

**PREDIKSI PLACEMENT MAHASISWA  
MENGGUNAKAN  
REGULARIZED LOGISTIC REGRESSION**

**UAS DATA SCIENCE**

Oleh  
Adi Subhandi Subiyono  
227006516087



**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNOLOGI KOMUNIKASI DAN  
INFORMATIKA  
UNIVERSITAS NASIONAL**

**2025**

# DAFTAR ISI

<b>DAFTAR ISI .....</b>	<b>ii</b>
<b>DAFTAR GAMBAR .....</b>	<b>iv</b>
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1    Latar Belakang .....	1
1.2    Rumusan Masalah .....	2
1.3    Tujuan Penelitian.....	2
1.4    Output yang Diharapkan .....	3
<b>BAB II DATA DAN METODOLOGI.....</b>	<b>4</b>
2.1    Dataset dan Variabel Penelitian.....	4
2.2    Tahapan Pengolahan Data .....	4
2.3    Model yang Digunakan.....	4
2.4    Evaluasi Performa Model.....	5
2.5    Analisis dan Kontribusi Variabel .....	5
<b>BAB III HASIL DAN PEMBAHASAN .....</b>	<b>6</b>
3.1    Dataset dan Struktur Variabel.....	6
3.2    Kualitas Data dan Missing Values.....	7
3.3    Setup Fitur, Target, dan Split Data .....	8
3.4    Modeling dan Hyperparameter Tuning .....	9
3.5    Evaluasi Model.....	10
3.6    Interpretasi Kontribusi Variabel .....	12
3.7    Insight dan Rekomendasi .....	14
<b>BAB IV KESIMPULAN .....</b>	<b>15</b>

4.1	Kesimpulan .....	15
<b>LAMPIRAN</b>	.....	<b>16</b>

## DAFTAR GAMBAR

Gambar 3. 1 Tipe data .....	6
Gambar 3. 2 Kelas target.....	6
Gambar 3. 3 Missing values.....	7
Gambar 3. 4 Kategori gender.....	7
Gambar 3. 5 Split data.....	8
Gambar 3. 6 Train test.....	8
Gambar 3. 7 Pipeline numerik & kategorikal .....	8
Gambar 3. 8 Best parameter.....	9
Gambar 3. 9 Hasil cross-validation.....	9
Gambar 3. 10 Performa model .....	10
Gambar 3. 11 ROC Curve .....	10
Gambar 3. 12 Precision-Recall Curve.....	10
Gambar 3. 13 Confusin matrix.....	11
Gambar 3. 14 Confusion Matrix Heatmap.....	11
Gambar 3. 15 Koefisien Positif.....	12
Gambar 3. 16 Koefisien Negatif .....	12
Gambar 3. 17 Top kontribusi variabel.....	13
Gambar 3. 18 Variabel paling berkontribusi .....	13
Gambar 3. 19 Rekomendasi Variabel.....	14

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Tingkat keberhasilan lulusan dalam mendapatkan pekerjaan setelah lulus sering digunakan sebagai ukuran kualitas sebuah perguruan tinggi atau universitas. Indikator ini bukan hanya soal reputasi; ini juga menunjukkan apakah proses pembelajaran, kurikulum, dan layanan pengembangan karier benar-benar mempersiapkan mahasiswa untuk memasuki pasar kerja. Bagi mahasiswa, apakah mereka memiliki pekerjaan atau tidak setelah lulus secara langsung memengaruhi stabilitas keuangan dan arah karier mereka. Oleh karena itu, kampus memerlukan cara yang lebih objektif untuk menilai risiko dan peluang sebelum kelulusan.

Masalahnya, banyak intervensi karier dilakukan secara umum dan reaktif, artinya pelatihan diberikan secara merata tanpa memperhatikan mahasiswa yang paling membutuhkan. Padahal kampus biasanya punya data yang bisa jadi tanda awal, seperti nilai di berbagai tingkatan, hasil selama kuliah, jurusan yang dipilih, pengalaman kerja sebelum lulus, serta hasil tes kemampuan bekerja. Jika sinyal ini dianalisis dengan cara yang teratur, kampus bisa membuat sistem peringatan dini untuk mengenali mahasiswa yang mungkin kesulitan mencari pekerjaan dan merancang bantuan yang lebih tepat, seperti program untuk meningkatkan kemampuan kerja atau memperkuat pengalaman berkerja.

