LAPORAN PRAKTIKUM
UTS DATA SCIENCE
"Analisis Data Student Academic Status"

Kelompok 26:

41425078 Daniel Siahaan
41425079 Jessica Pasaribu
41425080 Novrael Gabriel Louis Marbun

FAKULTAS VOKASI
INSTITUT TEKNOLOGI DEL

A. Pendahuluan

1. Latar Belakang

Data Science merupakan bidang yang berfokus pada pengolahan, analisis, dan interpretasi data untuk menghasilkan informasi dan wawasan yang dapat digunakan dalam pengambilan keputusan. Dalam konteks pendidikan, analisis data mahasiswa menjadi hal penting untuk memahami faktor-faktor yang memengaruhi keberhasilan akademik, seperti nilai masuk, latar belakang pendidikan, dan performa selama perkuliahan.

Proyek ini merupakan implementasi praktis dari tahapan analisis data menggunakan metode data science pipeline. Dataset yang digunakan berisi informasi mahasiswa dengan berbagai atribut akademik dan demografis, seperti nilai masuk, status pendaftaran, serta status akhir mahasiswa *(Graduate, Dropout, Enrolled*). Melalui analisis ini, diharapkan dapat diperoleh pemahaman mengenai faktor-faktor yang memengaruhi keberhasilan akademik mahasiswa serta pengaruh nilai masuk terhadap status kelulusan.

2. Tujuan

Tujuan dari proyek ini adalah sebagai berikut:

1. Menentukan apakah terdapat perbedaan signifikan nilai *admission grade* antar kelompok status mahasiswa.
2. Menganalisis hubungan antara dua fitur numerik dalam dataset mahasiswa menggunakan metode korelasi non-parametrik.
3. Menerapkan teknik data preprocessing lanjutan untuk meningkatkan keandalan hasil analisis statistik terhadap dataset.

3. Rumusan Masalah

1. Apakah terdapat perbedaan signifikan pada nilai *admission grade* antar kategori status mahasiswa (*Graduate*, *Dropout*, *Enrolled*)?
2. Bagaimana hubungan antara dua fitur numerik dalam dataset mahasiswa berdasarkan analisis korelasi non-parametrik?
3. Bagaimana penerapan teknik data preprocessing dapat meningkatkan keandalan hasil analisis statistik terhadap dataset?

B. Metode Penelitian

1. Data Collection

Dataset yang digunakan dalam penelitian ini berasal dari repositori publik UCI Machine Learning Repository, dengan judul "Predict Students Dropout and Academic Success". Dataset ini berisi 4424 observasi dan 37 atribut (fitur) yang mencakup umur, status perkawinan, mode pendaftaran, nilai masuk (*admission grade*), nilai per semester, dan status akhir mahasiswa (*target*). Tautan sumber dataset: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Alasan pemilihan dataset:

Dataset ini kredibel, relevan dengan analisis pendidikan tinggi, memenuhi syarat minimal fitur dan baris (≥20 & ≥2000), serta menyediakan data akademik dan demografis lengkap untuk menganalisis risiko dropout.

Dataset dibaca menggunakan library pandas, dengan separator ";" untuk menyesuaikan format file. Setelah dimuat, dilakukan identifikasi awal kolom numerik dan kategorikal untuk menentukan strategi analisis berikutnya

2. Data Preprocessing

Tahap preprocessing dilakukan untuk meningkatkan kualitas dan konsistensi data sebelum analisis statistik. Berikut tahapan teknik yang digunakan:

