribusi probabilitas diskrit yang muncul ketika melakukan serangkaian uji coba independen, di mana setiap uji coba memiliki dua hasil yang mungkin (biasanya disebut sebagai sukses dan gagal), dan probabilitas sukses tetap sama untuk setiap uji coba[12]. Distribusi ini digunakan dalam konteks regresi logistik karena variabel dependen dalam regresi logistik adalah biner, yaitu memiliki dua kemungkinan hasil (biasanya 1 untuk sukses dan 0 untuk gagal). Dengan distribusi binomial, kita dapat menghitung probabilitas sukses dalam jumlah uji coba tertentu[12].

Distribusi binomial menyatakan probabilitas untuk mendapatkan k kejadian sukses dalam n uji coba, dengan probabilitas sukses p pada setiap uji coba. Fungsi massa probabilitas (PMF) untuk distribusi binomial dinyatakan sebagai berikut:

$$P(X = k) = (\frac{n}{k}) \cdot p^{k} \cdot (1 - p)^{n-k} \quad (1.5)$$

di mana :

- $P(X = k)$ adalah probabilitas untuk mendapatkan kkk kejadian sukses,
- n adalah jumlah total uji coba,
- p adalah probabilitas sukses pada setiap uji coba, dan
- $(\frac{n}{k})$ adalah simbol kombinasi yang menunjukkan jumlah cara untuk memilih elemen k dari total n elemen [12].

Generalized Linear Model (GLM) adalah model regresi yang terdiri dari komponen acak (sering disebut sebagai error) dan fungsi dari faktor desain (x) serta beberapa parameter (β). Dikembangkan oleh Nelder dan Wedderburn pada tahun 1972, Model linear ini berfungsi untuk menangani situasi di mana respons tidak berdistribusi Normal tetapi tetap bebas secara statistic [13]. Dalam kasus prediksi kelulusan mahasiswa, GLM dapat digunakan untuk memahami hubungan antara variabel-variabel independen (seperti jumlah jam belajar, kehadiran, dan nilai tugas) dengan probabilitas kelulusan mahasiswa.
Dalam teori regresi normal klasik, model regresi linier berganda dapat dituliskan sebagai:

$$Y = \beta_0 + \beta_i X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon \quad (1.6)$$

di mana:

- Y adalah variabel dependen,
- $X_1, X_2, ..., X_k$ adalah variabel independen,
- $\beta_0, \beta_2, ..., \beta_k$ adalah koefisien regresi,
- ϵ berdistribusi normal dengan rata-rata 0 dan varian konstan.

Rata-rata dari variabel respon y adalah

$$E(y) = \mu = \beta_0 + \beta_i X_1 + \beta_2 X_2 + ... + \beta_k X_k = x\beta \quad (1.7)$$

Ada tiga hal yang terkandung dalam GLM yaitu :

1. Memiliki komponen tetap yang disebut prediktor linear yang dinotasikan dengan η_i, yang merupakan bentuk dari linear tergeneralisasi, yaitu $\eta_i = x_i\beta$ dimana x_i merupakan vektor regresi untuk unit sebanyak i dengan *fixed effect* β.
2. $Respon\ y_i$ yang dalam keluarga eksponensial berdistribusi secara independen.
3. fungsi g(.) yang disebut fungsi link menunjukkan hubungan antara mean dengan prediktor linear, sedemikian sehingga $g(\mu_i) = \eta_i$ [13].

Outlier merupakan sebuah hasil observasi data yang secara signifikan memiliki nilai sangat jauh dari garis kuadrat terkecil atau dari data point lainnya. Mendeteksi outlier sangat penting karena memiliki dampak yang signifikan terhadap hasil analisis data, outlier negatif dapat mengurangi tingkat akurasi dari hasil prediksi, sedangkan outlier positif dapat dijadikan sebuah penemuan baru [14]. Berikut beberapa alasan lain mengapa mendeteksi outlier sangat penting :

1. Outlier dapat menyebabkan distorsi dalam analisis data dan model prediksi, mengarah pada kesimpulan yang tidak akurat [15].
2. Mendeteksi dan mengatasi outlier penting untuk memastikan pemahaman yang tepat tentang pola dan tren dalam dataset [15].
3. Analisis yang valid terjamin dengan mengelola outlier, menjaga konsistensi dan validitas hasil analisis [15].
4. Outlier juga dapat menjadi indikator adanya kesalahan atau kecurangan dalam data, memicu investigasi lebih lanjut untuk menyelesaikan masalah yang mungkin muncul [15].

Dengan demikian, mendeteksi outlier tidak hanya penting untuk meningkatkan kualitas analisis dan mengurangi risiko, tetapi juga untuk memastikan kepercayaan dalam pengambilan keputusan dan mendapatkan pemahaman yang lebih mendalam tentang data.

Principal Component Analysis (PCA) merupakan suatu teknik analisis yang bertujuan untuk mengurangi informasi dari sebuah dataset besar tanpa mengurangi informasi yang signifikan dan tetap mempertahankan struktur informasi dari data tersebut[16]. PCA membantu dalam mengidentifikasi pola yang signifikan yang ada dalam data untuk menunjukkan persamaan dan perbedaan mereka [17]. PCA memiliki beberapa tahapan yang harus dilalui sebagai berikut:

1. Standarisasi data dengan rata-rata dan varians

   Standarisasi data dengan mengurangkan setiap nilai dalam fitur dengan rata-rata (mean) kemudian dibagi dengan Varians [18].

$$x_{new} = \frac{x-\bar{x}}{\sigma} \tag{1.8}$$

2. Menghitung *covariance* matrix

   Beberapa rumus yang digunakan dalam menghitung *covariance* matrix adalah sebagai berikut:

Tabel 1.1 Rumus Covariance Matrix

	Populasi	Sample
Variance	$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Covariance	$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Kemudian diimplementasikan dengan aturan pada dibawah ini[18].

$$covmat = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$$

Gambar 1.1 Aturan PCA

3. Menghitung *Eigen-values* dan *Eigen-vectors*

   Vektor eigen adalah vektor bukan nol yang berubah paling banyak sebesar faktor skalar ketika diterapkannya transformasi linear kepadanya. Eigen-values yang sesuai berfungsi sebagai faktor yang digunakan untuk menskalakan vektor eigen. Jika kita memiliki sebuah matriks persegi *A* (dalam kasus kita, matriks kovarians), sebuah vektor v, dan sebuah skalar λ yang memenuhi persamaan $Av = \lambda v$, maka λ disebut sebagai nilai eigen yang terkait dengan vektor eigen v dari matriks *A*. Berdasarkan persamaan diatas maka didapat:

$$Av - \lambda v = 0;\ (A - \lambda I)v = 0 \tag{1.9}$$

   Karena v adalah vektor bukan nol maka untuk mencari nilai Eigen-values berlaku ketentuan:

$$det(A - \lambda I) = 0 \tag{1.10}$$

4. Penentuan Prinsipal Komponen (PK)

   Penentuan vektor eigen yang terkait dengan nilai eigen (lambda) di mana proporsi kumulatifnya tidak melebihi nilai maksimum proporsi yang telah ditetapkan atau diinginkan (PK) [18].

5. Standar Dataset Terbaru

   Setelah PCA dilakukan pada dataset pelatihan untuk mengurangi dimensi dan mengekstraksi fitur-fitur utama, langkah selanjutnya adalah menerapkan transformasi yang sama pada dataset baru [18].

Bab II
Pembahasan Masalah

2.1 Penjelasan Dataset

Penelitian ini menekankan pada analisis data pendidikan dengan fokus utama pada tingginya tingkat putus kuliah di kalangan mahasiswa. Dataset yang digunakan dalam penelitian ini diambil dari website Kaggle, berjudul "Predict students' dropout and academic success" dan mencakup berbagai variabel yang relevan untuk menganalisis risiko putus kuliah di kalangan mahasiswa. Dataset ini berisi informasi penting seperti latar belakang keluarga, kualifikasi orang tua, tingkat pengangguran di wilayah sekitar, serta prestasi akademik mahasiswa.

Dataset 'Predict Students.xlsx' yang digunakan dalam penelitian ini berisi informasi mengenai karakteristik mahasiswa dan hasil akademik mereka, mencakup 35 kolom termasuk variabel target (Target) yang terdiri dari tiga kategori: Dropout, Enrolled, dan Graduate. Dari analisis deskriptif, mayoritas mahasiswa memiliki status perkawinan belum menikah dan menghadiri kelas di siang hari. Kualifikasi pendidikan sebelumnya serta kualifikasi pendidikan ibu dan ayah menunjukkan variasi yang luas. Tingkat pengangguran di wilayah tempat tinggal mahasiswa bervariasi dari 3.7% hingga 16.5%, dengan tingkat inflasi berkisar antara -0.8% hingga 1.4% dan GDP bervariasi dari -3.12 hingga 1.74. Statistik deskriptif ini memberikan gambaran awal tentang distribusi dan variasi dalam data, membantu dalam memahami konteks analisis regresi logistik yang dilakukan untuk mengidentifikasi faktor-faktor yang mempengaruhi kelulusan mahasiswa.

Variabel-variabel independen dalam dataset ini meliputi kualifikasi orang tua yang merepresentasikan pendidikan tertinggi yang dicapai oleh orang tua mahasiswa, tingkat pengangguran yang menunjukkan persentase pengangguran di wilayah tempat tinggal mahasiswa, dan prestasi akademik yang diukur melalui nilai atau indikator kinerja akademik mahasiswa selama periode studi. Variabel dependen adalah status mahasiswa yang bernilai biner, yaitu "Graduate" atau "Dropout."

