LAPORAN UAS DATA SCIENCE

Studi Kasus: Faktor Akademik yang Mempengaruhi Penempatan Kerja (Campus Recruitment)

Link GitHub: https://github.com/ibrahim57448/UAS-Data-Science-Campus-Recruitment

Dosen Pengampu : Nur Hayati, S.SI., MTI

Disusun oleh :

SLAMET IBRAHIM

227006416154

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNOLOGI KOMUNIKASI DAN INFORMATIKA

UNIVERSITAS NASIONAL

2026

LAPORAN UAS DATA SCIENCE

Studi Kasus: Faktor Akademik yang Mempengaruhi Penempatan Kerja (Campus Recruitment)

1. Latar Belakang

Perguruan tinggi memiliki tanggung jawab untuk memastikan lulusannya siap memasuki dunia kerja. Salah satu indikator keberhasilan tersebut adalah tingkat keberhasilan penempatan kerja (placement) melalui program campus recruitment. Namun, tidak semua mahasiswa berhasil mendapatkan pekerjaan melalui campus recruitment dan tingkat keberhasilan placement bervariasi antar mahasiswa. Karena itu, dibutuhkan pendekatan Data Science untuk menganalisis data historis mahasiswa, membangun model prediksi peluang placement, serta mengidentifikasi faktor akademik dan non-akademik yang paling berpengaruh.

2. Tujuan Bisnis

Tujuan bisnis dari studi kasus ini adalah membantu institusi pendidikan dalam meningkatkan tingkat keberhasilan placement mahasiswa. Dengan mengetahui faktor yang paling berpengaruh, kampus dapat menyusun strategi pembinaan yang lebih tepat sasaran (misalnya pada aspek nilai akademik, kesiapan tes kemampuan kerja, atau penguatan program pascasarjana).

3. Tujuan Teknis Data Science

Secara teknis, pekerjaan Data Science pada UAS ini mencakup: (1) memahami data dan struktur variabel, (2) melakukan EDA untuk melihat pola dan anomali, (3) preprocessing (menangani missing value, encoding kategori, standarisasi numerik), (4) membangun model klasifikasi untuk memprediksi status placement, (5) mengevaluasi performa model menggunakan Confusion Matrix dan ROC Curve, serta (6) menginterpretasikan faktor dominan melalui feature importance.

4. Deskripsi Dataset dan Variabel

Dataset yang digunakan berjumlah 215 baris dan 15 kolom. Target prediksi adalah kolom ****'status kelulusan (Bekerja/Belum)'**.**

Berikut ringkasan variabel (nama kolom mengikuti file CSV):

Variabel	Tipe	Deskripsi Singkat
ID	int64	Deskripsi variabel belum tersedia (mengikuti nama kolom).
Jenis Kelamin	object	Jenis kelamin mahasiswa (M/F).
Nilai rata-rata SMP	float64	Nilai akademik setara secondary school (kelas 10) dalam persen.
Lembaga pendidikan kelas 10	object	Jenis sekolah kelas 10 (Negeri/Swasta/Internasional).
Nilai rata-rata SMA	float64	Nilai akademik setara higher secondary school (kelas 12) dalam persen.
Lembaga pendidikan kelas 12	object	Jenis sekolah kelas 12 (Negeri/Swasta/Internasional).
Jurusan saat SMA	object	Jurusan ketika SMA (Science/Commerce/Arts).
IPK	float64	Nilai/performansi akademik pada jenjang sarjana (di dataset ini berupa angka).
Program studi sarjana	object	Kelompok program studi sarjana (mis. Sci&Tech, Comm&Mgmt).
Pengalaman kerja sebelum lulus	object	Apakah memiliki pengalaman kerja sebelum lulus (Yes/No).
Nilai tes kemampuan kerja	float64	Nilai employability test (etest_p).
Pendidikan pascasarjana	object	Spesialisasi pascasarjana/MBA (mis. Mkt&HR, Mkt&Fin).