Dataset pada penelitian ini memuat variabel akademik dan employability serta label status penempatan kerja (placed atau not placed), sehingga secara alami membentuk persoalan klasifikasi. Tujuan penelitian bukan hanya menghasilkan prediksi, tetapi juga memahami variabel mana yang paling berkaitan dengan peluang placement agar rekomendasi kebijakan dapat berbasis bukti. Oleh karena itu, digunakan *Regularized Logistic Regression* untuk menyeimbangkan kebutuhan performa prediksi dan interpretabilitas, sekaligus mengurangi risiko overfitting pada data berukuran relatif kecil. Hasil yang diperoleh diperlakukan sebagai

hubungan statistik, bukan bukti sebab-akibat, namun tetap berguna untuk pengambilan keputusan yang lebih terukur.

## **1.2 Rumusan Masalah**

1. Bagaimana membangun model klasifikasi untuk memprediksi status penempatan kerja mahasiswa (placed atau not placed) berdasarkan variabel akademik dan employability yang tersedia pada dataset?
2. Variabel apa saja yang paling berkontribusi terhadap peluang mahasiswa menjadi placed, dan bagaimana arah pengaruhnya berdasarkan interpretasi model (koefisien/odds ratio)?
3. Seberapa baik performa model klasifikasi yang dibangun, ditinjau dari metrik evaluasi yang relevan (misalnya ROC-AUC, PR-AUC, F1-score, confusion matrix) dengan skema validasi yang memadai (stratified cross-validation)?

## **1.3 Tujuan Penelitian**

1. Membangun model klasifikasi untuk memprediksi status penempatan kerja mahasiswa (placed atau not placed) berdasarkan variabel akademik dan employability yang tersedia pada dataset.
2. Mengidentifikasi faktor-faktor yang paling berkontribusi terhadap peluang mahasiswa menjadi placed melalui interpretasi model (koefisien dan odds ratio), sehingga terlihat variabel mana yang paling relevan secara praktis.
3. Mengevaluasi performa model secara kuantitatif dan transparan menggunakan metrik yang sesuai untuk klasifikasi biner (misalnya ROC-AUC, PR-AUC, F1-score, dan confusion matrix) dengan skema validasi yang memadai (stratified cross-validation).
4. Menyusun rekomendasi berbasis hasil analisis yang dapat digunakan sebagai masukan untuk perbaikan strategi pembinaan akademik maupun pengembangan kesiapan kerja mahasiswa, tanpa mengklaim hubungan sebab-akibat.

## **1.4 Output yang Diharapkan**

1. Model klasifikasi final (Regularized Logistic Regression) yang sudah terintegrasi dengan pipeline preprocessing (imputasi, encoding kategorikal, scaling numerik) sehingga dapat direplikasi dan diuji ulang.
2. Ringkasan performa model dalam bentuk tabel/angka metrik utama (ROC-AUC, PR-AUC, F1-score, accuracy) serta confusion matrix untuk menunjukkan jenis kesalahan prediksi yang dominan.
3. Peringkat kontribusi variabel berupa tabel koefisien dan odds ratio (serta arah pengaruh), termasuk interpretasi singkat untuk variabel yang paling berpengaruh.
4. Insight dan rekomendasi praktis berdasarkan variabel kunci, misalnya area yang perlu diperkuat (akademik tertentu, employability, pengalaman kerja), serta konsekuensi jika kampus mengabaikan sinyal risiko yang terdeteksi.

## **BAB II**

### **DATA DAN METODOLOGI**

#### **2.1 Dataset dan Variabel Penelitian**

Dataset yang digunakan berisi profil mahasiswa dan hasil status penempatan kerja (*placed* atau *not placed*). Variabel yang tersedia mencakup informasi akademik (nilai rata-rata pada beberapa jenjang pendidikan, performa sarjana dan pascasarjana), serta variabel yang berkaitan dengan kesiapan kerja seperti pengalaman kerja sebelum lulus dan nilai tes kemampuan kerja. Target analisis adalah status penempatan kerja yang dibentuk sebagai klasifikasi biner (Placed = 1, Not Placed = 0). Variabel Gaji tidak digunakan dalam pemodelan klasifikasi karena hanya tersedia pada mahasiswa yang berstatus *placed*, sehingga berpotensi menimbulkan data leakage dan membuat evaluasi model menjadi tidak valid.