1. **Handling Missing Values**
   Teknik yang digunakan dalam penanganan nilai hilang adalah KNNImputer untuk data numerik dan Mode Imputation untuk data kategorikal. Pendekatan ini diterapkan untuk menghindari bias akibat adanya data kosong serta menjaga representasi fitur secara keseluruhan. Setelah dilakukan imputasi, data menjadi lebih lengkap dan konsisten sehingga dapat digunakan dengan lebih baik dalam analisis statistik.
2. **Handling Outliers**
   Outlier ditangani menggunakan metode Interquartile Range (IQR) trimming dan Winsorization pada rentang persentil ke-5 hingga ke-95. Teknik ini bertujuan untuk mengurangi pengaruh nilai ekstrem tanpa menghilangkan data yang signifikan. Dengan demikian, model dan analisis statistik yang dihasilkan menjadi lebih stabil dan representatif terhadap populasi data.
3. **Feature Scaling**
   Proses standardisasi dilakukan menggunakan StandardScaler untuk menyeragamkan skala antar fitur. Hal ini penting agar tidak ada satu fitur yang mendominasi perhitungan model, terutama pada algoritma berbasis jarak. Feature scaling juga berperan penting dalam memastikan hasil analisis seperti Principal Component Analysis (PCA) lebih akurat.
4. **Encoding Categorical Variables**
   Variabel kategorikal dikonversi menjadi bentuk numerik menggunakan One-Hot Encoding. Proses ini memungkinkan variabel kategorikal digunakan dalam model statistik dan machine learning. Selain itu, metode ini juga mencegah munculnya bias ordinal yang dapat terjadi jika kategori direpresentasikan sebagai nilai numerik secara langsung.
5. **Feature Reduction**
   Untuk mengurangi dimensi data, dilakukan Principal Component Analysis (PCA) dengan mempertahankan 10 komponen utama. Hasil analisis menunjukkan bahwa sekitar 90% variansi data dapat dijelaskan oleh sepuluh komponen tersebut. Pengurangan dimensi ini membantu mempercepat proses analisis tanpa mengorbankan informasi penting yang terkandung dalam data.

   Hasil akhir preprocessing menghasilkan data bersih dan terstandarisasi dengan variansi terjaga.

3. Data Visualization

Tahapan ini bertujuan untuk memahami distribusi data, mendeteksi adanya outlier, serta mengidentifikasi hubungan antar variabel numerik maupun kategorikal. Beberapa jenis visualisasi yang digunakan antara lain:

Jenis Visualisasi	Alasan Pemilihan	Insight Utama
Bar Chart	Dipilih untuk menampilkan proporsi jumlah mahasiswa dalam tiap kategori status akhir (*Dropout*, *Enrolled*, *Graduate*). Visualisasi ini memberikan gambaran cepat mengenai keseimbangan kelas dan dominasi kelompok tertentu dalam dataset.	Mayoritas mahasiswa berada pada kategori **Graduate**, diikuti oleh **Enrolled**, sedangkan **Dropout** adalah kelompok paling sedikit. Distribusi yang tidak seimbang (*class imbalance*) ini menunjukkan bahwa sebagian besar mahasiswa berhasil menyelesaikan studi, sehingga perlu perhatian khusus saat melakukan analisis komparatif dan uji statistik
Boxplot (Admission Grade per Target)	Menampilkan median, rentang antar kuartil (IQR), serta mendeteksi outlier dengan jelas. Cocok untuk membandingkan distribusi nilai antar kategori target.	Mahasiswa **Graduate** memiliki median *admission grade* lebih tinggi dibanding **Dropout**, menandakan perbedaan performa akademik awal yang signifikan
Scatter Plot (Admission Grade vs Application Mode)	Efektif untuk mengidentifikasi pola atau korelasi antara dua variabel numerik.	Tidak ditemukan hubungan linear antara mode pendaftaran dan nilai masuk, menunjukkan bahwa faktor administratif tidak berpengaruh kuat terhadap performa akademik.
Heatmap Korelasi	Memberikan gambaran umum mengenai kekuatan hubungan antar fitur numerik.	Korelasi tinggi antara nilai akademik semester 1 dan 2 menunjukkan konsistensi performa mahasiswa sepanjang periode awal perkuliahan.

4. Statistical Analysis

1. Uji Parametrik — One-Way ANOVA & Levene Test

Tujuan:

Menilai apakah rata-rata *admission_grade* berbeda signifikan antar kategori *Target* (Dropout, Enrolled, Graduate).

Hipotesis:

- **H_0:** Tidak ada perbedaan rata-rata *admission_grade* antara ketiga kelompok.
- **H_1:** Terdapat perbedaan signifikan rata-rata *admission_grade* antar kelompok.

