Data Science

Analisis Data Student Academic Status

Kelompok - 26

41425078 Daniel Siahaan
41425079 Jessica Pasaribu
41425080 Novrael Gabriel Louis Marbun

https://github.com/nielshn/UTS-Prak-41425078_41425079_41425080

Pendahuluan

Data Science berfokus pada pengolahan dan analisis data untuk menghasilkan wawasan dalam pengambilan keputusan.

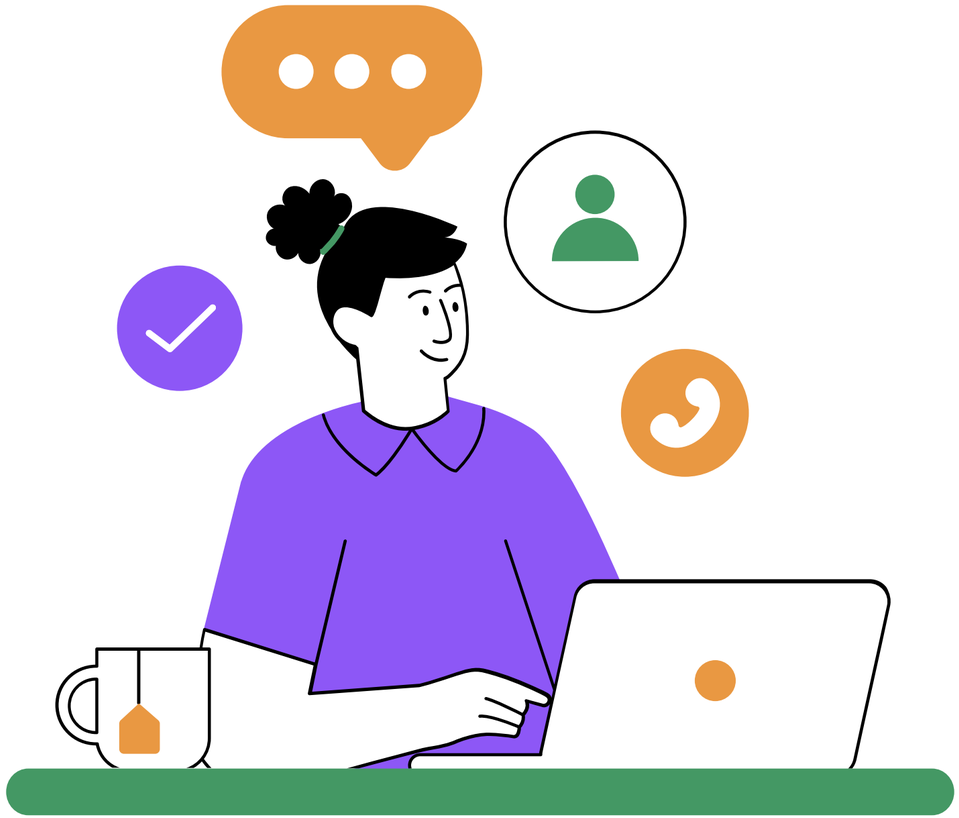
Dalam konteks pendidikan, analisis data mahasiswa membantu memahami faktor-faktor yang memengaruhi keberhasilan akademik, seperti nilai masuk dan performa studi.

Proyek ini menerapkan tahapan data science pipeline untuk menganalisis dataset mahasiswa dan melihat pengaruh nilai masuk terhadap status kelulusan.