Dalam penelitian ini, regresi logistik digunakan untuk memodelkan probabilitas kejadian biner tersebut berdasarkan variabel-variabel independen yang ada. Pendekatan ini memungkinkan evaluasi sistematis terhadap pengaruh masing-masing faktor terhadap risiko putus kuliah dan prediksi mahasiswa yang membutuhkan intervensi dini. Kemampuan untuk mengidentifikasi mahasiswa ini secara akurat memungkinkan pendidik dan pembuat kebijakan untuk mengalokasikan sumber daya secara efektif, memberikan dukungan yang tepat kepada mahasiswa berisiko tinggi, dan pada akhirnya meningkatkan tingkat kelulusan.

Penerapan regresi logistik dalam konteks ini juga menemui sejumlah tantangan, termasuk pengumpulan data berkualitas tinggi, mengatasi potensi bias dalam data, dan memastikan interpretasi yang akurat dari hasil analisis. Oleh karena itu, penelitian ini tidak hanya memusatkan perhatian pada penggunaan regresi logistik sebagai teknik analisis tetapi juga pada pengembangan strategi untuk meningkatkan keakuratan dan relevansi prediksi yang dihasilkan. Dataset ini menjadi pondasi penting dalam penelitian karena menyediakan informasi yang diperlukan untuk membangun model prediksi yang handal dan memberikan wawasan mendalam mengenai faktor-faktor yang berkontribusi terhadap risiko putus kuliah di kalangan mahasiswa.

2.2 Penyelesaian Kasus

Dalam laporan ini, kasus yang diberikan diselesaikan menggunakan software MATLAB sebagai alat utama untuk melakukan analisis regresi linear pada dataset regresi logistik. MATLAB menawarkan berbagai fungsi statistik dan pemodelan yang kuat dan memungkinkan untuk mengelola dataset dengan efisien, membangun model analisis, serta menghasilkan hasil yang akurat untuk laporan ini.

Dataset dimuat dan diproses menggunakan kode pada Gambar 2.1 untuk analisis lebih lanjut. Tujuan utama dari kode ini adalah memuat dataset dari file 'Predict Students.xlsx' ke dalam bentuk tabel, dengan mempertahankan nama-nama variabel asli. Setelah memuat data, kode menampilkan nilai-nilai unik dari variabel target untuk memberikan gambaran awal tentang kategori yang ada dalam dataset. Selanjutnya, variabel target diubah menjadi tipe data 'categorical', yang penting untuk analisis statistik dan pemodelan di MATLAB. Akhirnya, kode ini memverifikasi bahwa konversi variabel target ke tipe data categorical telah dilakukan dengan benar dengan menampilkan kategori-kategori unik yang telah dihasilkan.

```
clear; clc;

% Memuat dataset dengan nama variabel yang dipertahankan
data = readtable('Predict Students.xlsx', 'VariableNamingRule', 'preserve');

% Menampilkan nilai unik dalam variabel Target
disp('Nilai unik dalam variabel Target:');
disp(unique(data.Target));

% Mengodekan variabel 'Target' (Dropout = 0, Graduate = 1)
data.Target = categorical(data.Target);

% Menampilkan kategori unik dalam variabel Target
disp('Kategori unik dalam variabel Target:');
disp(categories(data.Target));
```

Gambar 2.1 Mempersiapkan data awal untuk dianalisis lebih lanjut

Proses transformasi data pada Gambar 2.2 menggambarkan langkah-langkah untuk mengonversi variabel kategori menjadi nilai numerik dan dummy variables. Langkah pertama adalah mengubah kategori dalam variabel 'Target' menjadi nilai numerik, di mana 'Graduate' diberi nilai 1 dan 'Dropout' diberi nilai 0. Proses ini menggunakan fungsi categories untuk mengidentifikasi kategori dan membandingkan nilai dalam 'Target' dengan kategori 'Graduate'. Hasilnya adalah variabel 'Target' yang terdiri dari nilai 0 dan 1, yang siap untuk digunakan dalam analisis regresi.

Langkah berikutnya adalah memvalidasi bahwa variabel 'Target' hanya berisi nilai 0 dan 1 menggunakan pernyataan assert. Selanjutnya, kode mengidentifikasi kolom-kolom kategori dan numerik dalam dataset menggunakan fungsi varfun untuk memeriksa tipe data dari setiap kolom. Hasilnya adalah dua vektor logika, categoricalVars dan numericVars, yang menunjukkan posisi kolom-kolom kategori dan numerik dalam tabel data.

Setelah mengidentifikasi variabel-variabel kategori, langkah terakhir adalah mengonversi setiap variabel kategori menjadi dummy variables menggunakan fungsi varfun dan categorical. Dummy variables adalah representasi biner dari kategori yang dapat digunakan dalam analisis regresi, dan nama asli kolom dummy disimpan untuk referensi lebih lanjut.

```
% Mengubah kategori menjadi nilai numerik
targetCategories = categories(data.Target);
data.Target = double(data.Target == targetCategories{2}); % Diasumsikan 'Graduate' adalah 1 dan

% Memastikan 'Target' hanya berisi nilai 0 dan 1
assert(all(ismember(data.Target, [0, 1])), 'Variabel Target harus hanya berisi nilai 0 dan 1.');

% Mengidentifikasi kolom kategori dan numerik
categoricalVars = varfun(@iscategorical, data, 'OutputFormat', 'uniform');
numericVars = ~categoricalVars;

% Mengonversi variabel kategori menjadi variabel dummy
categoricalData = data(:, categoricalVars);
numericData = data(:, numericVars);

% Mengonversi data kategori menjadi dummy variables dan menyimpan nama asli kolom
dummyVars = varfun(@(x) double(categorical(x)), categoricalData);
dummyVarNames = dummyVars.Properties.VariableNames;
```

Gambar 2.2 Proses Transformasi Data Kategori Menjadi Nilai Numerik dan Dummy Variables

Kemudian bagian ini akan menjelaskan langkah-langkah dalam mempersiapkan data untuk analisis lebih lanjut, serta menambahkan bahwa langkah-langkah ini dapat dilihat dalam detail pada Gambar 2.3. Pertama, nama-nama kolom asli disimpan dalam variabel colNames sebelum data kategori dikonversi menjadi array. Selanjutnya, data dummy dan numerik digabungkan menjadi satu matriks prediktor X, dengan menghapus kolom terakhir yang berisi variabel target. Variabel target y disimpan secara terpisah. Untuk memastikan hasil analisis yang konsisten, seed acak ditetapkan menggunakan fungsi rng(1).

```
% Menyimpan nama-nama kolom asli sebelum mengonversi menjadi array
colNames = [dummyVarNames, numericData.Properties.VariableNames];

% Menggabungkan data dummy dan numerik
X = [table2array(dummyVars), table2array(numericData)];

% Menghapus kolom terakhir (Target) dari prediktor
X = X(:, 1:end-1);
y = data.Target; % Variabel respons

% Menetapkan seed acak untuk reproduktibilitas
rng(1);
```

Gambar 2.3 Penggabungan Data Dummy dan Numerik serta Persiapan Data untuk Analisis

Bagian ini menjelaskan beberapa langkah penting dalam persiapan data sebelum analisis. Pertama, data dipisahkan menjadi set pelatihan (70%) dan set pengujian (30%) menggunakan fungsi cvpartition. Pemisahan ini bertujuan untuk memastikan bahwa model yang dibangun dapat dievaluasi secara independen menggunakan data yang tidak terlihat selama pelatihan. Data pelatihan dan pengujian disimpan dalam variabel XTrain, yTrain, XTest, dan yTest.

Langkah kedua adalah mendeteksi dan menghapus outlier dari data pelatihan. Outlier dapat mempengaruhi kualitas model yang dibangun, sehingga penting untuk mengidentifikasinya. Fungsi arrayfun digunakan untuk menerapkan metode IQR (Interquartile Range) pada setiap kolom prediktor, yang didefinisikan dalam fungsi detectOutliers. Outlier yang terdeteksi kemudian dihapus dari XTrain dan yTrain untuk memastikan bahwa data yang digunakan dalam pelatihan model bersih dan representatif.