#### **2.2 Tahapan Pengolahan Data**

Tahapan pengolahan dilakukan untuk memastikan data siap digunakan dalam pemodelan dan evaluasi. Proses yang dilakukan meliputi:

1. Pembersihan data: menghapus kolom identitas (ID) dan menyeragamkan label kategorikal yang tidak konsisten (misalnya variasi penulisan gender).
2. Penanganan missing values: nilai kosong pada fitur numerik diisi menggunakan median, sedangkan fitur kategorikal diisi menggunakan nilai yang paling sering muncul atau kategori “Unknown”.
3. Transformasi fitur:
  - Variabel kategorikal diubah menggunakan one-hot encoding.
  - Variabel numerik distandardisasi menggunakan StandardScaler agar perbedaan skala tidak memengaruhi model.

#### **2.3 Model yang Digunakan**

Model yang digunakan adalah Regularized Logistic Regression, yaitu model klasifikasi yang memprediksi probabilitas suatu observasi termasuk dalam kelas placed berdasarkan kombinasi variabel input. Regularisasi digunakan untuk

mengurangi risiko overfitting, terutama setelah proses one-hot encoding yang dapat menambah jumlah fitur. Parameter regularisasi ( $C$ ) ditentukan melalui proses tuning sederhana menggunakan cross-validation.

## 2.4 Evaluasi Performa Model

Performa model dievaluasi menggunakan beberapa metrik untuk memberikan gambaran yang lebih lengkap:

1. ROC-AUC untuk mengukur kemampuan model membedakan dua kelas.
2. PR-AUC (Average Precision) untuk menilai kualitas prediksi pada kondisi proporsi kelas tidak seimbang.
3. F1-score sebagai ukuran keseimbangan precision dan recall.
4. Confusion Matrix untuk melihat jenis kesalahan prediksi (false positive dan false negative) secara langsung.

## 2.5 Analisis dan Kontribusi Variabel

Kontribusi variabel dianalisis menggunakan koefisien model Logistic Regression. Koefisien diubah ke bentuk odds ratio untuk memudahkan interpretasi, sehingga dapat dijelaskan apakah suatu variabel meningkatkan atau menurunkan peluang mahasiswa berstatus placed. Untuk variabel kategorikal yang diubah menjadi one-hot, interpretasi dilakukan relatif terhadap kategori acuan (baseline).

## BAB III

### HASIL DAN PEMBAHASAN

#### 3.1 Dataset dan Struktur Variabel

Dataset terdiri dari 215 observasi dengan target status kelulusan (Bekerja/Belum). Distribusi kelas menunjukkan Placed 148 (68,84%) dan Not Placed 67 (31,16%), artinya ada ketidakseimbangan kelas ringan yang masih wajar untuk klasifikasi biner. Informasi ini menjadi dasar pemilihan skema validasi dan pengaturan pembobotan kelas pada model.

Tabel 2. Ringkasan Tipe Data & Jumlah Nilai Unik

Kolom	Tipe Data	Jumlah Unik
ID	int64	215
Jenis Kelamin	object	4
Nilai rata-rata SMP	float64	103
Lembaga pendidikan kelas 10	object	3
Nilai rata-rata SMA	float64	97
Lembaga pendidikan kelas 12	object	3
Jurusan saat SMA	object	3
IPK	float64	89
Program studi sarjana	object	3
Pengalaman kerja sebelum lulus	object	2
Nilai tes kemampuan kerja	float64	97
Pendidikan pascasarjana	object	2
Nilai rata-rata pascasarjana	float64	205
status kelulusan (Bekerja/Belum)	object	2
Gaji	float64	45

Gambar 3. 1 Tipe data

Di sisi struktur data, variabel numerik dan kategorikal muncul bersamaan. Tipe data dan jumlah nilai unik memberi sinyal bahwa sebagian fitur kategorikal memiliki kategori terbatas (mis. lembaga pendidikan, jurusan, program studi), sehingga cocok ditangani dengan one-hot encoding.

Tabel 4. Distribusi Kelas Target

Kelas	Jumlah	Persentase
Placed	148	68.84%
Not Placed	67	31.16%

Gambar 3. 2 Kelas target

### 3.2 Kualitas Data dan Missing Values

Pemeriksaan missing values menunjukkan kekosongan terbesar terdapat pada Gaji (31,16%). Nilai ini kosong terutama pada kelas Not Placed, sehingga Gaji berpotensi menjadi leakage jika dipakai sebagai fitur. Karena itu, Gaji dikeluarkan dari fitur untuk menjaga integritas prediksi. Missing values lain relatif kecil, misalnya pada nilai tes kemampuan kerja (2,79%) serta beberapa nilai akademik di bawah 1%, sehingga masih aman ditangani dengan imputasi.