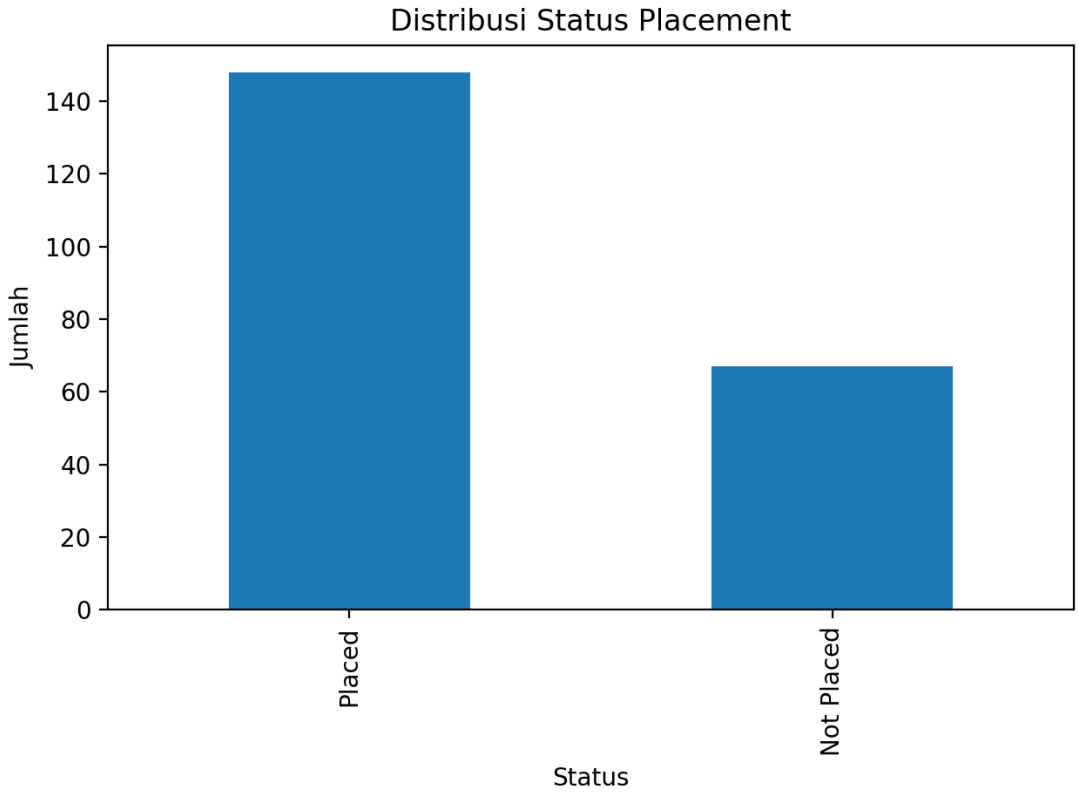
Variabel	Tipe	Deskripsi Singkat
Nilai rata-rata pascasarjana	float64	Nilai MBA/pascasarjana (mba_p).
status kelulusan (Bekerja/Belum)	object	Label target: Placed / Not Placed.
Gaji	float64	Gaji (hanya terisi untuk yang Placed).

Catatan: kolom ****Gaji**** hanya terisi untuk mahasiswa yang 'Placed', sehingga tidak digunakan sebagai fitur prediksi karena berpotensi menyebabkan ***data leakage*** (informasi yang seharusnya tidak diketahui saat memprediksi).