Langkah ketiga adalah memeriksa korelasi antar prediktor dalam data pelatihan untuk menghindari masalah multikolinieritas yang dapat menyebabkan overfitting. Matriks korelasi dihitung menggunakan fungsi corr dan hasilnya ditampilkan di konsol. Korelasi tinggi antara prediktor dapat menyebabkan model untuk memberikan bobot yang berlebihan pada variabel yang sangat berkorelasi, sehingga mengurangi generalisasi model. Untuk visualisasi yang lebih baik dan interpretasi yang lebih mudah, matriks korelasi divisualisasikan menggunakan heatmap dengan judul 'Matriks Korelasi' dan skema warna jet, dengan batas warna antara -1 dan 1.

```
% Memisahkan data menjadi set pelatihan 70% dan pengujian 30%
cv = cvpartition(size(X, 1), 'HoldOut', 0.3);
XTrain = X(training(cv), :);
yTrain = y(training(cv));
XTest = X(test(cv), :);
yTest = y(test(cv));

% Langkah 2: Mendeteksi dan menghapus outlier dari data pelatihan
outliers = arrayfun(@(col) detectOutliers(XTrain(:, col)), 1:size(XTrain, 2), 'UniformOutput', false);
outliers = any(cat(2, outliers{:}), 2);
XTrain = XTrain(~outliers, :);
yTrain = yTrain(~outliers);

% Langkah 3: Memeriksa korelasi untuk menghindari overfitting
correlationMatrix = corr(XTrain);
disp('Matriks Korelasi:');
disp(correlationMatrix);

% Visualisasi matriks korelasi
figure;
heatmap(correlationMatrix, 'Title', 'Matriks Korelasi', 'Colormap', jet, 'ColorLimits', [-1 1]);
```

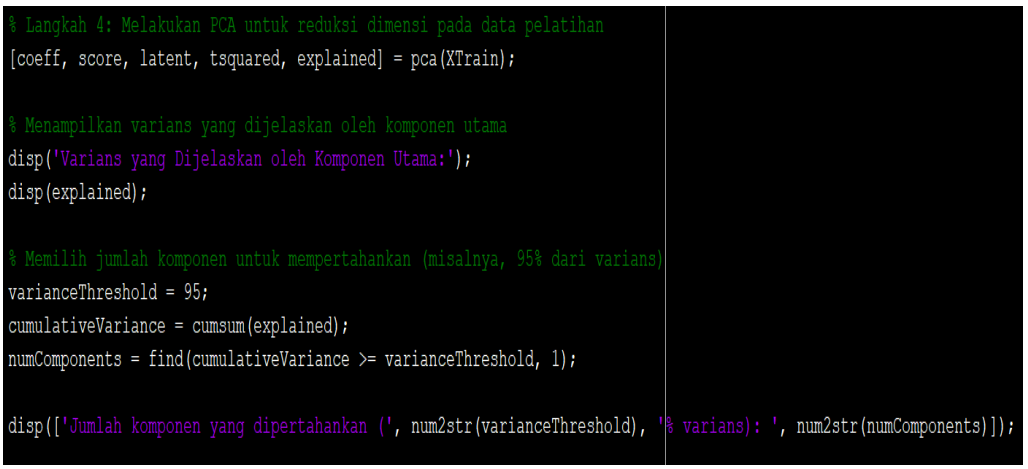
Gambar 2.4 Pemisahan Data, Penghapusan Outlier, dan Pemeriksaan Korelasi

Pada Gambar 2.5, Principal Component Analysis (PCA) diterapkan pada data pelatihan untuk melakukan reduksi dimensi. Reduksi dimensi bertujuan untuk mengurangi jumlah variabel prediktor tanpa mengorbankan informasi yang signifikan, yang dapat membantu dalam menghindari overfitting dan meningkatkan efisiensi model.

Pertama, PCA dijalankan pada data pelatihan XTrain menggunakan fungsi pca. Fungsi ini menghasilkan beberapa output penting: coeff (koefisien komponen utama), score (representasi data dalam ruang komponen utama), latent (varians dari masing-masing komponen utama), tsquared (statistik T^2 Hotelling), dan explained (persentase varians yang dijelaskan oleh masing-masing komponen utama).

Varians yang dijelaskan oleh setiap komponen utama kemudian ditampilkan di konsol untuk memberikan gambaran tentang seberapa banyak informasi yang ditangkap oleh masing-masing komponen. Ini membantu dalam memahami kontribusi setiap komponen utama terhadap total varians dalam data.

Selanjutnya, jumlah komponen utama yang perlu dipertahankan ditentukan berdasarkan ambang varians kumulatif yang diinginkan, dalam hal ini 95%. Varians kumulatif dihitung menggunakan fungsi cumsum, dan jumlah komponen utama yang diperlukan untuk mencapai atau melebihi ambang 95% diidentifikasi dengan fungsi find. Informasi ini ditampilkan di konsol untuk memberikan wawasan tentang jumlah komponen utama yang cukup untuk mempertahankan mayoritas informasi dari data asli.



```
% Langkah 4: Melakukan PCA untuk reduksi dimensi pada data pelatihan
[coeff, score, latent, tsquared, explained] = pca(XTrain);

% Menampilkan varians yang dijelaskan oleh komponen utama
disp('Varians yang Dijelaskan oleh Komponen Utama:');
disp(explained);

% Memilih jumlah komponen untuk mempertahankan (misalnya, 95% dari varians)
varianceThreshold = 95;
cumulativeVariance = cumsum(explained);
numComponents = find(cumulativeVariance >= varianceThreshold, 1);

disp(['Jumlah komponen yang dipertahankan (', num2str(varianceThreshold), '% varians): ', num2str(numComponents)]);
```

Gambar 2.5 Reduksi Dimensi Menggunakan PCA

Setelah menentukan jumlah komponen utama yang diperlukan untuk mempertahankan 95% varians data, dapat dilihat pada Gambar 2.6, langkah berikutnya adalah mengurangi data pelatihan ke dalam ruang komponen utama yang dipilih. Dalam langkah ini, data pelatihan (XTrain) diubah menjadi representasi baru dengan menggunakan jumlah komponen yang dipilih (numComponents). Data pelatihan yang telah direduksi disimpan dalam XTrainReduced. Kemudian, transformasi yang sama diterapkan pada data uji (XTest) dengan menggunakan koefisien komponen utama (coeff) dan nilai rata-rata dari data pelatihan. Data uji yang telah direduksi disimpan dalam XTestReduced. Proses ini memastikan bahwa data uji diubah dengan cara yang konsisten dengan data pelatihan.

Selanjutnya, model regresi logistik dibangun menggunakan data pelatihan yang telah direduksi (XTrainReduced) dan variabel respons (yTrain). Model ini dibangun dengan fungsi fitglm menggunakan distribusi binomial dan link logit, yang cocok untuk prediksi variabel biner seperti kelulusan siswa. Setelah model regresi logistik terbentuk, ringkasan model ditampilkan di konsol untuk memberikan gambaran tentang performa dan koefisien model. Ringkasan ini mencakup informasi penting tentang estimasi koefisien, kesalahan standar, nilai t, dan nilai p untuk setiap prediktor dalam model.

Prediksi terhadap data uji dilakukan dengan model yang telah dibentuk, menghasilkan probabilitas prediksi (yPred). Probabilitas ini kemudian dikonversi menjadi hasil biner (yPredBinary) berdasarkan ambang batas 0.5, di mana nilai di atas 0.5 dianggap sebagai kelulusan (1) dan di bawah 0.5 dianggap sebagai non-kelulusan (0). Proses ini memungkinkan evaluasi kinerja model dalam memprediksi kelulusan siswa.

Terakhir, tipe data dari yTest dan yPredBinary disamakan untuk memastikan konsistensi dalam evaluasi model. Proses ini memastikan bahwa prediksi dan nilai aktual dapat dibandingkan secara akurat dalam langkah evaluasi model berikutnya, yang penting untuk menilai keandalan dan akurasi model regresi logistik yang telah dibangun.

```
% Mengurangi data pelatihan ke jumlah komponen yang dipilih
XTrainReduced = score(:, 1:numComponents);

% Menerapkan transformasi yang sama pada data uji
XTestReduced = (XTest - mean(XTrain)) * coeff(:, 1:numComponents);

% Model regresi logistik
mdl = fitglm(XTrainReduced, yTrain, 'Distribution', 'binomial', 'Link', 'logit');

% Menampilkan ringkasan model regresi logistik
disp(mdl);

% Memprediksi respons untuk set pengujian
yPred = predict(mdl, XTestReduced);

% Mengonversi probabilitas menjadi hasil biner
yPredBinary = yPred >= 0.5;

% Memastikan kedua yTest dan yPredBinary memiliki tipe yang sama
yTest = double(yTest);
yPredBinary = double(yPredBinary);
```

Gambar 2.6 Penerapan PCA pada Data dan Pembangunan Model Regresi Logistik

Langkah selanjutnya pada Gambar 2.7 adalah mengevaluasi kinerja model regresi logistik yang telah dibangun. Pertama, akurasi model dihitung dengan membandingkan prediksi biner (yPredBinary) dengan nilai aktual (yTest). Akurasi ini dihitung sebagai rata-rata dari prediksi yang benar, dan hasilnya ditampilkan dalam bentuk persentase. Akurasi ini memberikan gambaran umum tentang seberapa baik model mampu memprediksi kelulusan siswa.

Selain akurasi, Area Under the Curve (AUC) juga dihitung untuk mengevaluasi kinerja model secara lebih mendalam. AUC adalah ukuran kinerja model klasifikasi biner yang mempertimbangkan sensitivitas (true positive rate) dan spesifisitas (false positive rate). Kurva ROC (Receiver Operating Characteristic) diplot untuk menggambarkan trade-off antara sensitivitas dan spesifisitas pada berbagai ambang batas. Nilai AUC yang lebih tinggi menunjukkan kinerja model yang lebih baik dalam membedakan antara kelas-kelas yang berbeda.

Selanjutnya, confusion matrix diplot untuk memberikan gambaran visual tentang performa model dalam hal prediksi yang benar dan salah. Confusion matrix ini dinormalisasi per baris dan kolom untuk memudahkan interpretasi. Confusion matrix menunjukkan jumlah true positives, true negatives, false positives, dan false negatives yang dihasilkan oleh model. Informasi ini sangat berguna untuk memahami kesalahan prediksi dan kinerja keseluruhan model.