**Tabel 3. Kolom dengan Missing Values**

Kolom	Missing Count	Missing (%)
Gaji	67	31.16%
Nilai tes kemampuan kerja	6	2.79%
Nilai rata-rata SMP	2	0.93%
Jenis Kelamin	1	0.47%
Nilai rata-rata SMA	1	0.47%

Gambar 3. 3 Missing values

Selain itu, kategori pada variabel Jenis Kelamin dinormalisasi (misalnya “Male/Female” menjadi “M/F”), agar encoding tidak memecah kategori yang sebenarnya sama.

**Tabel 5. Perbandingan Kategori Gender (Sebelum vs Sesudah Normalisasi)**

Kategori (sebelum)	Jumlah	Kategori (sesudah)	Jumlah
M	132	M	139.000000
F	73	F	75.000000
Male	7	nan	1.000000
Female	2	nan	nan
nan	1	nan	nan

Gambar 3. 4 Kategori gender

### 3.3 Setup Fitur, Target, dan Split Data

**Tabel 6. Ringkasan Data untuk Modeling**

Item	Nilai
Jumlah observasi	215
Jumlah fitur (X)	12
Kolom target	status kelulusan (Bekerja/Belum)
Placed (1)	148
Not Placed (0)	67

Gambar 3. 5 Split data

Setelah pembersihan, jumlah fitur yang digunakan adalah 12 fitur dan target tetap dua kelas (Placed vs Not Placed). Data kemudian dibagi menggunakan train-test split stratified dengan proporsi 172 data train dan 43 data test, dengan komposisi kelas yang tetap terjaga (proporsi Placed/Not Placed di train dan test tidak berubah jauh).

**Tabel 8. Train-Test Split (Stratified)**

Subset	Jumlah Data	Placed (1)	Not Placed (0)	Placed (%)	Not Placed (%)
Train	172	118	54	68.60%	31.40%
Test	43	30	13	69.77%	30.23%

Gambar 3. 6 Train test

Pipeline preprocessing disusun agar proses modeling dapat direplikasi:

- Numerik: imputasi median → standardisasi
- Kategorikal: imputasi modus → one-hot encoding

**Tabel 9. Ringkasan Preprocessing**

Komponen	Langkah	Jumlah Kolom
Numerik	Imputer (median) → StandardScaler	5
Kategorikal	Imputer (most_frequent) → OneHotEncoder (handle_unknown='ignore')	7

Gambar 3. 7 Pipeline numerik & kategorikal

### 3.4 Modeling dan Hyperparameter Tuning

Model yang digunakan adalah Regularized Logistic Regression (L2) dengan tuning parameter C menggunakan Stratified K-Fold Cross Validation. Hasil tuning menunjukkan parameter terbaik C = 1.0 dengan CV ROC-AUC (mean) = 0,9536.

**Tabel 10. Parameter Terbaik dari GridSearchCV**

Parameter	Nilai
classifier_C	1.000000
CV ROC-AUC (mean)	0.953600

Gambar 3. 8 Best parameter

**Tabel 11. Hasil Cross-Validation untuk Berbagai Nilai C**

C	ROC-AUC (CV Mean)	ROC-AUC (CV Std)	Train ROC-AUC (Mean)	Train ROC-AUC (Std)	Rank
1.000000	0.9536	0.0288	0.9769	0.0048	1
0.500000	0.9513	0.0275	0.9751	0.0042	2
2.000000	0.9503	0.0291	0.9788	0.0052	3
0.100000	0.9498	0.0353	0.9662	0.0045	4
0.200000	0.9489	0.0350	0.9713	0.0045	5
5.000000	0.9480	0.0258	0.9801	0.0051	6
10.000000	0.9441	0.0265	0.9803	0.0058	7
0.050000	0.9420	0.0468	0.9591	0.0048	8
0.010000	0.9198	0.0549	0.9317	0.0069	9

Gambar 3. 9 Hasil cross-validation

Rentang nilai C di sekitar 0,5–2,0 menghasilkan skor yang relatif kompetitif, sedangkan regularisasi yang terlalu kuat (C kecil) menurunkan performa, dan C yang terlalu besar mulai menunjukkan gap train vs validasi yang lebih besar.

