Nama & NPM : Raian Naufal Rahmat (227006516028)
Nama & NPM : Andra Cesario Febriansyah (227006516077)
Matakuliah : Data Science

Link Colab:
https://colab.research.google.com/drive/1PcjRJRP0GX6drLZ9hIe1QRpqyO9wqN8e?usp=sharing

UAS DATA SCIENCE (Laporan)

```
df = pd.read_csv("Campus Recruitment.csv")
df.head()
```

Kode ini menggunakan pustaka Pandas (diasumsikan sudah diimpor sebagai pd) untuk melakukan dua langkah dasar dalam analisis data. Baris pertama, df = pd.read_csv("Campus Recruitment.csv"), berfungsi untuk membaca file CSV bernama "Campus Recruitment.csv" dan memuat data tersebut ke dalam sebuah struktur data Pandas yang disebut DataFrame, yang disimpan dalam variabel df. Baris kedua, df.head(), kemudian menampilkan lima baris pertama dari DataFrame tersebut, yang memungkinkan Anda untuk melakukan inspeksi cepat terhadap struktur, nama kolom, dan isi data untuk memastikan file telah dimuat dengan benar.

```
df.info()
df.describe()
```

Kode ini melakukan eksplorasi data menyeluruh dengan menggunakan df.info() untuk menampilkan struktur teknis data seperti jumlah baris, tipe data, dan nilai yang hilang, serta menggunakan df.describe() untuk menghasilkan ringkasan statistik deskriptif—seperti rata-rata, standar deviasi, dan nilai kuartil—khusus untuk kolom-kolom numerik dalam DataFrame.

```
df.columns
```

```
Index(['ID', 'Jenis Kelamin', 'Nilai rata-rata SMP',
       'Lembaga pendidikan kelas 10', 'Nilai rata-rata SMA',
       'Lembaga pendidikan kelas 12', 'Jurusan saat SMA', 'IPK',
       'Program studi sarjana', 'Pengalaman kerja sebelum lulus',
       'Nilai tes kemampuan kerja', 'Pendidikan pascasarjana',
       'Nilai rata-rata pascasarjana', 'status kelulusan (Bekerja/Belum)',
       'Gaji'],
      dtype='object')
```

Kode ini digunakan untuk melakukan pemeriksaan mendalam terhadap struktur dan karakteristik data dalam DataFrame. Fungsi df.info() memberikan gambaran teknis mengenai tipe data dan keberadaan nilai yang hilang pada setiap kolom, sementara df.describe() menyajikan ringkasan statistik untuk kolom numerik. Perintah df.columns kemudian menampilkan daftar seluruh nama kolom yang tersedia, seperti 'ID', 'Jenis Kelamin', 'IPK', hingga 'status kelulusan (Bekerja/Belum)', sehingga Anda dapat mengidentifikasi variabel mana saja yang akan dianalisis lebih lanjut.

```
sns.countplot(x='status kelulusan (Bekerja/Belum)', data=df)
plt.title("Distribusi Placement")
plt.xticks(rotation=45)
plt.show()
```

Kode ini bertujuan untuk memvisualisasikan data kategori menggunakan pustaka Seaborn (sns) dan Matplotlib (plt). Melalui fungsi sns.countplot(), kode ini membuat sebuah grafik batang yang menghitung frekuensi atau jumlah kemunculan setiap kategori dalam kolom 'status kelulusan (Bekerja/Belum)', sehingga Anda bisa melihat perbandingan jumlah alumni yang sudah bekerja versus yang belum secara visual. Penambahan fungsi plt.title() memberi label judul "Distribusi Placement", sementara plt.xticks(rotation=45) digunakan untuk memutar label pada sumbu X agar lebih mudah dibaca jika teksnya terlalu panjang.

```
sns.countplot(x='Jenis Kelamin', data=df)
plt.title("Distribusi Gender")
plt.show()
```

Seluruh rangkaian kode ini berfungsi untuk melakukan analisis data eksploratif (EDA) dan visualisasi pada dataset rekrutmen kampus. Langkah awal melibatkan penggunaan df.info() untuk mengecek struktur data serta nilai yang hilang, df.describe() untuk mendapatkan ringkasan statistik seperti rata-rata dan kuartil pada kolom numerik, serta df.columns untuk mengidentifikasi variabel yang tersedia seperti 'Jenis Kelamin', 'IPK', dan 'Gaji'. Setelah memahami struktur datanya, kode kemudian beralih ke tahap visualisasi menggunakan pustaka Seaborn dan Matplotlib untuk membuat grafik batang (countplot) yang menunjukkan distribusi data kategoris; grafik pertama menampilkan perbandingan jumlah mahasiswa berdasarkan 'status kelulusan (Bekerja/Belum)', sementara grafik kedua memvisualisasikan 'Distribusi Gender' untuk melihat komposisi jenis kelamin dalam dataset tersebut.

```
sns.boxplot(x='status kelulusan (Bekerja/Belum)', y='IPK', data=df)
plt.title("IPK vs Placement")
plt.show()
```

keseluruhan, rangkaian kode ini menjalankan proses Analisis Data Eksploratif (EDA) untuk memahami faktor-faktor yang memengaruhi keberhasilan rekrutmen mahasiswa. Proses dimulai dengan pemeriksaan struktur data menggunakan df.info(), df.describe(), dan df.columns untuk memahami tipe variabel (seperti 'IPK', 'Jenis Kelamin', dan 'Gaji') serta statistik dasarnya. Selanjutnya, kode beralih ke visualisasi menggunakan Seaborn untuk memetakan distribusi frekuensi melalui *countplot* pada variabel 'status kelulusan' dan 'Jenis Kelamin'. Puncaknya, penggunaan sns.boxplot() pada bagian akhir bertujuan untuk membandingkan distribusi nilai 'IPK' terhadap 'status kelulusan', yang memungkinkan Anda mengidentifikasi secara visual apakah mahasiswa dengan IPK lebih tinggi cenderung memiliki peluang lebih besar untuk mendapatkan status 'Bekerja'.

```
sns.countplot(x='Pengalaman kerja sebelum lulus', hue='status kelulusan (Bekerja/Belum)', data=df)
plt.title("Pengalaman Kerja vs Placement")
plt.show()
```