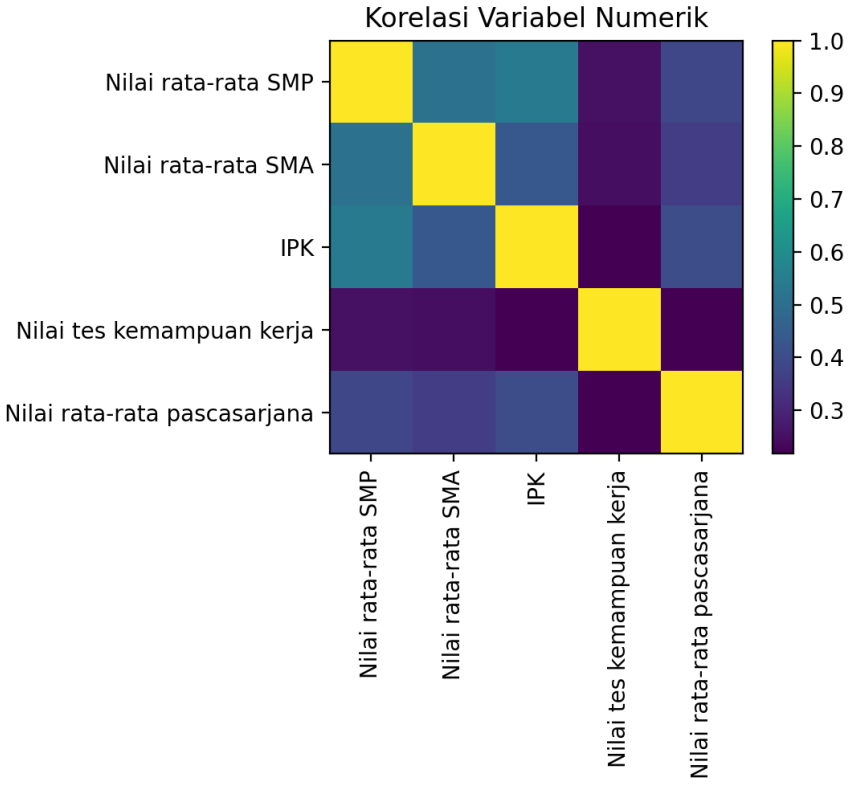
5. Exploratory Data Analysis (EDA)

Pada tahap EDA, tujuan utama adalah memahami distribusi target, melihat korelasi antar nilai numerik, dan membandingkan perbedaan karakteristik antara kelompok 'Placed' dan 'Not Placed'. Grafik-grafik berikut membantu menjelaskan pola awal sebelum pemodelan.