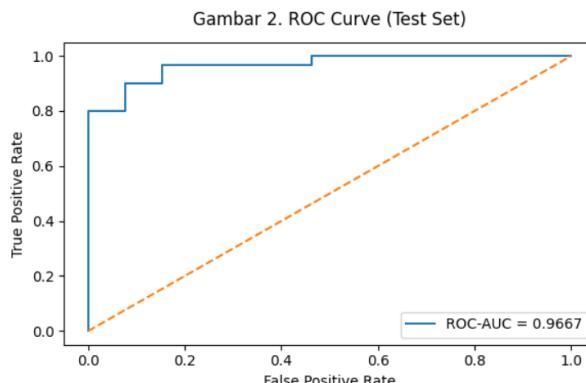
### 3.5 Evaluasi Model

**Tabel 12. Performa Model pada Test Set**

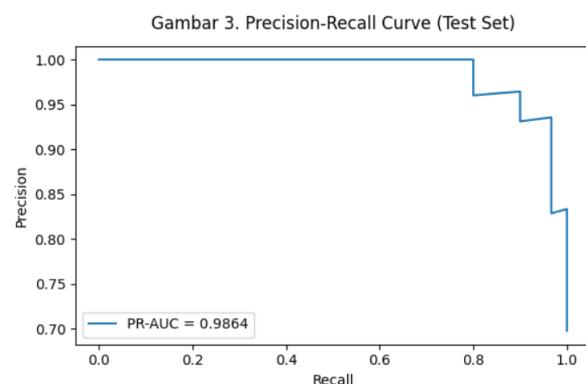
Metrik	Nilai
ROC-AUC	0.966700
PR-AUC (Avg Precision)	0.986400
F1-score	0.888900
Accuracy	0.860500

Gambar 3. 10 Performa model

Pada test set, model menunjukkan performa yang kuat dengan ROC-AUC = 0,9667, PR-AUC = 0,9864, F1-score = 0,8889, dan Accuracy = 0,8605. Kurva ROC dan Precision–Recall konsisten dengan metrik tersebut, yaitu model mampu memisahkan kelas dengan baik dan tetap stabil meskipun terdapat ketidakseimbangan kelas.



Gambar 3. 11 ROC Curve



Gambar 3. 12 Precision-Recall Curve

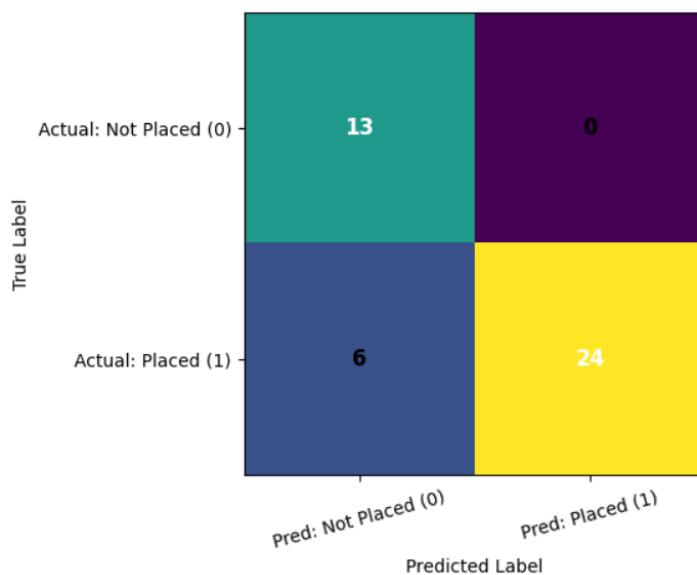
Confusion matrix memperlihatkan  $TN = 13$ ,  $FP = 0$ ,  $FN = 6$ ,  $TP = 24$ . Pola ini menunjukkan model sangat ketat dalam memprediksi Placed (karena tidak ada false positive), tetapi masih melewakan sebagian mahasiswa yang sebenarnya Placed (false negative). Dengan kata lain: ketika model bilang “Placed”, hasilnya cenderung dapat dipercaya, namun sebagian kasus Placed masih diklasifikasikan sebagai Not Placed. Ini perlu dicatat karena konsekuensi FN dan FP berbeda tergantung tujuan penggunaan (misalnya untuk early warning, FN biasanya lebih merugikan daripada FP).

**Tabel 13. Confusion Matrix (Test Set)**

	Pred: Not Placed (0)	Pred: Placed (1)
Actual: Not Placed (0)	13	0
Actual: Placed (1)	6	24

Gambar 3. 13 Confusin matrix

Gambar 4. Confusion Matrix Heatmap (Test Set)



Gambar 3. 14 Confusion Matrix Heatmap

### 3.6 Interpretasi Kontribusi Variabel

Interpretasi dilakukan melalui koefisien dan odds ratio (OR). Koefisien positif mengarah pada peningkatan peluang Placed, sedangkan koefisien negatif menurunkan peluang Placed. Karena fitur numerik dstandardisasi, interpretasi numerik dibaca sebagai efek kenaikan 1 standar deviasi pada variabel tersebut.