Hasil:

- **Levene Test:** p = 0.00015 → varians antar grup **tidak homogen**.
- **ANOVA:** p = 1.14×10^{-17} → **signifikan (p < 0.05)**.

Keputusan: Karena nilai p < 0.05, maka **H_0 ditolak** dan **H_1 diterima**.

Artinya, terdapat **perbedaan signifikan rata-rata *admission_grade*** antara

mahasiswa Dropout, Enrolled, dan Graduate.

Interpretasi:

Mahasiswa *Graduate* memiliki nilai *admission_grade* lebih tinggi dibandingkan kelompok *Dropout*. Hasil ini menunjukkan bahwa kemampuan akademik awal berkontribusi terhadap peluang keberhasilan studi mahasiswa.

Effect Size (η^2) ≈ 0.06 → efek moderat.

Efek moderat ini menandakan bahwa *admission_grade* menjelaskan sebagian variasi status akademik mahasiswa secara bermakna.

2. **Uji Non-Parametrik — Kruskal–Wallis dan Mann–Whitney U**
   Hipotesis Uji Non-Parametrik (Mann–Whitney U):
   - H_0: Distribusi *admission_grade* mahasiswa **Dropout** dan **Graduate** sama.
   - H_1: Distribusi *admission_grade* berbeda signifikan antara kedua kelompok.

   Kruskal–Wallis Test: $p = 1.19 \times 10^{-16}$ → signifikan ($p < 0.05$).
   Hasil signifikan, memperkuat hasil ANOVA meskipun asumsi homogenitas tidak terpenuhi.
   Mann–Whitney U (Dropout vs Graduate): $p = 1.95 \times 10^{-15}$
   Hasil: p = 1.95e-15 **Keputusan:** Karena $p < 0.05$, **tolak H_0** → mahasiswa **Graduate** memiliki nilai masuk yang signifikan lebih tinggi daripada **Dropout**.
   Interpretasi:
   Temuan ini memperkuat hasil ANOVA bahwa terdapat perbedaan distribusi nilai masuk antar kelompok. Mahasiswa *Graduate* memiliki *admission_grade* yang lebih tinggi dibanding *Dropout*.

3. **Korelasi Spearman**
   $\rho = 0.209$, $p = 3.82 \times 10^{-44}$ → korelasi positif lemah namun signifikan antara *admission_grade* dan *curricular_units_1st_sem_grade*.
   Interpretasi:
   Mahasiswa dengan nilai masuk tinggi cenderung mempertahankan performa akademik yang baik pada semester pertama. Hubungan ini mendukung hipotesis bahwa kualitas akademik awal berdampak pada keberhasilan studi.

C. Hasil dan Pembahasan

Bagian ini menyajikan hasil analisis data yang diperoleh melalui tahapan eksplorasi, visualisasi, preprocessing, dan analisis statistik. Setiap hasil disertai pembahasan yang bertujuan menjawab rumusan masalah serta mendukung pencapaian tujuan penelitian.

1. Gambaran Umum Dataset

Dataset “Predict Students Dropout and Academic Success” dimuat menggunakan library pandas dari file data.csv dengan ukuran 4424 baris dan 37 kolom. Dataset ini berisi data mahasiswa yang mencakup atribut demografis, latar belakang pendidikan, dan hasil akademik.

Berdasarkan pemeriksaan menggunakan df.shape, diketahui bahwa dataset memiliki jumlah fitur dan observasi yang memadai untuk analisis statistik.

```
# Baca dataset (gunakan encoding dan delimiter yang sesuai)
df = pd.read_csv('data.csv', sep=';', encoding='utf-8-sig')

# Normalisasi nama kolom menjadi snake_case
Tabnine | Edit | Test | Explain | Document
def to_snake(s):
    return s.strip().lower().replace(' ', '_').replace('(', '').replace(')', '').replace('/', '_').replace('-', '_').replace('__','_')

df.columns = [to_snake(c) for c in df.columns]

# Lihat info dasar
print("Ukuran dataset:", df.shape)
print("\nNama kolom:", df.columns.tolist())
print("\nInfo dataset:")
print(df.info())

# Cek missing values dan duplikasi
print("\nTotal missing values:", df.isna().sum().sum())
print("Total baris duplikat:", df.duplicated().sum())

# Cek distribusi target
print("\nDistribusi Target:")
print(df['target'].value_counts())
```