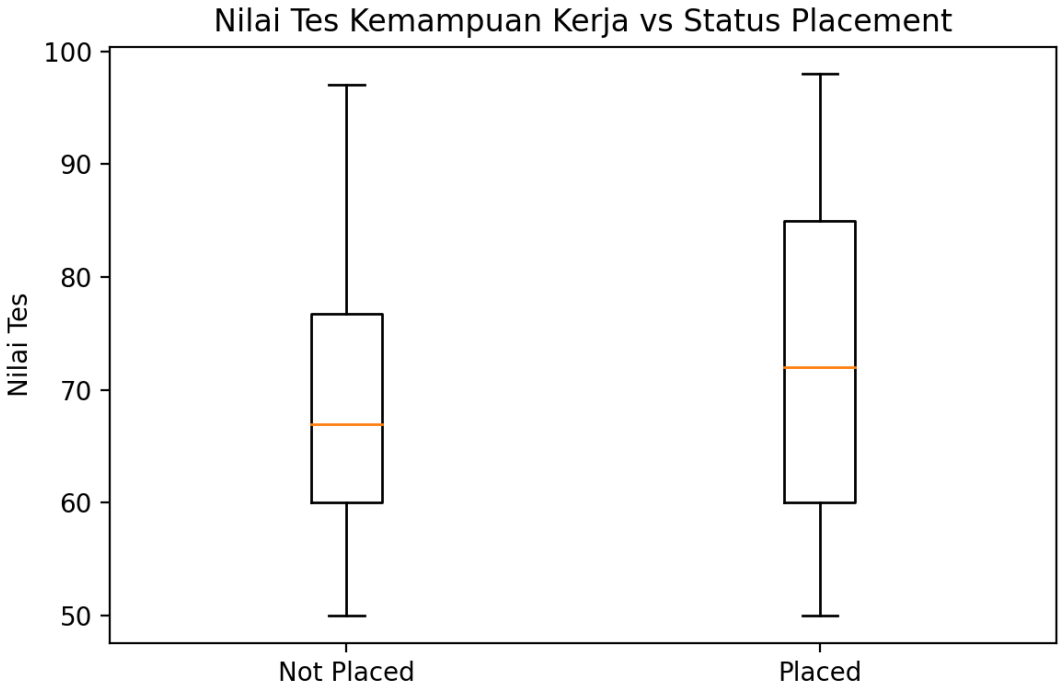
Gambar 1. Distribusi Status Placement



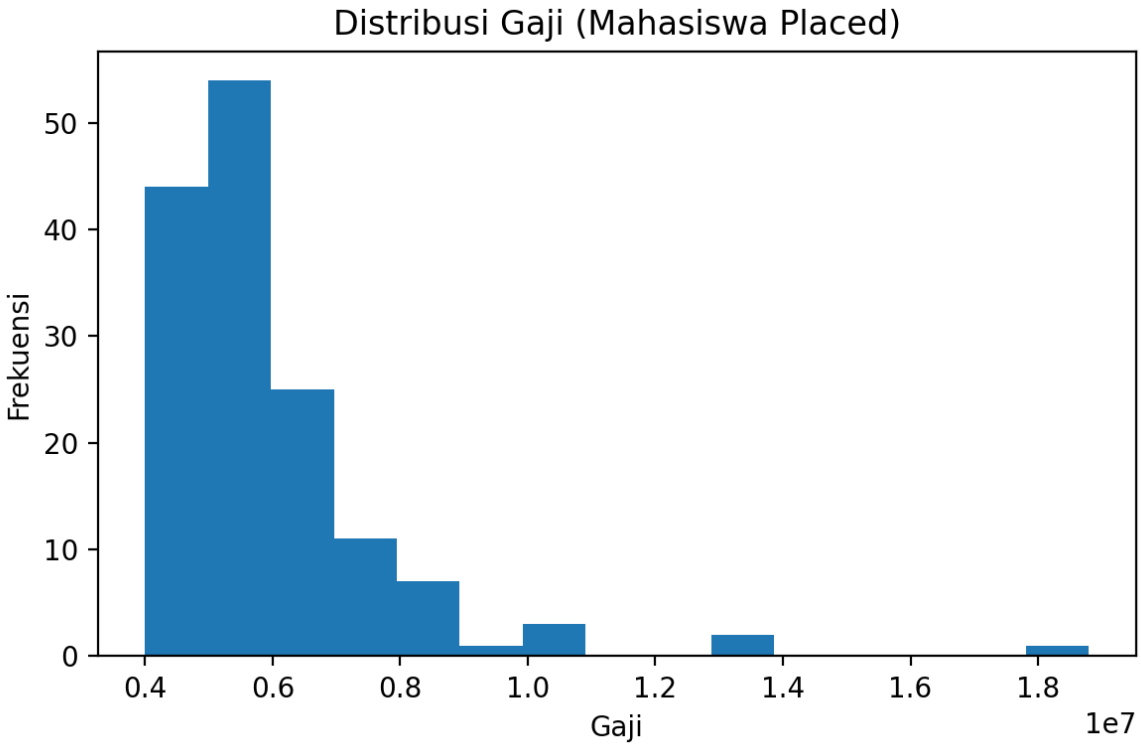
Gambar 2. Korelasi Variabel Numerik (Heatmap)



Gambar 3. Boxplot Nilai Tes Kemampuan Kerja vs Status Placement



Gambar 4. Distribusi Gaji untuk Mahasiswa yang Placed



Interpretasi singkat: distribusi target menunjukkan ketidakseimbangan kelas (Placed cenderung lebih banyak). Korelasi numerik memberikan gambaran hubungan antar nilai akademik. Boxplot memperlihatkan kecenderungan nilai tes kemampuan kerja yang lebih tinggi pada kelompok Placed. Histogram gaji digunakan sebagai insight tambahan (bukan fitur model).

6. Preprocessing

Preprocessing dilakukan agar data siap digunakan untuk model machine learning. Langkah yang dilakukan: (1) menghapus kolom identitas (ID) dan tidak memakai Gaji sebagai fitur, (2) menangani missing value (numerik diisi median, kategori diisi modus), (3) mengubah variabel kategori menjadi numerik dengan One-Hot Encoding, dan (4) melakukan standarisasi fitur numerik menggunakan StandardScaler.