Tabel 15. Top 10 Koefisien Positif

Fitur	Koefisien	Odds Ratio
Nilai rata-rata SMP	1.9804	7.2453
IPK	1.0802	2.9454
Pengalaman kerja sebelum lulus = Yes	1.0151	2.7596
Program studi sarjana = Comm&Mgmt	0.7504	2.1178
Nilai rata-rata SMA	0.7157	2.0457
Lembaga pendidikan kelas 12 = Internasional	0.6684	1.9510
Jenis Kelamin = M	0.6667	1.9478
Lembaga pendidikan kelas 10 = Internasional	0.6322	1.8817
Lembaga pendidikan kelas 12 = Negeri	0.5441	1.7231
Jurusan saat SMA = Arts	0.5349	1.7073

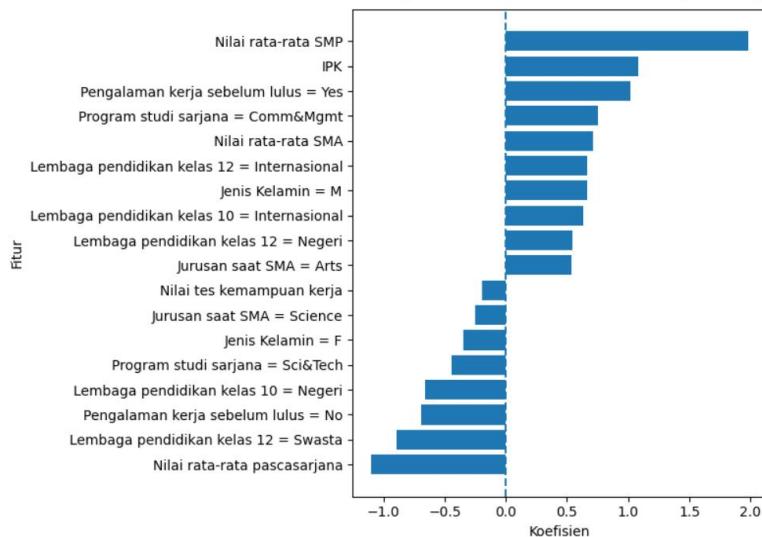
Gambar 3. 15 Koefisien Positif

Tabel 16. Top 10 Koefisien Negatif

Fitur	Koefisien	Odds Ratio
Nilai rata-rata pascasarjana	-1.1010	0.3325
Lembaga pendidikan kelas 12 = Swasta	-0.8935	0.4092
Pengalaman kerja sebelum lulus = No	-0.6961	0.4985
Lembaga pendidikan kelas 10 = Negeri	-0.6615	0.5161
Program studi sarjana = Sci&Tech	-0.4444	0.6412
Jenis Kelamin = F	-0.3477	0.7063
Jurusan saat SMA = Science	-0.2496	0.7791
Nilai tes kemampuan kerja	-0.1979	0.8205

Gambar 3. 16 Koefisien Negatif

Gambar 4. Top Kontribusi Variabel (Koefisien Logistic Regression)



Gambar 3. 17 Top kontribusi variabel

Variabel dengan kontribusi positif paling kuat adalah Nilai rata-rata SMP (OR  $\approx 7,25$ ), diikuti IPK (OR  $\approx 2,95$ ) dan Pengalaman kerja sebelum lulus = Yes (OR  $\approx 2,76$ ). Ini memberi sinyal bahwa kombinasi kemampuan akademik dan exposure dunia kerja berhubungan kuat dengan placement dalam dataset ini.

Pada sisi negatif, Nilai rata-rata pascasarjana muncul dengan koefisien negatif (OR  $\approx 0,33$ ). Temuan semacam ini perlu dibaca hati-hati karena efeknya bersifat kondisional terhadap variabel lain. Jadi posisinya lebih tepat sebagai indikator pola yang layak dianalisis lanjut (misalnya korelasi antarfaktor, segmentasi), bukan bahan klaim sebab-akibat.