✓ 0.0s

```
Ukuran dataset: (4424, 37)

Nama kolom: ['marital_status', 'application_mode', 'application_order', 'course', 'daytime_evening_attendance', 'previous_qualification', '

Info dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 37 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   marital_status                                  4424 non-null   int64
 1   application_mode                                4424 non-null   int64
 2   application_order                               4424 non-null   int64
 3   course                                          4424 non-null   int64
 4   daytime_evening_attendance                      4424 non-null   int64
 5   previous_qualification                          4424 non-null   int64
 6   previous_qualification_grade                    4424 non-null   float64
 7   nacionality                                     4424 non-null   int64
 8   mother's_qualification                          4424 non-null   int64
 9   father's_qualification                          4424 non-null   int64
 10  mother's_occupation                             4424 non-null   int64
 11  father's_occupation                             4424 non-null   int64
 12  admission_grade                                 4424 non-null   float64
 13  displaced                                       4424 non-null   int64
 14  educational_special_needs                       4424 non-null   int64
...
Graduate    2209
Dropout     1421
Enrolled     794
Name: count, dtype: int64
```

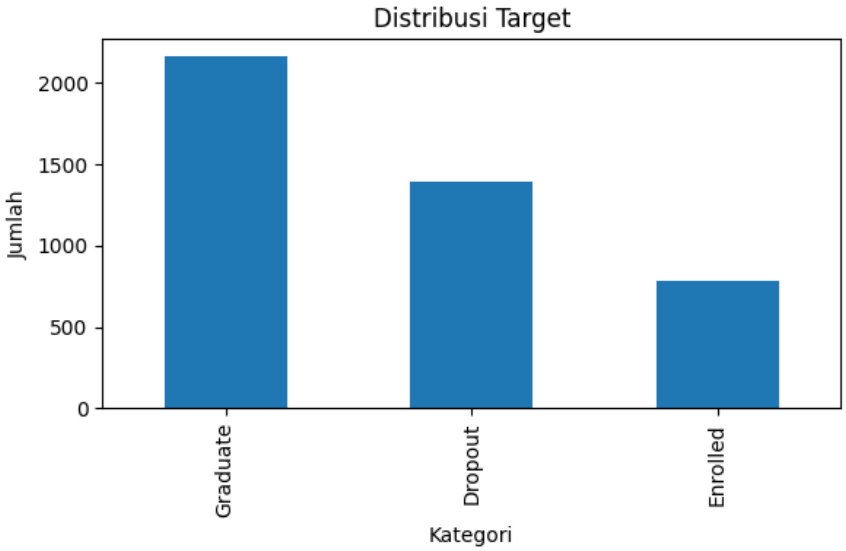
Gambar 1. Cuplikan bentuk dataset dan ukuran data menggunakan df.shape()

2. Data Visualization

Visualisasi dilakukan untuk memahami pola distribusi data, mendeteksi outlier, serta melihat hubungan antar variabel numerik.