Terakhir, faktor-faktor signifikan yang mempengaruhi kelulusan siswa diidentifikasi berdasarkan nilai p dari koefisien model. Tabel koefisien (coeffTable) yang berisi estimasi, kesalahan standar, nilai t, dan nilai p dari setiap prediktor digunakan untuk menentukan faktor-faktor yang signifikan. Faktor-faktor yang memiliki nilai p kurang dari 0.05 dianggap signifikan dan disimpan dalam variabel significantFactors. Identifikasi faktor-faktor signifikan ini penting untuk memahami variabel mana yang memiliki dampak terbesar pada probabilitas kelulusan siswa, dan informasi ini dapat digunakan untuk pengambilan keputusan berbasis data dalam konteks pendidikan.

```
% Menghitung akurasi model
accuracy = mean(yPredBinary == yTest);
fprintf('Akurasi: %.2f%%\n', accuracy * 100);

% Menghitung Area Under the Curve (AUC)
[X_ROC, Y_ROC, ~, AUC] = perfcurve(yTest, yPred, 1);
fprintf('AUC: %.2f\n', AUC);

% Memplot confusion matrix
figure;
cm = confusionchart(yTest, yPredBinary);
cm.Title = 'Confusion Matrix untuk Regresi Logistik';
cm.RowSummary = 'row-normalized'; % Normalisasi per baris
cm.ColumnSummary = 'column-normalized'; % Normalisasi per kolom

% Memplot kurva ROC
figure;
plot(X_ROC, Y_ROC)
xlabel('False positive rate')
ylabel('True positive rate')
title(['Kurva ROC (AUC = ' num2str(AUC) ')'])

% Menentukan faktor-faktor signifikan berdasarkan nilai p
coeffTable = mdl.Coefficients;
significantFactors = coeffTable(coeffTable.pValue < 0.05, :);
```

Gambar 2.7 Evaluasi Kinerja Model dan Identifikasi Faktor-Faktor Signifikan

Sebagai langkah terakhir dalam analisis ini, Gambar 2.8 menampilkan faktor-faktor signifikan yang mempengaruhi kelulusan siswa berdasarkan hasil model regresi logistik.

Pertama, nama-nama faktor signifikan ditampilkan. Nama-nama faktor ini diambil dari daftar kolom asli yang sebelumnya telah dikonversi menjadi variabel dummy dan numerik. Faktor-faktor yang memiliki nilai p kurang dari 0.05 dianggap signifikan, dan nama-nama faktor tersebut disimpan dalam variabel significantFactorNames. Menampilkan nama-nama faktor ini penting untuk memahami variabel spesifik yang berpengaruh terhadap hasil kelulusan siswa.

Selanjutnya, tabel faktor-faktor signifikan beserta nilai p masing-masing ditampilkan. Tabel ini memberikan detail lengkap mengenai estimasi koefisien, kesalahan standar, nilai t, dan nilai p untuk setiap faktor yang signifikan. Informasi ini sangat berguna untuk mengevaluasi sejauh mana setiap faktor mempengaruhi probabilitas kelulusan siswa. Tabel ini juga membantu dalam mengidentifikasi variabel mana yang paling penting dan dapat digunakan untuk mengarahkan intervensi dan kebijakan pendidikan.

Terakhir, fungsi untuk mendeteksi outlier menggunakan metode Interquartile Range (IQR) juga disertakan. Fungsi ini mendeteksi nilai-nilai yang berada di luar batas bawah dan atas (lowerBound dan upperBound) yang dihitung berdasarkan Q1 dan Q3. Nilai-nilai yang berada di luar rentang ini dianggap sebagai outlier. Deteksi dan penghapusan outlier penting untuk memastikan bahwa analisis data tidak dipengaruhi oleh nilai-nilai ekstrim yang dapat mendistorsi hasil.

```
% Menampilkan nama faktor-faktor signifikan
disp('Faktor-faktor yang signifikan mempengaruhi kelulusan:');
significantFactorNames = colNames(coeffTable.pValue(2:end) < 0.05);
disp(significantFactorNames);

% Menampilkan faktor signifikan beserta nilai p
disp('Tabel faktor-faktor signifikan:');
disp(significantFactors);

% Mendefinisikan fungsi untuk mendeteksi outlier menggunakan metode IQR
function isOutlier = detectOutliers(data)
    Q1 = quantile(data, 0.25);
    Q3 = quantile(data, 0.75);
    IQR = Q3 - Q1;
    lowerBound = Q1 - 1.5 * IQR;
    upperBound = Q3 + 1.5 * IQR;
    isOutlier = (data < lowerBound) | (data > upperBound);
end
```

Gambar 2.8 Identifikasi dan Penampilan Faktor-Faktor Signifikan

Bab III
Hasil dan Analisis

Bab ini menguraikan hasil dari implementasi kode yang telah dijelaskan pada bab sebelumnya serta analisis mendalam terhadap data yang diperoleh. Berdasarkan pengujian dan eksperimen yang dilakukan, berbagai temuan penting terkait faktor-faktor penentu dan performa algoritma dalam analisis data ini akan dipaparkan. Analisis ini bertujuan untuk mengidentifikasi kekuatan dan kelemahan dari metode yang digunakan, mengevaluasi akurasi, presisi, dan recall dari model, serta memberikan wawasan mengenai faktor-faktor yang paling berpengaruh dalam menentukan hasil akhir dari data yang dianalisis.

Variabel Target memiliki tiga nilai unik: 'Dropout', 'Enrolled', dan 'Graduate'. 'Dropout' menunjukkan mahasiswa yang keluar sebelum menyelesaikan program, 'Enrolled' menunjukkan mahasiswa yang masih aktif dalam program, dan 'Graduate' menunjukkan mahasiswa yang telah menyelesaikan program dengan sukses. Analisis terhadap distribusi kategori ini penting untuk memahami pola keberhasilan dan kegagalan dalam pendidikan, serta untuk mengidentifikasi faktor-faktor yang mempengaruhi keberlanjutan dan keberhasilan akademis mahasiswa. Dapat dilihat dalam Gambar 3.1 bahwa distribusi kategori unik dalam variabel Target memperlihatkan status akhir mahasiswa setelah mengikuti program pendidikan.

```
Nilai unik dalam variabel Target:
    {'Dropout' }
    {'Enrolled'}
    {'Graduate'}

Kategori unik dalam variabel Target:
    {'Dropout' }
    {'Enrolled'}
    {'Graduate'}
```

Gambar 3.1 Distribusi Kategori Unik dalam Variabel Target

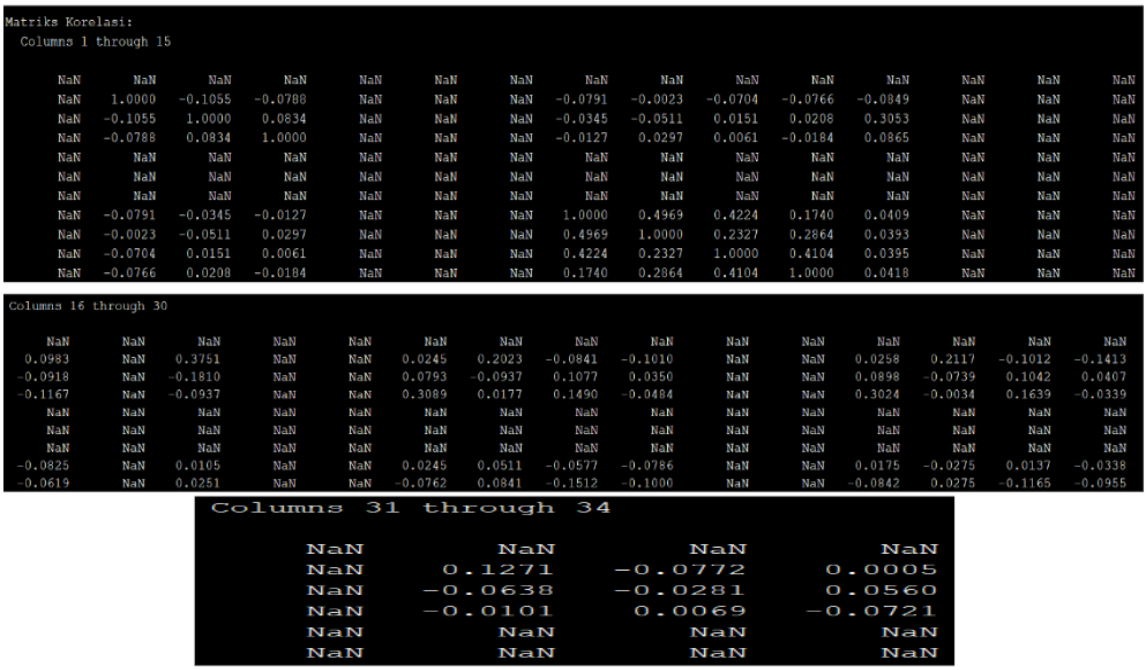
Matriks korelasi pada Gambar 3.2 menggambarkan hubungan antar variabel dalam dataset yang digunakan. Beberapa poin penting dari matriks korelasi ini adalah sebagai berikut:

1. Korelasi Antar Variabel
   Beberapa variabel menunjukkan korelasi yang signifikan dengan variabel lainnya. Misalnya, terdapat korelasi positif sebesar 0.4969 antara variabel pada kolom ke-8 dan ke-9, yang menunjukkan hubungan yang cukup kuat antara kedua variabel tersebut. Korelasi negatif juga ditemukan, seperti korelasi sebesar -0.1810 antara variabel pada kolom ke-2 dan ke-18, yang menunjukkan bahwa ketika satu variabel meningkat, variabel lainnya cenderung menurun.
2. Korelasi dengan Variabel Target

Dalam konteks ini, variabel target adalah yang digunakan untuk menentukan output model, seperti 'Dropout', 'Enrolled', atau 'Graduate'. Sebagai contoh, variabel pada kolom ke-2 memiliki korelasi positif sebesar 0.3751 dengan variabel target pada kolom ke-18, menunjukkan bahwa ada hubungan yang cukup kuat antara kedua variabel tersebut.