Tabel 17. Top 12 Variabel Paling Berkontribusi (Regularized Logistic Regression)

Fitur	Jenis	Koefisien	Odds Ratio	Arah	Interpretasi singkat
Nilai rata-rata SMP	Numerik	1.980400	7.245300	Positif	Kenaikan 1 SD pada variabel ini cenderung meningkatkan peluang Placed (odds Placed sekitar 7.25x).
Nilai rata-rata pascasarjana	Numerik	-1.101000	0.332500	Negatif	Kenaikan 1 SD pada variabel ini cenderung menurunkan peluang Placed (odds Placed menjadi sekitar 0.33x).
IPK	Numerik	1.080200	2.945400	Positif	Kenaikan 1 SD pada variabel ini cenderung meningkatkan peluang Placed (odds Placed sekitar 2.95x).
Pengalaman kerja sebelum lulus = Yes	Kategorikal	1.015100	2.759600	Positif	Kategori ini cenderung meningkatkan peluang Placed (odds Placed sekitar 2.76x) dibanding kategori acuan.
Lembaga pendidikan kelas 12 = Swasta	Kategorikal	-0.893500	0.409200	Negatif	Kategori ini cenderung menurunkan peluang Placed (odds Placed menjadi sekitar 0.41x) dibanding kategori acuan.
Program studi sarjana = Comm&Mgmt	Kategorikal	0.750400	2.117800	Positif	Kategori ini cenderung meningkatkan peluang Placed (odds Placed sekitar 2.12x) dibanding kategori acuan.
Nilai rata-rata SMA	Numerik	0.715700	2.045700	Positif	Kenaikan 1 SD pada variabel ini cenderung meningkatkan peluang Placed (odds Placed sekitar 2.05x).
Pengalaman kerja sebelum lulus = No	Kategorikal	-0.696100	0.498500	Negatif	Kategori ini cenderung menurunkan peluang Placed (odds Placed menjadi sekitar 0.50x) dibanding kategori acuan.
Lembaga pendidikan kelas 12 = Internasional	Kategorikal	0.668400	1.951000	Positif	Kategori ini cenderung meningkatkan peluang Placed (odds Placed sekitar 1.95x) dibanding kategori acuan.
Jenis Kelamin = M	Kategorikal	0.666700	1.947800	Positif	Kategori ini cenderung meningkatkan peluang Placed (odds Placed sekitar 1.95x) dibanding kategori acuan.
Lembaga pendidikan kelas 10 = Negeri	Kategorikal	-0.661500	0.516100	Negatif	Kategori ini cenderung menurunkan peluang Placed (odds Placed menjadi sekitar 0.52x) dibanding kategori acuan.
Lembaga pendidikan kelas 10 = Internasional	Kategorikal	0.632200	1.881700	Positif	Kategori ini cenderung meningkatkan peluang Placed (odds Placed sekitar 1.88x) dibanding kategori acuan.

Gambar 3. 18 Variabel paling berkongruensi

### 3.7 Insight dan Rekomendasi

Berdasarkan tabel insight, rekomendasi difokuskan pada variabel aksiabile yang dapat ditangani lewat program kampus. Dua jalur yang paling masuk akal adalah:

1. Dukungan akademik lebih awal (berbasis sinyal nilai SMP/SMA dan IPK), melalui bimbingan, remedial, dan monitoring progres.
2. Penguatan employability terutama untuk mahasiswa tanpa pengalaman kerja, dengan magang wajib, mentoring karier, serta simulasi rekrutmen.

Variabel yang bersifat historis (misalnya latar lembaga pendidikan) lebih tepat dipakai untuk pemetaan risiko tingkat kelompok, bukan keputusan per individu. Untuk variabel sensitif seperti jenis kelamin, penggunaan yang masuk akal adalah audit fairness model, bukan intervensi personal.

Tabel 18. Insight dan Rekomendasi Praktis (Berbasis Hasil Model)