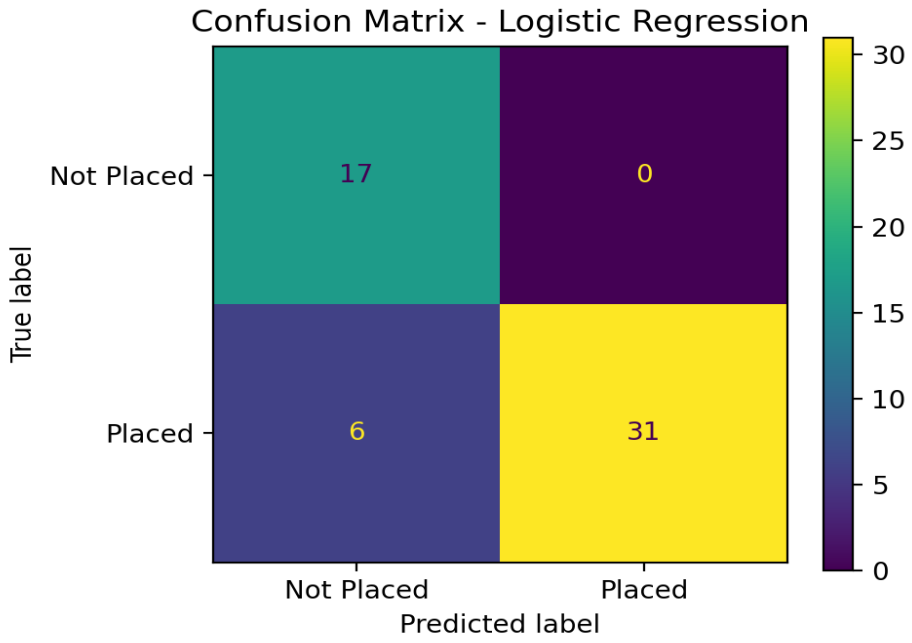
7. Pemodelan

Dua model klasifikasi digunakan untuk membandingkan performa: Logistic Regression (sebagai baseline yang mudah diinterpretasi) dan Random Forest (model ensemble yang sering kuat untuk data campuran numerik-kategorikal). Data dibagi menjadi train-test dengan rasio 75:25 dan stratifikasi berdasarkan label agar proporsi kelas tetap terjaga.