3. Korelasi Kuat dan Lemah
   Korelasi yang sangat kuat atau sangat lemah dapat memberikan wawasan penting. Sebagai contoh, korelasi sebesar 0.9519 antara variabel pada kolom ke-29 dan kolom ke-19 menunjukkan hubungan yang hampir linear sempurna. Disisi lain, beberapa variabel menunjukkan korelasi yang sangat lemah, seperti korelasi sebesar 0.0005 antara variabel pada kolom ke-2 dan kolom ke-31, menunjukkan bahwa hubungan antara kedua variabel tersebut sangat lemah atau tidak ada.
4. Variabel dengan Korelasi Tinggi
   Variabel yang menunjukkan korelasi tinggi dapat menjadi indikasi multikolinearitas, yang dapat mempengaruhi hasil analisis regresi dan model prediktif. Misalnya, kolom ke-19 dan ke-27 memiliki korelasi sebesar 0.7741, yang cukup tinggi dan mungkin perlu diperhatikan dalam analisis lebih lanjut.

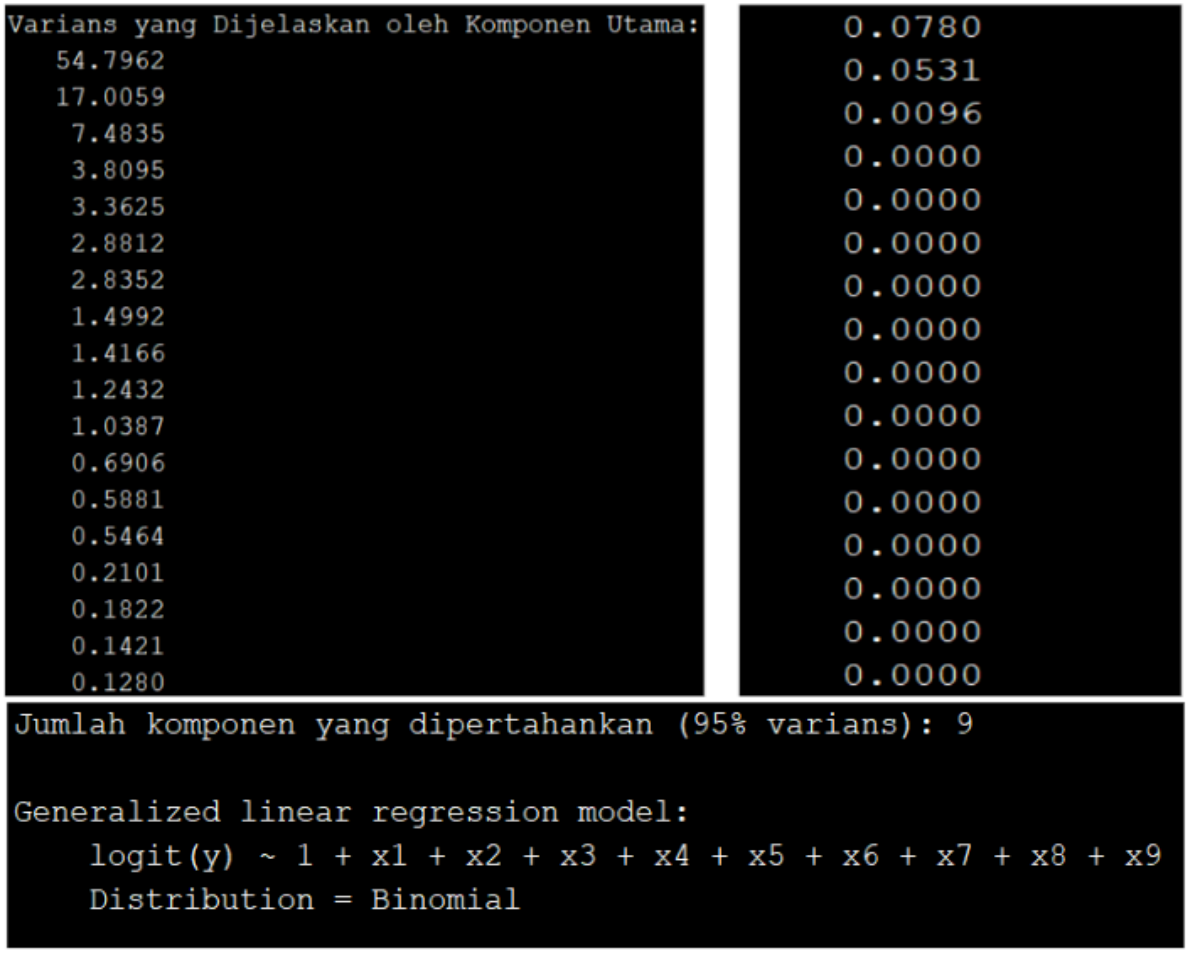
Matriks korelasi ini membantu dalam mengidentifikasi hubungan antar variabel dan dapat digunakan untuk menentukan variabel mana yang memiliki dampak signifikan terhadap variabel target. Informasi ini sangat penting dalam proses pemodelan dan analisis data lebih lanjut.



Gambar 3.2 Output Koefisien Regresi Logistik

Pada Gambar 3.3, analisis komponen utama (PCA) menunjukkan bahwa komponen pertama menjelaskan sebagian besar varians dalam data, yakni 54.7962%. Komponen kedua dan ketiga juga memberikan kontribusi signifikan terhadap varians yang dijelaskan, masing-masing sebesar 17.0059% dan 7.4835%. Secara keseluruhan, sembilan komponen

utama pertama menjelaskan sebagian besar varians dalam data, dengan total kontribusi lebih dari 95%. Setelah komponen kesembilan, kontribusi varians yang dijelaskan oleh masing-masing komponen menurun drastis, menunjukkan bahwa komponen-komponen ini tidak signifikan dalam menjelaskan variasi dalam data. Dengan demikian, dapat disimpulkan bahwa sebagian besar informasi dalam dataset dapat direpresentasikan oleh sembilan komponen utama pertama, yang berguna untuk tujuan reduksi dimensi dan analisis lebih lanjut.



Gambar 3.3 Analisis Komponen Utama (PCA) dan Model Regresi Linear Terumumkan

Tabel 3.1 menunjukkan hasil estimasi koefisien dari model regresi linear terumumkan (GLM) dengan fungsi logit, yang digunakan untuk memprediksi variabel respons biner (y) berdasarkan sembilan komponen utama (x1) hingga (x9). Model ini menggunakan distribusi binomial. Berikut adalah interpretasi dari hasil estimasi:

1. Intercept memiliki estimasi koefisien -1.4973 dengan standar error (SE) 0.10967, t-statistik -13.653, dan p-value 1.9322e-42. Nilai p-value yang sangat kecil menunjukkan bahwa intercept ini sangat signifikan.
2. x1 memiliki estimasi koefisien 0.0012493 dengan SE 0.0081359, t-statistik 0.15355, dan p-value 0.87796. P-value yang besar menunjukkan bahwa x1 tidak signifikan dalam model.

3. x2 memiliki estimasi koefisien -0.037446 dengan SE 0.014674, t-statistik -2.5519, dan p-value 0.010714. P-value yang kecil menunjukkan bahwa x2 signifikan pada tingkat signifikansi 5%.
4. x3 memiliki estimasi koefisien 0.051953 dengan SE 0.020867, t-statistik 2.4897, dan p-value 0.012785. Ini menunjukkan bahwa x3 signifikan pada tingkat signifikansi 5%.
5. x4 memiliki estimasi koefisien -0.030464 dengan SE 0.031033, t-statistik -0.98167, dan p-value 0.32626. P-value yang besar menunjukkan bahwa x4 tidak signifikan.
6. x5 memiliki estimasi koefisien -0.040556 dengan SE 0.031742, t-statistik -1.2777, dan p-value 0.20135. P-value yang besar menunjukkan bahwa x5 tidak signifikan.
7. x6 memiliki estimasi koefisien -0.077418 dengan SE 0.034544, t-statistik -2.2411, dan p-value 0.025017. P-value yang kecil menunjukkan bahwa x6 signifikan pada tingkat signifikansi 5%.
8. x7 memiliki estimasi koefisien 0.26095 dengan SE 0.035985, t-statistik 7.2515, dan p-value 4.1211e-13. Nilai p-value yang sangat kecil menunjukkan bahwa x7 sangat signifikan.
9. x8 memiliki estimasi koefisien -0.10062 dengan SE 0.047443, t-statistik -2.1209, dan p-value 0.033931. P-value yang kecil menunjukkan bahwa x8 signifikan pada tingkat signifikansi 5%.
10. x9 memiliki estimasi koefisien 0.026529 dengan SE 0.050137, t-statistik 0.52914, dan p-value 0.59671. P-value yang besar menunjukkan bahwa x9 tidak signifikan.

Secara keseluruhan, dari sembilan komponen utama yang digunakan dalam model, x2, x3, x6, x7, dan x8 menunjukkan signifikansi statistik dalam memprediksi variabel respons biner (y).