<b>Fitur</b>	<b>Arah</b>	<b>Odds Ratio</b>	<b>Prioritas</b>	<b>Kategori rekomendasi</b>	<b>Rekomendasi</b>
Nilai rata-rata SMP	Positif	7.250000	Tinggi	Aksiabile (dukungan akademik)	Gunakan sebagai sinyal awal pemetaan kebutuhan dukungan akademik; intervensi berupa bimbingan, remedial, dan monitoring progres.
Nilai rata-rata pascasarjana	Negatif	0.330000	Tinggi	Kurang aksiabile (lebih historis)	Hindari interpretasi kausal; gunakan sebagai indikator risiko dan lakukan analisis lanjutan (korelasi, multikolinearitas, segmentasi).
IPK	Positif	2.950000	Tinggi	Aksiabile (langsung bisa diintervensi)	Terapkan early support akademik (tutoring/clinic) untuk mahasiswa dengan IPK rendah, fokus pada mata kuliah bottleneck.
Pengalaman kerja sebelum lulus = Yes	Positif	2.760000	Tinggi	Aksiabile (langsung bisa diintervensi)	Skalakan program magang/kerja praktik;jadikan jalur utama peningkatan employability sebelum lulus.
Lembaga pendidikan kelas 12 = Swasta	Negatif	0.410000	Sedang	Kurang aksiabile (lebih historis)	Gunakan untuk pemetaan risiko tingkat kelompok; rekomendasi bersifat programatik (bridging, pelatihan tambahan), bukan keputusan per individu.
Program studi sarjana = Comm&Mgmt	Positif	2.120000	Sedang	Kurang aksiabile (lebih historis)	Gunakan untuk pemetaan risiko tingkat kelompok; rekomendasi bersifat programatik (bridging, pelatihan tambahan), bukan keputusan per individu.
Nilai rata-rata SMA	Positif	2.050000	Sedang	Aksiabile (dukungan akademik)	Gunakan sebagai sinyal awal pemetaan kebutuhan dukungan akademik; intervensi berupa bimbingan, remedial, dan monitoring progres.
Pengalaman kerja sebelum lulus = No	Negatif	0.500000	Sedang	Aksiabile (langsung bisa diintervensi)	Prioritaskan mahasiswa tanpa pengalaman kerja untuk magang wajib, mentoring karier, dan simulasi rekrutmen.
Lembaga pendidikan kelas 12 = Internasional	Positif	1.950000	Rendah	Kurang aksiabile (lebih historis)	Gunakan untuk pemetaan risiko tingkat kelompok; rekomendasi bersifat programatik (bridging, pelatihan tambahan), bukan keputusan per individu.
Jenis Kelamin = M	Positif	1.950000	Rendah	Sensitif (audit fairness)	Gunakan untuk audit bias model (mis. perbandingan error rate antar gender), bukan untuk intervensi individual.
Lembaga pendidikan kelas 10 = Negeri	Negatif	0.520000	Rendah	Kurang aksiabile (lebih historis)	Gunakan untuk pemetaan risiko tingkat kelompok; rekomendasi bersifat programatik (bridging, pelatihan tambahan), bukan keputusan per individu.
Lembaga pendidikan kelas 10 = Internasional	Positif	1.880000	Rendah	Kurang aksiabile (lebih historis)	Gunakan untuk pemetaan risiko tingkat kelompok; rekomendasi bersifat programatik (bridging, pelatihan tambahan), bukan keputusan per individu.

Gambar 3. 19 Rekomendasi Variabel

## **BAB IV**

### **KESIMPULAN**

#### **4.1 Kesimpulan**

1. Pembangunan model klasifikasi berhasil dilakukan dengan pendekatan Regularized Logistic Regression yang terintegrasi dalam pipeline preprocessing (imputasi, one-hot encoding, dan scaling), serta pembagian data menggunakan stratified train-test split sehingga distribusi kelas tetap terjaga. Model kemudian dituning menggunakan Stratified K-Fold Cross Validation, dan menghasilkan parameter terbaik  $C = 1.0$  dengan CV ROC-AUC (mean) = 0,9536, menunjukkan model cukup stabil dalam validasi.
2. Variabel yang paling berkontribusi terhadap peluang mahasiswa menjadi Placed dapat diinterpretasikan melalui koefisien dan odds ratio. Hasil menunjukkan kontribusi positif terkuat berasal dari Nilai rata-rata SMP ( $OR \approx 7,25$ ), diikuti IPK ( $OR \approx 2,95$ ) dan Pengalaman kerja sebelum lulus = Yes ( $OR \approx 2,76$ ). Ini mengindikasikan bahwa kombinasi performa akademik dan pengalaman kerja sebelum lulus berkaitan kuat dengan peluang placement pada dataset. Pada sisi negatif, Nilai rata-rata pascasarjana muncul dengan koefisien negatif ( $OR \approx 0,33$ ) dan perlu dibaca sebagai pola kondisional dalam model, bukan dasar klaim sebab-akibat.
3. Performa model pada test set tergolong kuat dengan ROC-AUC = 0,9667, PR-AUC = 0,9864, F1-score = 0,8889, dan Accuracy = 0,8605, sehingga model mampu membedakan kelas Placed dan Not Placed dengan baik pada evaluasi akhir. Confusion matrix menunjukkan TN = 13, FP = 0, FN = 6, TP = 24, yang berarti model cenderung “ketat” memberi label Placed (tidak ada false positive), namun masih melewatkannya beberapa kasus Placed (false negative).

## **LAMPIRAN**

REPOSITORY GITHUB:

<https://github.com/diionlyone/UAS-DATA-SCIENCE>