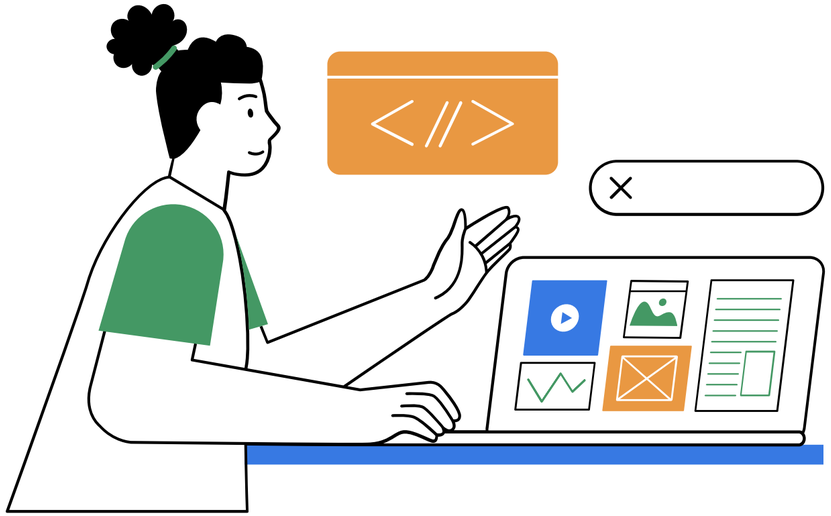
Tujuan Penelitian

- Menentukan apakah terdapat perbedaan signifikan nilai admission grade antar status mahasiswa.
- Menganalisis hubungan antar fitur numerik dengan metode korelasi non-parametrik.
- Menerapkan teknik data preprocessing lanjutan untuk meningkatkan keandalan hasil analisis statistik.

Rumusan Masalah



1. Apakah terdapat perbedaan signifikan nilai admission grade antar kategori status mahasiswa?
2. Bagaimana hubungan antar dua fitur numerik berdasarkan analisis korelasi non-parametrik?
3. Bagaimana teknik data preprocessing meningkatkan keandalan analisis?



Dataset: Predict Students Dropout and Academic Success

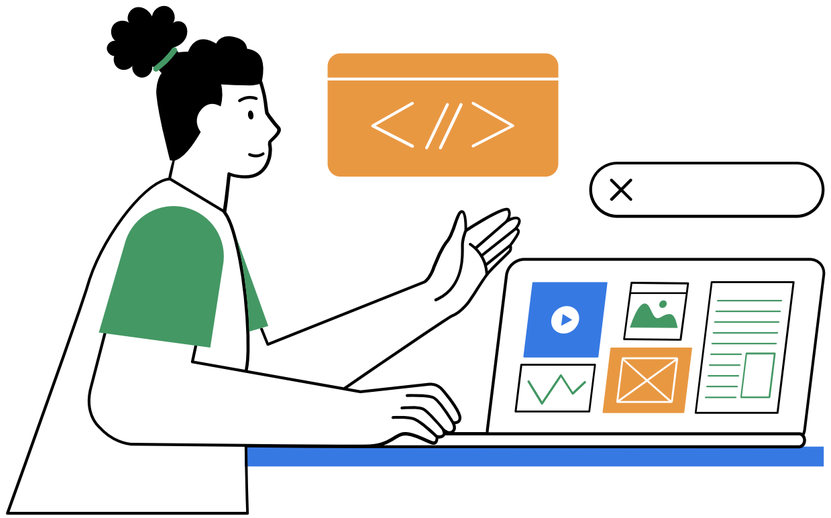
Sumber: UCI Machine Learning Repository

- **4424 observasi, 37 atribut**
- **Berisi data demografis, latar belakang pendidikan, dan performa akademik**
- **Format: CSV dengan separator ;**
- **Alasan pemilihan dataset: Kredibel, relevan, memenuhi syarat analisis, dan lengkap.**

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Data Collection

Data Preprocessing

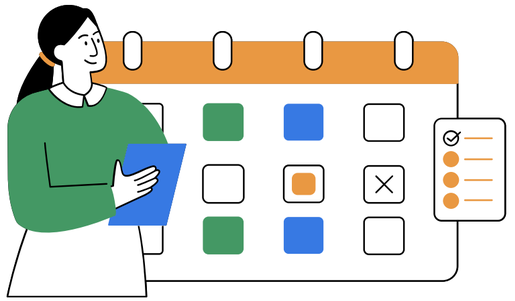


Tahapan dilakukan untuk meningkatkan kualitas dan konsistensi data sebelum analisis.