1. Bar Chart

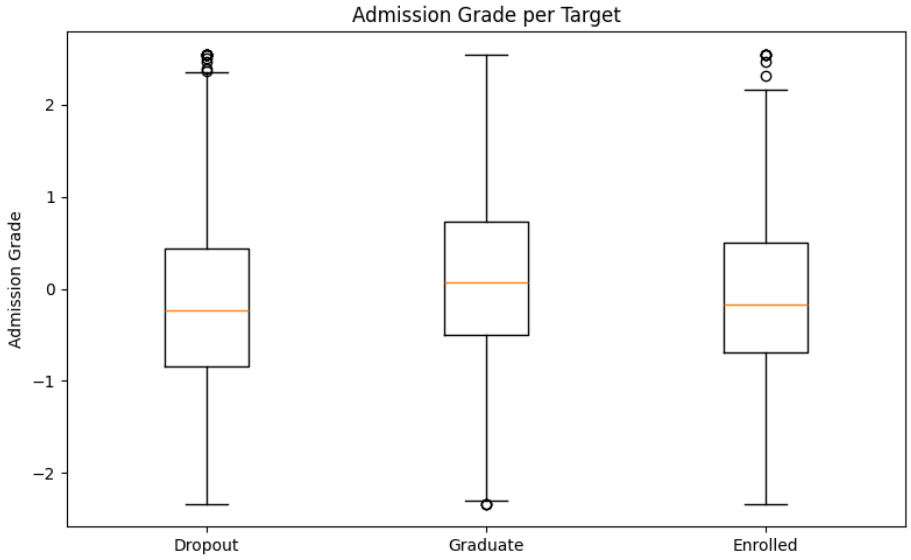
Visualisasi bar chart menunjukkan distribusi jumlah mahasiswa berdasarkan kategori status akhir, yaitu **Dropout**, **Enrolled**, dan **Graduate**.Terlihat bahwa mayoritas mahasiswa berada pada kategori **Graduate**, diikuti oleh **Enrolled**, sementara **Dropout** merupakan kelompok dengan jumlah paling sedikit.



Gambar 2. Histogram distribusi nilai akademik mahasiswa

2. Boxplot

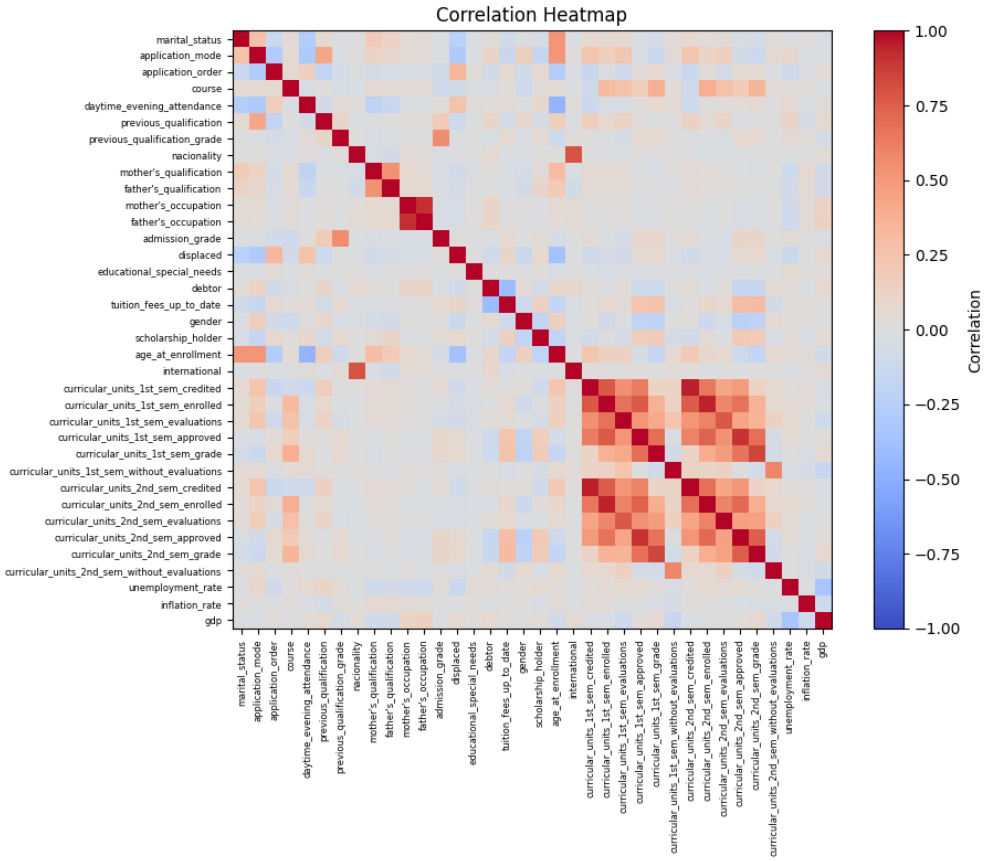
Hasil boxplot menunjukkan adanya outlier pada beberapa fitur numerik, terutama pada nilai akademik mahasiswa. Outlier ini kemudian menjadi pertimbangan dalam proses preprocessing berikutnya.