Rangkaian kode ini menjalankan alur Analisis Data Eksploratif (EDA) yang komprehensif untuk mengidentifikasi faktor-faktor penentu dalam rekrutmen kampus. Proses dimulai dengan fungsi df.info(), df.describe(), dan df.columns untuk membedah profil data, mulai dari tipe variabel hingga statistik kunci pada kolom numerik seperti 'IPK' dan 'Gaji'. Analisis kemudian diperdalam melalui visualisasi menggunakan Seaborn; selain memetakan distribusi frekuensi sederhana untuk 'status kelulusan' dan 'Jenis Kelamin', kode ini menggunakan sns.boxplot() untuk membandingkan rentang 'IPK' terhadap status bekerja guna mendeteksi pengaruh prestasi akademik secara visual. Terakhir, penggunaan grafik batang dengan parameter hue pada variabel 'Pengalaman kerja sebelum lulus' memungkinkan Anda melihat secara spesifik bagaimana pengalaman kerja nyata berkontribusi terhadap peluang alumni untuk mendapatkan status 'Bekerja' dibandingkan mereka yang belum memiliki pengalaman.

```
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```

Rangkaian kode ini menjalankan alur Analisis Data Eksploratif (EDA) yang komprehensif untuk mengidentifikasi faktor-faktor penentu dalam rekrutmen kampus. Proses dimulai dengan pemeriksaan profil data melalui df.info(), df.describe(), dan df.columns untuk memahami tipe variabel serta statistik kunci pada kolom numerik seperti 'IPK' dan 'Gaji'. Analisis kemudian divisualisasikan menggunakan Seaborn untuk memetakan distribusi frekuensi pada variabel 'status kelulusan' dan 'Jenis Kelamin', serta menggunakan sns.boxplot() untuk mendeteksi pengaruh prestasi akademik (IPK) terhadap peluang kerja secara visual. Kode ini juga mengeksplorasi hubungan antar variabel lebih dalam dengan membandingkan 'Pengalaman kerja sebelum lulus' terhadap status rekrutmen, dan ditutup dengan pembuatan Correlation Matrix menggunakan sns.heatmap() untuk melihat kekuatan hubungan linear antar semua variabel numerik, sehingga Anda dapat menentukan faktor mana yang paling berkorelasi dengan kesuksesan karier alumni.

```
sns.histplot(df['Nilai tes kemampuan kerja'], kde=True)
plt.title("Distribusi Nilai Employability Test")
plt.show()
```

Rangkaian kode ini menyajikan alur kerja Analisis Data Eksploratif (EDA) yang sangat mendalam untuk membedah dataset rekrutmen kampus. Proses dimulai dengan audit data menggunakan df.info(), df.describe(), dan df.columns untuk memahami struktur teknis, statistik variabel numerik (seperti 'IPK' dan 'Gaji'), serta daftar kolom yang tersedia. Analisis kemudian berlanjut ke tahap visualisasi menggunakan Seaborn untuk memetakan distribusi frekuensi pada variabel 'status kelulusan' dan 'Jenis Kelamin', serta menggunakan sns.boxplot() guna membandingkan performa akademik (IPK) terhadap peluang kerja. Lebih jauh lagi, kode ini mengevaluasi dampak 'Pengalaman kerja sebelum lulus' terhadap kesuksesan penempatan kerja dan menggunakan sns.heatmap() untuk menghitung Correlation Matrix, yang berfungsi mengukur kekuatan hubungan antar seluruh variabel numerik. Terakhir, kode ditutup dengan sns.histplot() pada kolom 'Nilai tes kemampuan kerja' lengkap dengan kurva Kernel Density Estimate (KDE) untuk memvisualisasikan sebaran skor kemampuan kerja para lulusan secara halus dan mendetail.

```
model = LogisticRegression(max_iter=5000)
model.fit(X_train, y_train)
```

Kode ini melakukan Analisis Data Eksploratif (EDA) dan pemodelan Machine Learning dengan memuat data, mengecek statistik, memvisualisasikan korelasi antar variabel (seperti IPK dan Gaji), serta melatih model Logistic Regression untuk memprediksi status kelulusan.

```
importance = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': model.coef_[0]
}).sort_values(by='Coefficient', ascending=False)

print(importance.head(10))
```

Kode ini merupakan alur lengkap Data Science yang mencakup analisis data eksploratif (EDA) untuk memahami pola dataset, pelatihan model Logistic Regression untuk prediksi kelulusan, dan ekstraksi Feature Importance untuk mengidentifikasi variabel yang paling berpengaruh terhadap peluang kerja.

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Kode ini melakukan alur kerja Data Science lengkap: memuat dan memeriksa data, memvisualisasikan korelasi dan distribusi (termasuk heatmap korelasi), melatih model Logistic Regression untuk memprediksi kelulusan, serta menampilkan metrik evaluasi seperti akurasi dan feature importance.