Langkah-langkah:

- **Handling Missing Values**
- **Handling Outliers**
- **Feature Scaling**
- **Encoding Categorical Variables**
- **Feature Reduction**

Handling Missing Values



Metode

KNNImputer(numerik) & Mode Imputation(kategorikal)

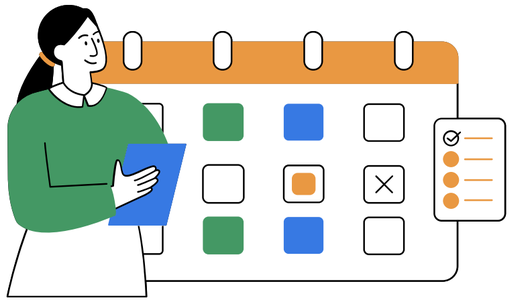
Tujuan

Menghindari bias akibat data kosong dan menjaga representasi fitur

Hasil

Data lebih lengkap dan konsisten untuk analisis statistik

Handling Outliers



Metode

IQR Trimming & Winsorization (5th–95th percentile)

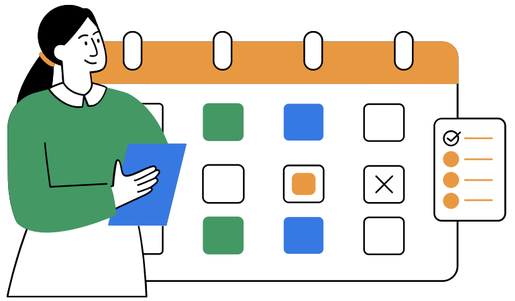
Tujuan

Mengurangi pengaruh nilai ekstrem tanpa kehilangan data penting.

Hasil

Model menjadi lebih stabil dan representatif terhadap populasi data.

Feature Scaling & Encoding



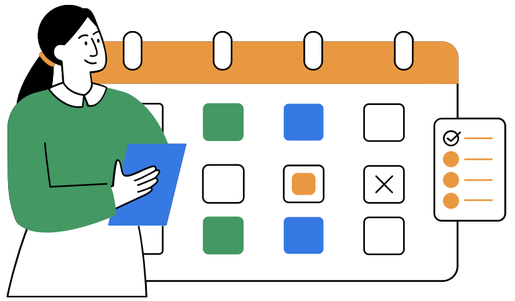
StandardScaler:

Menyeragamkan skala antar fitur agar setara dalam analisis berbasis jarak.

One-Hot Encoding

Mengubah variabel kategorikal menjadi numerik tanpa bias ordinal.

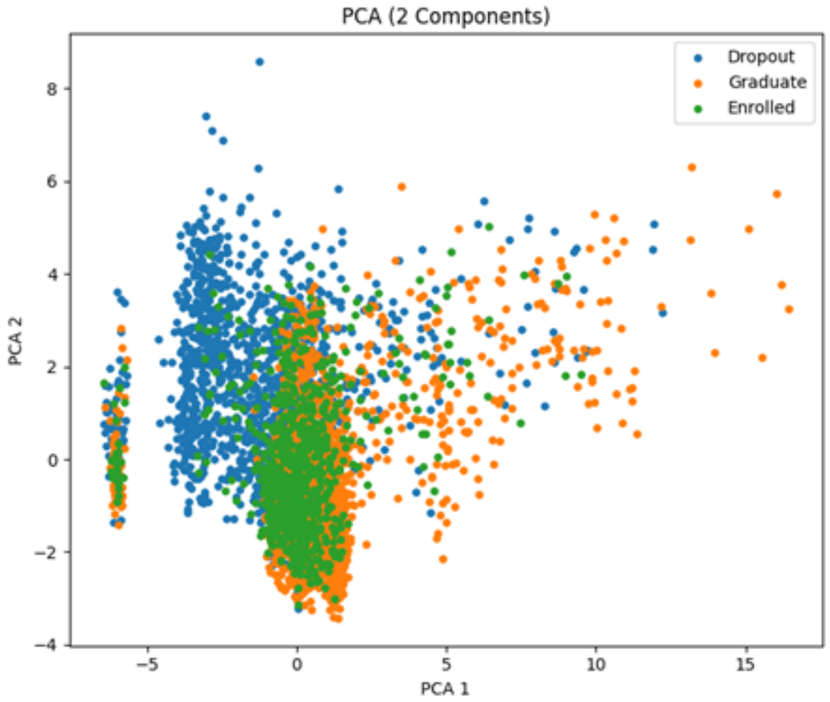
Feature Reduction



Metode : Principal Component Analysis (PCA)