Tabel 3.1 Koefisien yang Diestimasi dari Model Regresi Linear Terumumkan

	Estimate	SE	tStat	pValue
(Intercept)	-1.4973	0.10967	-13.653	1.9322e-42
x1	0.0012493	0.0081359	0.15355	0.87796
x2	-0.037446	0.014674	-2.5519	0.010714
x3	0.051953	0.020867	2.4897	0.012785
x4	-0.030464	0.031033	-0.98167	0.32626
x5	-0.040556	0.031742	-1.2777	0.20135
x6	-0.077418	0.034544	-2.2411	0.025017
x7	0.26095	0.035985	7.2515	4.1211e-13
x8	-0.10062	0.047443	-2.1209	0.033931
x9	0.026529	0.050137	0.52914	0.59671

Gambar 3.4 menunjukkan hasil dari model regresi logistik yang digunakan untuk memprediksi kelulusan. Model ini melibatkan 671 observasi dengan 661 derajat kebebasan error, menunjukkan bahwa model mengestimasi 10 parameter. Dispersi model adalah 1, menunjukkan bahwa model memenuhi asumsi variansi konstan. Chi^2-statistik sebesar 84 dengan p-value 2.58e-14 mengindikasikan bahwa model ini secara signifikan lebih baik dibandingkan model konstan. Akurasi model adalah 61.57%, menunjukkan bahwa prediksi model ini benar sekitar 61.57% dari waktu. AUC (Area Under the Curve) sebesar 0.60 menunjukkan kemampuan moderat model dalam membedakan antara kelas positif dan negatif. Faktor-faktor yang secara signifikan mempengaruhi kelulusan adalah mode aplikasi, urutan aplikasi, kualifikasi sebelumnya, kebangsaan, dan kualifikasi ibu. Kesimpulannya, model ini memberikan wawasan tentang variabel-variabel penting yang mempengaruhi kelulusan, meskipun dengan tingkat akurasi dan AUC yang moderat, masih ada ruang untuk perbaikan dalam prediksi.

```
671 observations, 661 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 84, p-value = 2.58e-14
Akurasi: 61.57%
AUC: 0.60
Faktor-faktor yang signifikan mempengaruhi kelulusan:
   {'Application mode'}    {'Application order'}    {'Previous qualification'}    {'Nacionality'}    {'Mother's qualification'}
```

Gambar 3.4 Hasil Model Regresi Logistik untuk Prediksi Kelulusan

Pada Tabel 3.2, beberapa faktor signifikan yang mempengaruhi kelulusan mahasiswa telah diidentifikasi. Faktor-faktor tersebut beserta koefisien estimasi, error standar (SE), statistik t (tStat), dan nilai p (pValue) dijelaskan sebagai berikut:

1. Intercept memiliki koefisien estimasi sebesar -1.4973 dengan error standar 0.10967. Statistik t untuk intercept adalah -13.653 dengan nilai p sebesar 1.9322e-42, menunjukkan bahwa intercept sangat signifikan secara statistik.
2. x2 (Application order) memiliki koefisien estimasi -0.037446 dengan error standar 0.014674. Statistik t sebesar -2.5519 dan nilai p sebesar 0.010714 menunjukkan bahwa faktor ini signifikan dalam mempengaruhi kelulusan.
3. x3 (Previous qualification) memiliki koefisien estimasi 0.051953 dengan error standar 0.020867. Statistik t sebesar 2.4897 dan nilai p sebesar 0.012785 menunjukkan bahwa faktor ini signifikan.
4. x6 memiliki koefisien estimasi -0.077418 dengan error standar 0.034544. Statistik t sebesar -2.2411 dan nilai p sebesar 0.025017 menunjukkan bahwa faktor ini signifikan.
5. x7 (Nationality) memiliki koefisien estimasi 0.26095 dengan error standar 0.035985. Statistik t sebesar 7.2515 dan nilai p sebesar 4.1211e-13 menunjukkan bahwa faktor ini sangat signifikan dalam mempengaruhi kelulusan.

6. x8 (Mother's qualification) memiliki koefisien estimasi -0.10062 dengan error standar 0.047443. Statistik t sebesar -2.1209 dan nilai p sebesar 0.033931 menunjukkan bahwa faktor ini juga signifikan.

Secara keseluruhan, faktor-faktor ini menunjukkan pengaruh yang signifikan terhadap probabilitas kelulusan mahasiswa, dengan beberapa faktor menunjukkan hubungan positif dan yang lainnya negatif.

Tabel 3.2 Faktor-faktor Signifikan yang Mempengaruhi Kelulusan

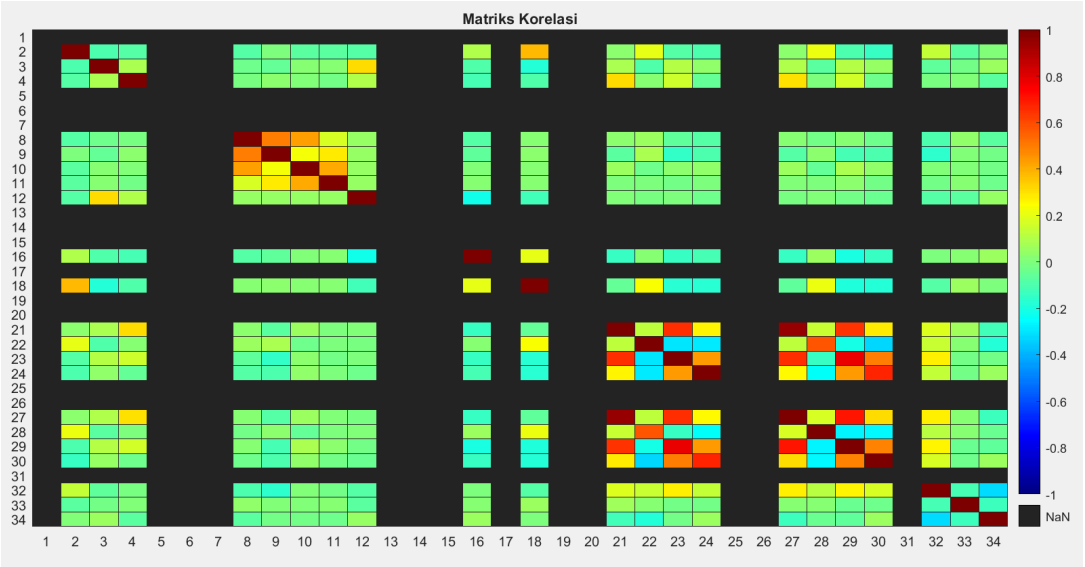
	Estimate	SE	tStat	pValue
(Intercept)	-1.4973	0.10967	-13.653	1.9322e-42
x2	-0.037446	0.014674	-2.5519	0.010714
x3	0.051953	0.020867	2.4897	0.012785
x6	-0.077418	0.034544	-2.2411	0.025017
x7	0.26095	0.035985	7.2515	4.1211e-13
x8	-0.10062	0.047443	-2.1209	0.033931

Pada Gambar 3.5, ditampilkan sebuah matriks korelasi yang menggambarkan hubungan antara 34 variabel. Matriks ini menggunakan skala warna untuk menunjukkan kekuatan dan arah korelasi antara pasangan variabel. Berikut adalah beberapa keterangan dari matriks tersebut:

1. Skala Warna
   Warna pada matriks berkisar dari merah tua (korelasi positif kuat, mendekati 1) hingga biru tua (korelasi negatif kuat, mendekati -1). Warna hijau menunjukkan korelasi rendah atau mendekati nol, sedangkan warna kuning dan oranye menunjukkan korelasi sedang. Warna hitam menunjukkan nilai 'NaN', menandakan bahwa korelasi tidak dapat dihitung untuk pasangan variabel tersebut.
2. Hubungan Positif Kuat
   Terdapat beberapa blok merah di matriks ini, terutama di sekitar variabel 2-5 dan 9-12, menunjukkan bahwa variabel-variabel ini memiliki korelasi positif yang kuat satu sama lain.
3. Hubungan Negatif Kuat
   Ada juga beberapa blok biru tua di sekitar variabel 27-34, menunjukkan korelasi negatif yang kuat antara variabel-variabel ini.
4. Korelasi Rendah atau Tidak Signifikan
   Banyak sel pada matriks berwarna hijau atau kuning, menunjukkan bahwa mayoritas variabel tidak memiliki korelasi yang signifikan satu sama lain atau memiliki korelasi rendah.
5. Kekosongan Data (NaN)

Warna hitam pada beberapa bagian matriks menunjukkan bahwa data tidak tersedia atau korelasi tidak dapat dihitung untuk pasangan variabel tersebut.

Secara keseluruhan, Gambar 3.5 memberikan visualisasi yang jelas tentang bagaimana setiap variabel berkorelasi satu sama lain, yang sangat berguna dalam analisis data untuk mengidentifikasi variabel mana yang memiliki hubungan kuat dan dapat digunakan dalam analisis lebih lanjut atau model prediktif.