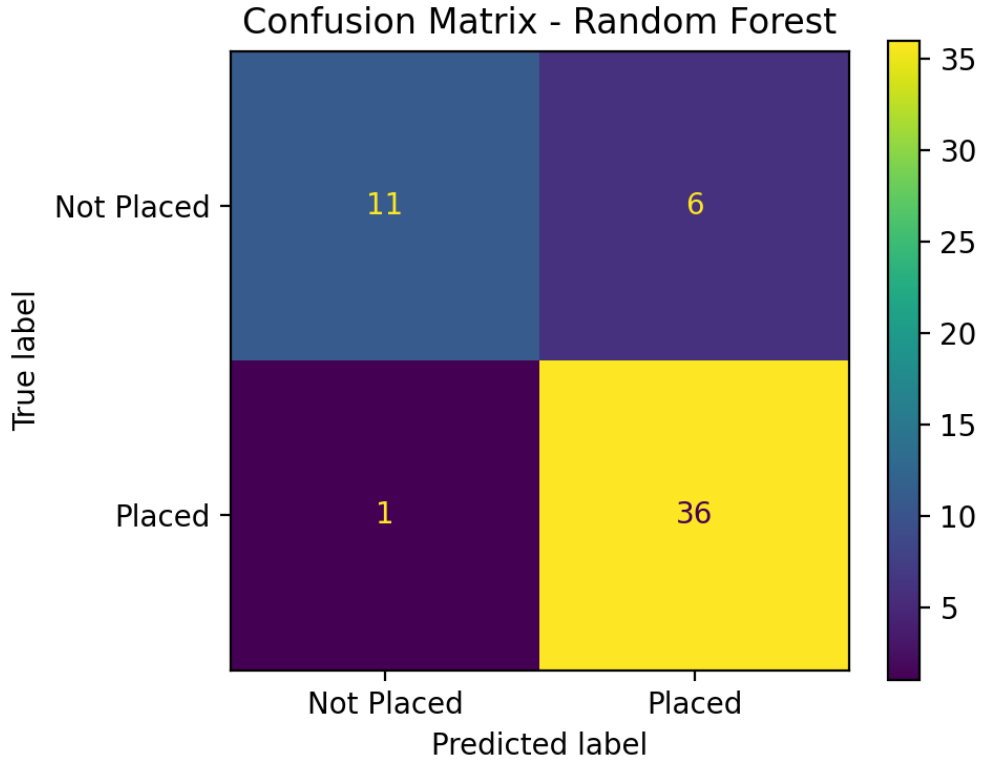
8. Evaluasi Model

Evaluasi dilakukan menggunakan Confusion Matrix, classification report (precision, recall, f1-score), dan ROC Curve (AUC). Semakin besar AUC, semakin baik model dalam membedakan kelas Placed vs Not Placed.

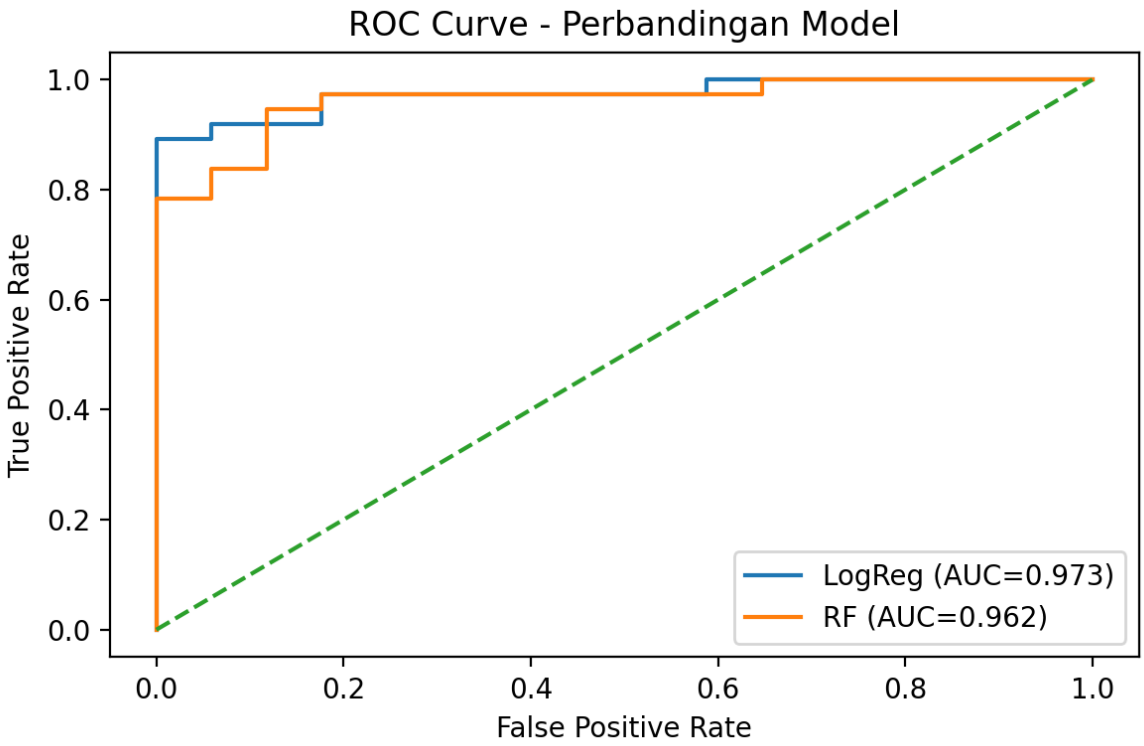
Gambar 5. Confusion Matrix - Logistic Regression