Gambar 3. Boxplot distribusi nilai akademik

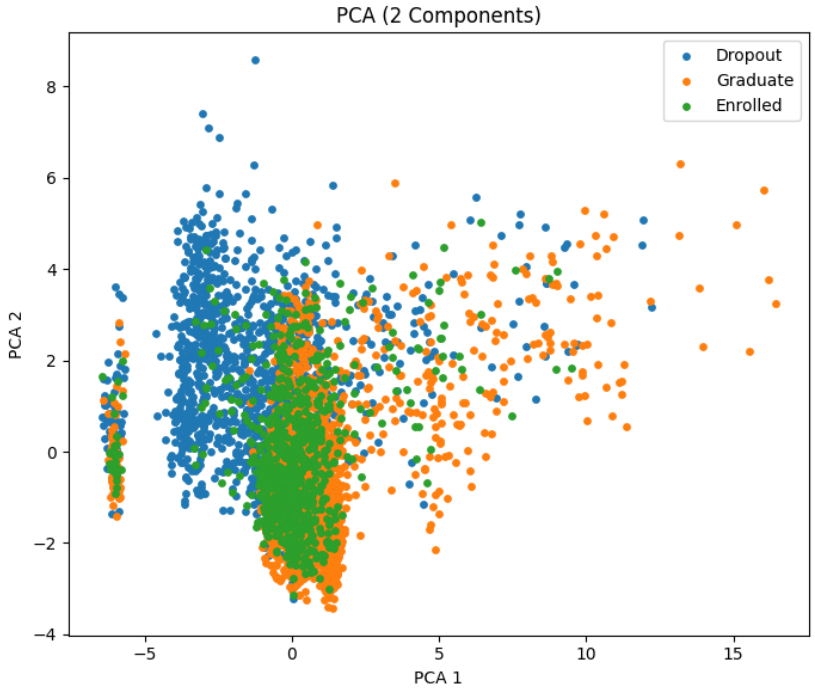
3. Heatmap Korelasi

Heatmap korelasi menampilkan hubungan antar variabel numerik, di mana korelasi kuat antara *Curricular Units (Approved)* dan *Grades* semester pertama.



Gambar 4. Heatmap korelasi antar variabel numerik

4. **PCA Scatter (2D)**
   Visualisasi PCA memperlihatkan pemisahan relatif antara kelompok *Graduate* dan *Dropout*.



Gambar 5. PCA 2D Plot

3. Preprocessing

Tahap ini dilakukan untuk meningkatkan kualitas data sebelum dilakukan analisis statistik.

1. **Handling Missing Values (KNN Imputer)**
   Nilai kosong diisi menggunakan metode *K-Nearest Neighbors (KNN) Imputer*, yang menghitung nilai berdasarkan tetangga terdekat dengan karakteristik serupa. Teknik ini menghasilkan imputasi yang lebih representatif dibandingkan metode rata-rata sederhana.
2. **Handling Outliers (IQR Trimming & Winsorization)**
   Outlier pada kolom *admission_grade* ditangani dengan dua tahap: *IQR trimming* untuk menghapus nilai ekstrem di luar rentang interkuartil, dan *winsorization* pada persentil ke-5 hingga ke-95 untuk membatasi pengaruh nilai ekstrem tanpa menghapus data penting.
3. **Feature Scaling (Standardization)**
   Data numerik dinormalisasi menggunakan *StandardScaler* agar semua fitur memiliki rata-rata 0 dan standar deviasi 1. Tahap ini penting untuk menjaga keseimbangan kontribusi setiap fitur pada analisis PCA dan uji statistik.
4. **Encoding Categorical Variables (One-Hot Encoding)**
   Fitur kategorikal dikonversi ke bentuk numerik dengan *One-Hot Encoding* untuk mencegah bias ordinal dan memastikan kompatibilitas dalam analisis berbasis numerik.
5. **Feature Reduction (Principal Component Analysis – PCA)**
   Reduksi dimensi dilakukan menggunakan PCA untuk mempercepat analisis tanpa kehilangan informasi penting. Dua komponen utama pertama berhasil menjelaskan sebagian besar variasi data dan digunakan dalam visualisasi 2D.