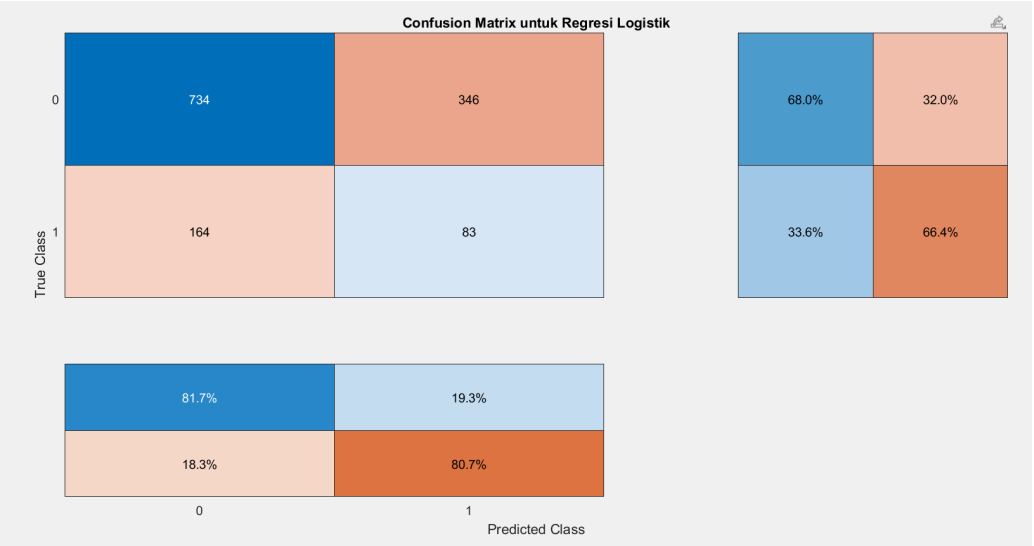
Gambar 3.5 Matriks Korelasi

Pada Gambar 3.6, ditampilkan sebuah matriks kebingungan (confusion matrix) untuk model regresi logistik. Matriks ini menyediakan informasi mengenai kinerja klasifikasi dari model tersebut. Berikut adalah rincian penjelasan dari matriks kebingungan ini:

1. Kelas Asli dan Kelas Prediksi:
    a. Sumbu vertikal (True Class) menunjukkan kelas asli atau sesungguhnya.
    b. Sumbu horizontal (Predicted Class) menunjukkan kelas yang diprediksi oleh model.
2. Jumlah Prediksi Benar dan Salah:
    a. True Negatives (TN): 734, menunjukkan jumlah prediksi negatif yang benar (model memprediksi kelas 0, dan sebenarnya adalah kelas 0).
    b. False Positives (FP): 346, menunjukkan jumlah prediksi positif yang salah (model memprediksi kelas 1, tapi sebenarnya adalah kelas 0).
    c. False Negatives (FN): 164, menunjukkan jumlah prediksi negatif yang salah (model memprediksi kelas 0, tapi sebenarnya adalah kelas 1).
    d. True Positives (TP): 83, menunjukkan jumlah prediksi positif yang benar (model memprediksi kelas 1, dan sebenarnya adalah kelas 1).
3. Persentase Prediksi:
    a. Persentase Prediksi Negatif yang Benar (TN): 81.7%.
    b. Persentase Prediksi Positif yang Salah (FP): 19.3%.
    c. Persentase Prediksi Negatif yang Salah (FN): 18.3%.
    d. Persentase Prediksi Positif yang Benar (TP): 80.7%.

4. Akurasi dan Error Rate:
    a. Akurasi untuk Kelas 0: 68.0%, menunjukkan persentase prediksi benar untuk kelas 0.
    b. Error Rate untuk Kelas 0: 32.0%, menunjukkan persentase prediksi salah untuk kelas 0.
    c. Akurasi untuk Kelas 1: 66.4%, menunjukkan persentase prediksi benar untuk kelas 1.
    d. Error Rate untuk Kelas 1: 33.6%, menunjukkan persentase prediksi salah untuk kelas 1.

Gambar 3.6 memberikan gambaran yang jelas tentang bagaimana model regresi logistik bekerja dalam mengklasifikasikan data. Dengan informasi ini, dapat dievaluasi kinerja model, termasuk tingkat akurasi dan area yang perlu perbaikan.

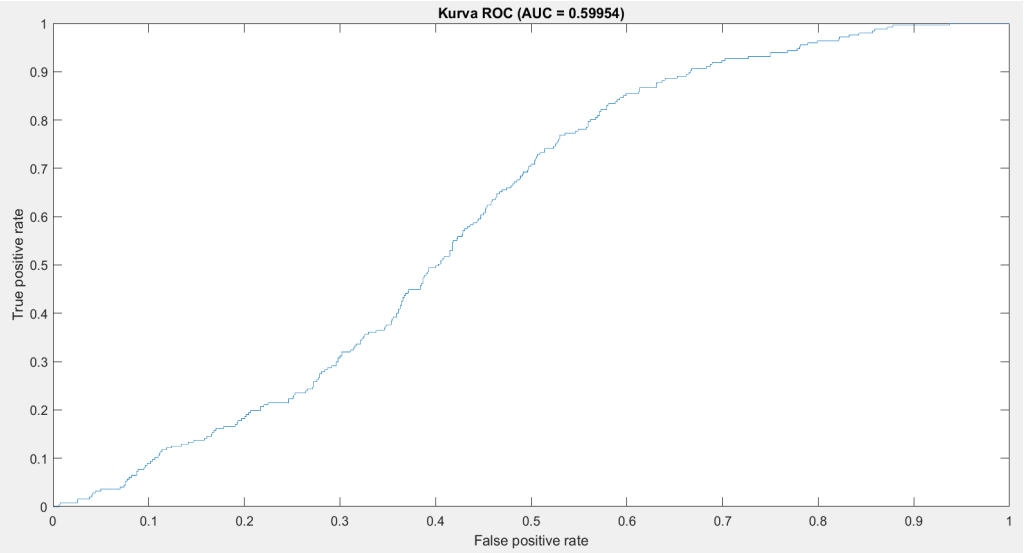


Gambar 3.6 Confusion Matrix untuk Regresi Logistik

Pada Gambar 3.7 ditampilkan Kurva ROC (Receiver Operating Characteristic) yang digunakan untuk mengevaluasi kinerja model regresi logistik. Kurva ROC menunjukkan hubungan antara tingkat positif palsu (false positive rate) pada sumbu horizontal dan tingkat positif sejati (true positive rate) pada sumbu vertikal.

AUC (Area Under the Curve) dari kurva adalah 0.59954, yang menunjukkan seberapa baik model dapat membedakan antara kelas positif dan negatif. AUC bernilai 0.5 menunjukkan bahwa model tidak lebih baik daripada tebak-tebakan acak, sedangkan AUC bernilai 1 menunjukkan model yang sempurna. Dalam hal ini, AUC sebesar 0.59954 menunjukkan bahwa model memiliki performa yang mendekati tebak-tebakan acak namun masih memiliki sedikit kemampuan dalam membedakan kelas.

Model dengan AUC sebesar 0.59954 ini menunjukkan bahwa model regresi logistik yang digunakan memiliki kinerja yang kurang memuaskan dalam memprediksi kelas yang benar dibandingkan dengan prediksi acak. Kurva ROC ini memberikan gambaran visual tentang seberapa baik model tersebut bekerja.



Gambar 3.7 Kurva ROC untuk Model Regresi Logistik

Bab IV
Kesimpulan

Berdasarkan hasil analisis yang dilakukan pada data ini, beberapa temuan penting dapat disimpulkan. Pertama, distribusi kategori unik dalam variabel target ('Dropout', 'Enrolled', dan 'Graduate') memberikan gambaran yang jelas tentang status akhir mahasiswa dalam program pendidikan. Analisis matriks korelasi menyoroti hubungan antar variabel dalam dataset, dengan beberapa korelasi signifikan antara variabel yang bisa mempengaruhi model prediksi. Model regresi logistik menunjukkan variabel seperti urutan aplikasi, kualifikasi sebelumnya, kebangsaan, dan kualifikasi ibu memiliki pengaruh signifikan terhadap prediksi kelulusan mahasiswa. Meskipun model ini memberikan wawasan yang berharga, evaluasi kinerja menunjukkan bahwa terdapat ruang untuk perbaikan, terutama dalam meningkatkan akurasi dan AUC model.

Berdasarkan hasil penelitian ini, disarankan untuk mempertimbangkan penggunaan teknik lain dalam pemodelan yang mungkin lebih cocok untuk karakteristik dataset ini, seperti penggunaan model ensemble atau pengoptimalan lebih lanjut terhadap fitur-fitur yang digunakan. Selain itu, memperluas dataset dengan variabel tambahan atau mempertimbangkan pengelompokan ulang variabel yang mungkin memiliki korelasi yang kuat dapat meningkatkan kemampuan prediktif model. Peningkatan dalam pengumpulan data, termasuk aspek-aspek kualitatif yang tidak terwakili dalam analisis ini, juga dapat memberikan pemahaman lebih mendalam terhadap faktor-faktor yang mempengaruhi keberhasilan akademik mahasiswa. Dengan langkah-langkah ini, diharapkan bahwa analisis dan prediksi dalam konteks pendidikan dapat lebih efektif dan informatif bagi pengambil keputusan di bidang ini.

Bab V
Lampiran

Dataset Predict Students.xlsx :
https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

Source Code :

```
clear; clc;

% Langkah 1: Memuat data dari file Excel dengan nama variabel yang dipertahankan
data = readtable('Predict Students.xlsx', 'VariableNamingRule', 'preserve');

% Menampilkan nilai unik dalam variabel 'Target'
disp('Nilai unik dalam variabel Target:');
disp(unique(data.Target));

% Mengodekan variabel 'Target' (Dropout = 0, Graduate = 1)
data.Target = categorical(data.Target);

% Menampilkan kategori unik dalam variabel 'Target'
disp('Kategori unik dalam variabel Target:');
disp(categories(data.Target));

% Mengubah kategori menjadi nilai numerik
targetCategories = categories(data.Target);
data.Target = double(data.Target == targetCategories{2}); % Diasumsikan 'Graduate' adalah
1 dan 'Dropout' adalah 0

% Memastikan 'Target' hanya berisi nilai 0 dan 1
assert(all(ismember(data.Target, [0, 1])), 'Variabel Target harus hanya berisi nilai 0 dan 1.');

% Mengidentifikasi kolom kategori dan numerik
categoricalVars = varfun(@iscategorical, data, 'OutputFormat', 'uniform');
numericVars = ~categoricalVars;

% Mengonversi variabel kategori menjadi variabel dummy
categoricalData = data(:, categoricalVars);
numericData = data(:, numericVars);

% Mengonversi data kategori menjadi dummy variables dan menyimpan nama asli kolom
dummyVars = varfun(@(x) double(categorical(x)), categoricalData);
dummyVarNames = dummyVars.Properties.VariableNames;
```

```
% Menyimpan nama-nama kolom asli sebelum mengonversi menjadi array
colNames = [dummyVarNames, numericData.Properties.VariableNames];

% Menggabungkan data dummy dan numerik
X = [table2array(dummyVars), table2array(numericData)];

% Menghapus kolom terakhir (Target) dari prediktor
X = X(:, 1:end-1);
y = data.Target; % Variabel respons

% Menetapkan seed acak untuk reproduktibilitas
rng(1);

% Memisahkan data menjadi set pelatihan 70% dan pengujian 30%
cv = cvpartition(size(X, 1), 'HoldOut', 0.3);
XTrain = X(training(cv), :);
yTrain = y(training(cv));
XTest = X(test(cv), :);
yTest = y(test(cv));

% Langkah 2: Mendeteksi dan menghapus outlier dari data pelatihan
outliers = arrayfun(@(col) detectOutliers(XTrain(:, col)), 1:size(XTrain, 2), 'UniformOutput',
false);
outliers = any(cat(2, outliers{:}), 2);
XTrain = XTrain(~outliers, :);
yTrain = yTrain(~outliers);

% Langkah 3: Memeriksa korelasi untuk menghindari overfitting
correlationMatrix = corr(XTrain);
disp('Matriks Korelasi:');
disp(correlationMatrix);

% Visualisasi matriks korelasi
figure;
heatmap(correlationMatrix, 'Title', 'Matriks Korelasi', 'Colormap', jet, 'ColorLimits', [-1 1]);

% Langkah 4: Melakukan PCA untuk reduksi dimensi pada data pelatihan
[coeff, score, latent, tsquared, explained] = pca(XTrain);

% Menampilkan varians yang dijelaskan oleh komponen utama
disp('Varians yang Dijelaskan oleh Komponen Utama:');
disp(explained);
```

```
% Memilih jumlah komponen untuk mempertahankan (misalnya, 95% dari varians)
varianceThreshold = 95;
cumulativeVariance = cumsum(explained);
numComponents = find(cumulativeVariance >= varianceThreshold, 1);

disp(['Jumlah komponen yang dipertahankan (', num2str(varianceThreshold), '% varians): ',
num2str(numComponents)]);

% Mengurangi data pelatihan ke jumlah komponen yang dipilih
XTrainReduced = score(:, 1:numComponents);

% Menerapkan transformasi yang sama pada data uji
XTestReduced = (XTest - mean(XTrain)) * coeff(:, 1:numComponents);

% Model regresi logistik
mdl = fitglm(XTrainReduced, yTrain, 'Distribution', 'binomial', 'Link', 'logit');

% Menampilkan ringkasan model regresi logistik
disp(mdl);

% Memprediksi respons untuk set pengujian
yPred = predict(mdl, XTestReduced);

% Mengonversi probabilitas menjadi hasil biner
yPredBinary = yPred >= 0.5;

% Memastikan kedua yTest dan yPredBinary memiliki tipe yang sama
yTest = double(yTest);
yPredBinary = double(yPredBinary);

% Menghitung akurasi model
accuracy = mean(yPredBinary == yTest);
fprintf('Akurasi: %.2f%%\n', accuracy * 100);

% Menghitung Area Under the Curve (AUC)
[X_ROC, Y_ROC, ~, AUC] = perfcurve(yTest, yPred, 1);
fprintf('AUC: %.2f\n', AUC);

% Memplot confusion matrix
figure;
cm = confusionchart(yTest, yPredBinary);
cm.Title = 'Confusion Matrix untuk Regresi Logistik';
cm.RowSummary = 'row-normalized'; % Normalisasi per baris
cm.ColumnSummary = 'column-normalized'; % Normalisasi per kolom
```

```
% Memplot kurva ROC
figure;
plot(X_ROC, Y_ROC)
xlabel('False positive rate')
ylabel('True positive rate')
title(['Kurva ROC (AUC = ' num2str(AUC) ')'])

% Menentukan faktor-faktor signifikan berdasarkan nilai p
coeffTable = mdl.Coefficients;
significantFactors = coeffTable(coeffTable.pValue < 0.05, :);

% Menampilkan nama faktor-faktor signifikan
disp('Faktor-faktor yang signifikan mempengaruhi kelulusan:');
significantFactorNames = colNames(coeffTable.pValue(2:end) < 0.05);
disp(significantFactorNames);

% Menampilkan faktor signifikan beserta nilai p
disp('Tabel faktor-faktor signifikan:');
disp(significantFactors);

% Mendefinisikan fungsi untuk mendeteksi outlier menggunakan metode IQR
function isOutlier = detectOutliers(data)
    Q1 = quantile(data, 0.25);
    Q3 = quantile(data, 0.75);
    IQR = Q3 - Q1;
    lowerBound = Q1 - 1.5 * IQR;
    upperBound = Q3 + 1.5 * IQR;
    isOutlier = (data < lowerBound) | (data > upperBound);
end
```

Daftar Pustaka

[1] Z. Setiawan, M. Fajar, A. Mudi Priyatno, A. Y. P. Putri, and M. Aryuni, *BUKU AJAR DATA MINING*. PT. Sonpedia Publishing Indonesia, 2023.

[2] G. Urva, Desyanti, isa Albanna, M. Sobri Sungkar, I. M. Made Agus Oka Gunawan, and I. Adhicandra, *PENERAPAN DATA MINING DI BERBAGAI BIDANG : Konsep, Metode, dan Studi Kasus*. PT. Sonpedia Publishing Indonesia, 2023.

[3] F. Marisa, S. Risnanto, R. Hardi, B. Pudjoatmojo, H. Tri Esti Endah, and A. L. Maukar, *Algoritma Populer Dalam Intelligent Sistem Beserta Contoh Kasus* . Deepublish, 2023.

[4] Muttaqin, W. W. Widiyanto, M. M. G. F. Mandias, S. R. Pungus, and A. W. W. K. Hapsari, "Pengenalan Data Mining," pp. 151–152, 2023, Accessed: Jun. 21, 2024. [Online]. Available: https://repository.upy.ac.id/4946/1/FullBook%20Pengenalan%20Data%20Mining.pdf

[5] K. W. Patunduk, Avini, Sumarni, A. Pratiwi, Harbianti, and R. Hidayat, "PEMODELAN PASIEN COVID-19 DI KOTA PALOPO DENGAN REGRESI LOGISTIK (Studi Perbandingan Regresi Logistik dan Analisis Survival)," *Jurnal Penelitian Matematika dan Pendidikan Matematika*, vol. 5, pp. 1–2, Nov. 2022.

[6] Ramli, Desi Yuniarti, and Rito Goejantoro, "Perbandingan Metode Klasifikasi Regresi Logistik Dengan Jaringan Saraf Tiruan (Studi Kasus: Pemilihan Jurusan Bahasa dan IPS pada SMAN 2 Samarinda Tahun Ajaran 2011/2012)," *Jurnal EKSPONENSIAL*, vol. 4, pp. 17–18, May 2013.

[7] Jeffry, S. Usman, and Aziz Firman, "Analisis Perilaku Pelanggan menggunakan Metode Ensemble Logistic Regression," *Jurnal Penelitian Teknik Informatika Universitas Prima Indonesia (UNPRI) Medan*, vol. 6, pp. 93–94, Nov. 2023.

[8] Iguazio Team, "What is the True Positive Rate in Machine Learning?" Accessed: Jun. 27, 2024. [Online]. Available: https://www.iguazio.com/glossary/true-positive-rate/

[9] A. A. Hidayah, "KLASIFIKASI GEN ESENSIAL PADA DROSOPHILA MELANOGASTER BERDASARKAN DNA SEQUENCE MENGGUNAKAN METODE GATED RECURRENT UNIT (GRU)," Lampung, 2023.

[10] J. C. Olamendy, "Memilih Metrik yang Tepat: Penarikan, Presisi, Kurva PR, dan Penjelasan Kurva ROC," Medium.

[11] Inc. T. Amazon Web Services, "Apa itu regresi logistik?" Accessed: Jun. 20, 2024. [Online]. Available: https://aws.amazon.com/id/what-is/logistic-regression/

[12] Moh. Yamin Darsyah and Dwi Haryo Ismunarti, “PERBANDINGAN KURVA PADA DISTRIBUSI UNIFORM DAN DISTRIBUSI BINOMIAL,” *Statistika*, vol. 1, pp. 21–23, May 2013.

[13] UPI EDU, “Model Linear Tergeneralisasi.” Accessed: Jun. 20, 2024. [Online]. Available: http://repository.upi.edu/111317/7/s_d505_033696_chapter3.pdf

[14] L. Hiryanto and D. Herwindiati, Writers, Multivariate Outliers in Multivariate Regression Model. Universitas Tarumanagara, 2024.

[15] PT Revolusi Cita Edukasi, "Outlier," Revoupedia, p. 1, 2024.

[16] PT. Algoritma Data Indonesia, "Mengenal Principal Component Analysis," Algoritma learn data science by building, p. 1, 2022.

[17] L. Hiryanto and D. Herwindiati, Writers, Principal Components Analysis (PCA). Universitas Tarumanagara, 2024.

[18] M. Billah, "Langkah-langkah Dasar PCA (Principal Component Analysis) menggunakan Perhitungan Manual dan Library Scikit-Learn," LinkedIn, p. 1, 2021.