Gambar 6. Confusion Matrix - Random Forest



Gambar 7. ROC Curve - Perbandingan Model



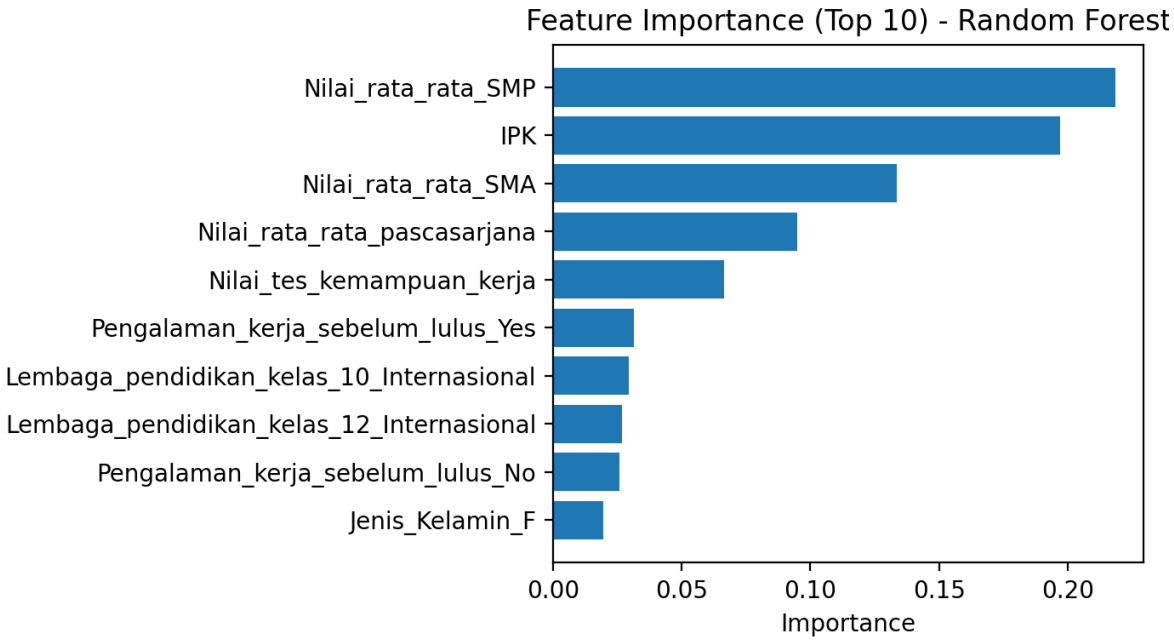
Model	Accuracy	Precision (Placed)	Recall (Placed)	F1 (Placed)	AUC
Logistic Regression	0.889	1.000	0.838	0.912	0.973
Random Forest	0.870	0.857	0.973	0.911	0.962

Dari hasil di atas, Logistic Regression memiliki AUC sedikit lebih tinggi pada data uji, sementara Random Forest memberikan performa yang cukup kompetitif dan unggul untuk interpretasi berbasis feature importance. Pemilihan model terbaik dapat mem-pertimbangkan keseimbangan akurasi dan kemampuan interpretasi.

9. Feature Importance (Top 10)

Feature importance digunakan untuk mengetahui fitur mana yang paling berpengaruh dalam prediksi placement (berdasarkan Random Forest). Karena terdapat variabel kategorikal, setelah One-Hot Encoding akan muncul fitur turunan seperti 'Jurusan_saat_SMA_Science' dan sebagainya.

Gambar 8. Feature Importance Top 10 - Random Forest



Secara umum, fitur numerik seperti nilai rata-rata SMP/SMA, IPK, nilai tes kemampuan kerja, serta nilai pascasarjana cenderung muncul sebagai faktor dominan. Hal ini konsisten dengan tujuan studi kasus yang menekankan pengaruh faktor akademik terhadap peluang placement.

10. Kesimpulan

Berdasarkan analisis dan pemodelan, faktor akademik (nilai rata-rata SMP/SMA, IPK, nilai tes kemampuan kerja, dan nilai pascasarjana) memiliki kontribusi signifikan terhadap peluang penempatan kerja. Model Logistic Regression dan Random Forest sama-sama mampu memprediksi status placement dengan performa yang baik pada data uji (AUC > 0.96). Untuk kebutuhan interpretasi dan pengambilan kebijakan, Random Forest dapat digunakan untuk melihat peringkat pengaruh fitur, sementara Logistic Regression berguna sebagai baseline yang sederhana.

11. Link GitHub (Wajib)

Link GitHub: https://github.com/ibrahim57448/UAS-Data-Science-Campus-Recruitment