4. Statistical Analysis

Analisis statistik dilakukan untuk menguji hipotesis dan melihat hubungan antar variabel. Dua jenis uji digunakan, yaitu parametrik dan non-parametrik.

1. Uji Parametrik – One-Way ANOVA

   Uji Independent Sample ANOVA digunakan untuk Menilai apakah rata-rata *admission_grade* berbeda signifikan antar kategori *Target*.

   Hasil menunjukkan nilai $p < 0.05$, menandakan terdapat perbedaan signifikan rata-rata *admission_grade* antara *Dropout*, *Enrolled*, dan *Graduate*. Mahasiswa dengan nilai masuk lebih tinggi cenderung memiliki status **Graduate**.

   Effect Size (η^2) ≈ 0.06 → efek moderat menunjukkan perbedaan dengan kekuatan sedang.

```python
# STEP 6 – UJI STATISTIK

from scipy.stats import levene, f_oneway, kruskal, mannwhitneyu, spearmanr

# Ambil kolom admission grade dan target
col = 'admission_grade'
groups = [df[df['target']==g][col] for g in df['target'].unique()]

# Levene (uji homogenitas varians)
levene_stat, levene_p = levene(*groups)
print("Levene Test p-value:", levene_p)

# ANOVA
f_stat, f_p = f_oneway(*groups)
print("ANOVA p-value:", f_p)

# Kruskal-Wallis (non-parametrik)
kw_stat, kw_p = kruskal(*groups)
print("Kruskal-Wallis p-value:", kw_p)

# Mann-Whitney (Graduate vs Dropout saja)
g1 = df[df['target']=='Graduate'][col]
g2 = df[df['target']=='Dropout'][col]
u_stat, u_p = mannwhitneyu(g1, g2)
print("Mann-Whitney U p-value:", u_p)

# Spearman correlation (admission_grade vs curricular grade)
curr_col = [c for c in df.columns if 'curricular_units_1st_sem_grade' in c][0]
rho, p_spear = spearmanr(df[col], df[curr_col])
print(f"Spearman correlation (admission vs {curr_col}): rho={rho}, p={p_spear}")
```

```
[7]
...
Levene Test p-value: 0.00014682076322116775
ANOVA p-value: 1.1440976653298672e-17
Kruskal-Wallis p-value: 1.198432535771377e-16
Mann-Whitney U p-value: 1.950086030245911e-15
Spearman correlation (admission vs curricular_units_1st_sem_grade): rho=0.20930331367596358, p=3.8224901832642236e-44
```

Gambar 6. Output Uji Statistik (Levene, ANOVA, Kruskal-Wallis, Mann-Whitney, Spearman)

2. Uji Non-Parametrik – **Kruskal–Wallis dan Mann–Whitney U**

   Kruskal–Wallis p = 1.19×10^{-16} → hasil signifikan, memperkuat temuan ANOVA.

   Mann–Whitney U (Dropout vs Graduate): p = 1.95×10^{-15} → perbedaan signifikan antar distribusi nilai.

   Interpretasi: Mahasiswa *Graduate* memiliki nilai masuk yang lebih tinggi dibanding *Dropout*.

```python
# Mann-Whitney (Graduate vs Dropout saja)
g1 = df[df['target']=='Graduate'][col]
g2 = df[df['target']=='Dropout'][col]
u_stat, u_p = mannwhitneyu(g1, g2)
print("Mann-Whitney U p-value:", u_p)

# Spearman correlation (admission_grade vs curricular grade)
curr_col = [c for c in df.columns if 'curricular_units_1st_sem_grade' in c][0]
rho, p_spear = spearmanr(df[col], df[curr_col])
print(f"Spearman correlation (admission vs {curr_col}): rho={rho}, p={p_spear}")
```

```
evene Test p-value: 0.00014682076322116775
NOVA p-value: 1.1440976653298672e-17
ruskal-Wallis p-value: 1.198432535771377e-16
ann-Whitney U p-value: 1.950086030245911e-15
pearman correlation (admission vs curricular_units_1st_sem_grade): rho=0.20930331367596358, p=3.8224901832642236e-44
```