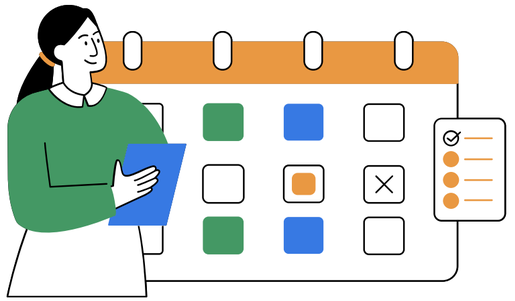
- Menyisakan 10 komponen utama
- Menjelaskan ±90% variansi data

Manfaat: Mengurangi dimensi tanpa kehilangan informasi penting.



Gambar : PCA 2D Plot

Data Visualization



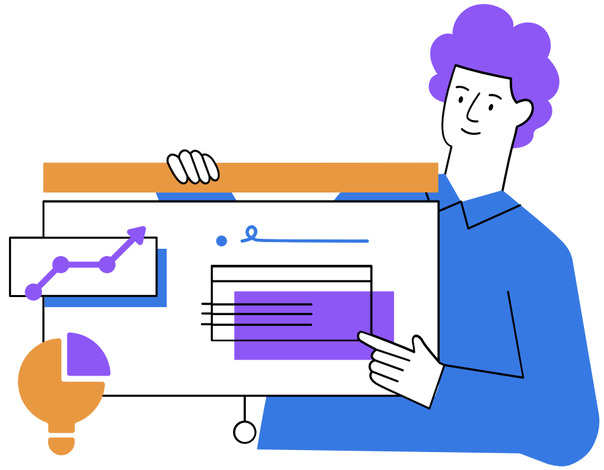
Tujuan:

- Memahami distribusi data
- Mendeteksi outlier
- Mengidentifikasi hubungan antar variabel

Jenis visualisasi yang digunakan:

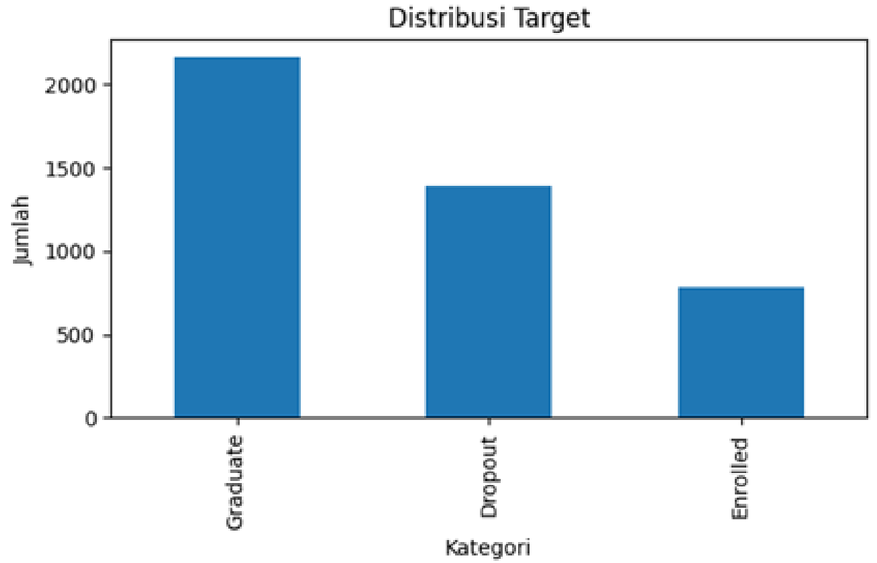
- Bar Chart
- Boxplot
- Scatter Plot
- Heatmap Korelasi

Bar Chart



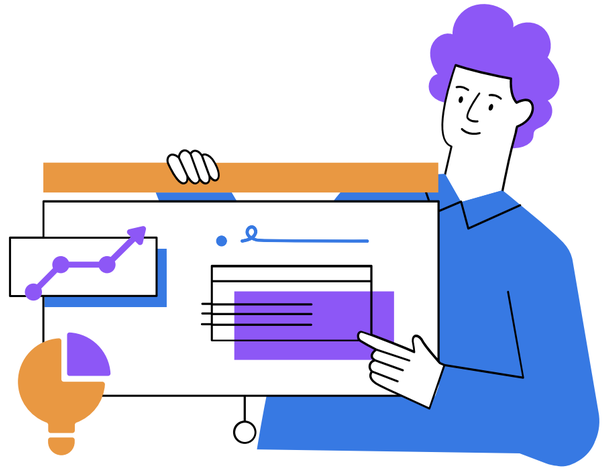
Menampilkan proporsi mahasiswa berdasarkan status akhir (Dropout, Enrolled, Graduate).

Insight: Mayoritas mahasiswa berstatus Graduate, disusul Enrolled, dan paling sedikit

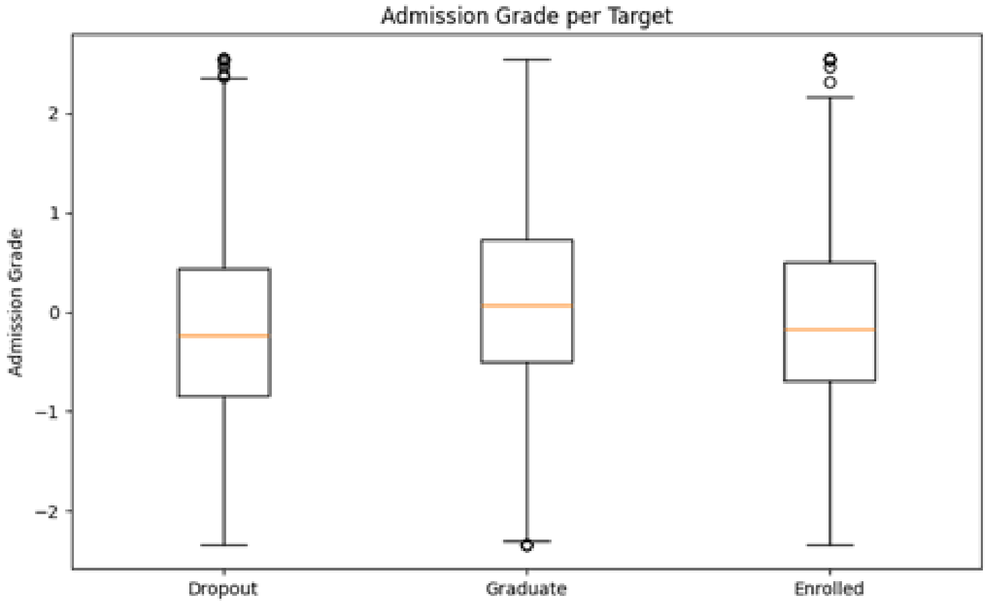


Gambar Bar Chart distribusi nilai akademik mahasiswa

Boxplot

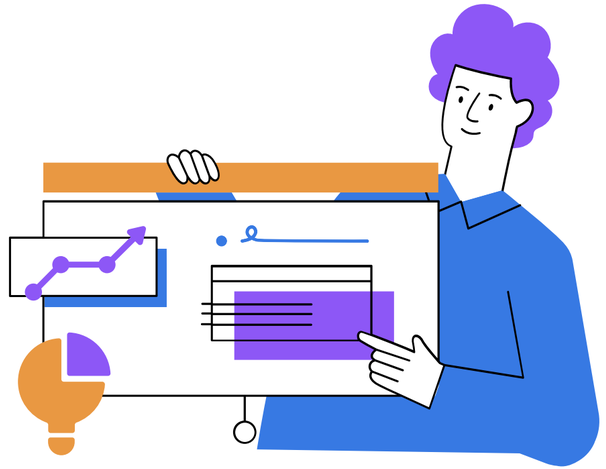


Menampilkan median, IQR, dan deteksi outlier antar kategori.
Insight: Mahasiswa Graduate memiliki median admission grade lebih tinggi dibanding Dropout.



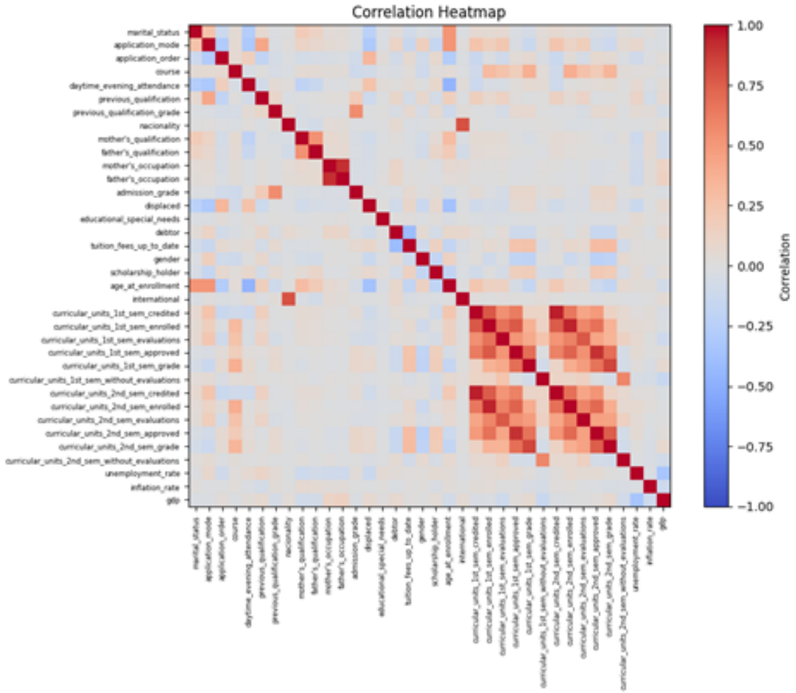
Gambar Boxplot distribusi nilai akademik

Heatmap Korelasi



Menunjukkan kekuatan hubungan antar variabel numerik.

Insight: Nilai akademik semester 1 dan 2 memiliki korelasi tinggi → konsistensi performa mahasiswa.



Gambar Heatmap korelasi antar variabel numerik

Uji Statistik Parametrik (ANOVA)

Tujuan:

Menilai perbedaan rata-rata admission_grade antar status mahasiswa.

Hasil:

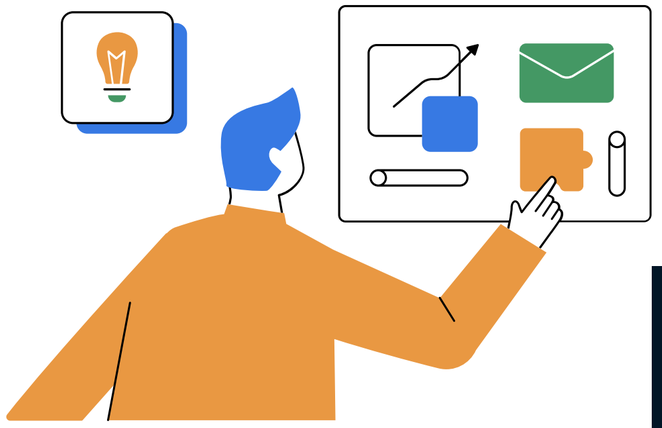
- Levene Test p = 0.00015 → Varians tidak homogen
- ANOVA p = 1.14×10^{-17} → Signifikan (p < 0.05)

Interpretasi: Ada perbedaan signifikan antar grup (Dropout, Enrolled, Graduate).
Effect Size (η^2) = 0.06 → Efek moderat.

```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}")
```

```
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
```

Uji Non-Parametrik



Metode: Kruskal–Wallis & Mann–Whitney U Test

- Kruskal–Wallis $p = 1.19 \times 10^{-16}$ → signifikan
- Mann–Whitney U $p = 1.95 \times 10^{-15}$ → signifikan

Interpretasi: Mahasiswa Graduate memiliki admission grade lebih tinggi dibanding Dropout.

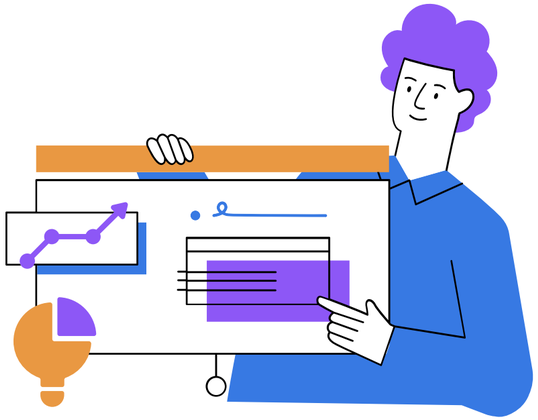
Korelasi Spearman

$\rho = 0.209$, $p = 3.82 \times 10^{-44}$ → korelasi positif lemah namun signifikan.
Artinya, mahasiswa dengan nilai masuk tinggi cenderung mempertahankan performa baik di semester awal.

```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}")
```

```
Levene Test p-value: 0.0001468207632211677 5
ANOVA p-value: 1.1440976653298672e-17
Kruskal-Wallis p-value: 1.198432535771377e-16
Mann-Whitney U p-value: 1.9500860302459 11e-15
Spearman correlation (admission vs curricular_units_1st_sem_grade): rho=0.20930331367596358, p=3.8224901832642236e-44
```

Pembahasan



1. Perbedaan signifikan antar kategori mahasiswa

Nilai $p = 1.14 \times 10^{-17}$ menunjukkan adanya perbedaan signifikan rata-rata admission grade antar kategori.

Mahasiswa dengan nilai masuk tinggi cenderung lulus.

2. Konsistensi performa akademik

Korelasi Spearman menunjukkan hubungan positif antara admission grade dan nilai semester pertama.

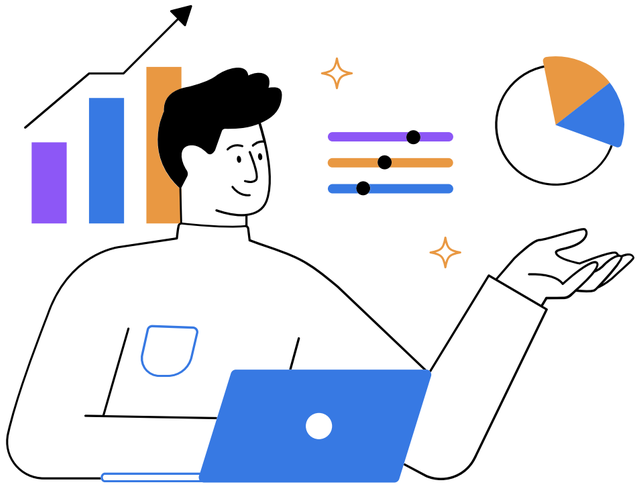
3. Efektivitas preprocessing

Metode trimming, winsorization, scaling, dan encoding membuat hasil analisis lebih stabil dan mudah diinterpretasikan.

Interpretasi:

Faktor nilai akademik awal berpengaruh signifikan terhadap keberhasilan studi mahasiswa.

Kesimpulan



1. Dataset memenuhi syarat kualitas untuk analisis statistik.
2. Nilai masuk (admission grade) berpengaruh signifikan terhadap status akhir mahasiswa.
3. Hasil ANOVA dan Kruskal–Wallis menunjukkan perbedaan nyata antar kategori.
4. Korelasi Spearman membuktikan hubungan positif dengan performa akademik.
5. Temuan ini dapat digunakan untuk sistem deteksi dini risiko dropout.

Thank You