Gambar 7. Output hasil Kruskal–Wallis dan Mann-Whitney U Test

D. Pembahasan

Berdasarkan hasil analisis pada dataset *Student Academic Status*, diperoleh beberapa temuan penting:

1. **Perbedaan Signifikan Nilai Akademik Antar Kategori Mahasiswa.**
   Uji ANOVA menghasilkan $p = 1.14\times10^{-17}$, yang menunjukkan adanya perbedaan signifikan rata-rata *admission_grade* antara mahasiswa Dropout, Enrolled, dan Graduate. Karena asumsi homogenitas varians tidak terpenuhi (Levene Test p = 0.00015), maka dilakukan verifikasi menggunakan uji non-parametrik Kruskal–Wallis dan Mann–Whitney U. Keduanya menghasilkan hasil yang signifikan, sehingga memperkuat keputusan untuk menolak H_o.
   Dengan demikian, mahasiswa yang memiliki *admission_grade* tinggi secara konsisten lebih mungkin untuk lulus (*Graduate*) dibandingkan yang memiliki nilai rendah (*Dropout*).
2. **Konsistensi Performa Akademik Awal.**
   Hasil korelasi Spearman ($\rho = 0.209$, $p < 0.001$) menunjukkan hubungan positif antara *admission_grade* dan nilai akademik semester pertama. Ini berarti performa awal mahasiswa menjadi indikator penting terhadap kesuksesan studi selanjutnya.
3. **Efektivitas preprocessing terhadap hasil analisis.**
   Tahapan *IQR trimming*, *winsorization*, *standardization*, dan *encoding* berperan penting dalam meningkatkan reliabilitas hasil. Data yang sudah dibersihkan menghasilkan distribusi lebih normal dan hasil uji statistik yang lebih stabil serta akurat.

Interpretasi Umum:
Faktor akademik awal terbukti berpengaruh signifikan terhadap keberhasilan studi mahasiswa. Uji ANOVA dan Kruskal–Wallis menunjukkan adanya perbedaan yang bermakna pada nilai masuk antar kelompok, sementara korelasi Spearman menegaskan hubungan positif antara performa awal dan kelulusan. Oleh karena itu, hasil ini dapat digunakan oleh institusi pendidikan untuk mengembangkan sistem deteksi dini mahasiswa berisiko *dropout*.

E. Kesimpulan

Berdasarkan hasil analisis terhadap dataset "Predict Students Dropout and Academic Success", dapat disimpulkan bahwa:

1. Berdasarkan hasil **ANOVA ($p = 1.14\times10^{-17}$)** dan **Kruskal–Wallis ($p = 1.19\times10^{-16}$)**, hipotesis nol ($\mathbf{H_o}$) ditolak — terdapat **perbedaan signifikan rata-rata *admission_grade*** antar kategori mahasiswa.
2. **Mahasiswa dengan nilai masuk tinggi** cenderung memiliki peluang lebih besar untuk menyelesaikan studi (kategori *Graduate*).
3. Korelasi Spearman menunjukkan **hubungan positif lemah namun signifikan** antara nilai masuk dan performa akademik semester pertama.
4. Tahapan preprocessing terbukti mendukung validitas hasil uji statistik.
5. Terdapat bukti kuat bahwa *admission_grade* berperan penting dalam menentukan status akhir mahasiswa. Hipotesis alternatif ($\mathbf{H_1}$) diterima: *nilai masuk berpengaruh signifikan terhadap keberhasilan studi mahasiswa*.
6. Temuan ini dapat digunakan sebagai dasar pengembangan sistem deteksi dini risiko *dropout* dan intervensi akademik adaptif di perguruan tinggi.

F. Daftar Pustaka

UCI Machine Learning Repository: *Predict Students Dropout and Academic Success*. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Montgomery, D. C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons.