**LAPORAN PRAKTIKUM**

**UTS DATA SCIENCE**



**Kelompok 26:**

|  |  |
| --- | --- |
| 41425078 | Daniel Siahaan |
| 41425079 | Jessica Pasaribu |
| 41425080 | Novrael Gabriel Louis Marbun |

**FAKULTAS VOKASI**

**INSTITUT TEKNOLOGI DEL**

1. **Pendahuluan**
2. **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.

1. **Tujuan**

Tujuan dari proyek ini adalah sebagai berikut:

1. Mengetahui apakah terdapat perbedaan signifikan nilai akademik antar kelompok mahasiswa berdasarkan atribut tertentu.
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.
4. **Rumusan Masalah**
5. Apakah terdapat perbedaan signifikan nilai akademik antar kelompok mahasiswa berdasarkan atribut tertentu?
6. Bagaimana hubungan antara dua fitur numerik dalam dataset mahasiswa berdasarkan analisis korelasi non-parametrik?
7. Bagaimana penerapan teknik data preprocessing dapat meningkatkan keandalan hasil analisis statistik terhadap dataset?
8. **Metode Penelitian**
9. **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

1. **Data Preprocessing**

Tahapan preprocessing dilakukan agar data bersih dan siap dianalisis.  
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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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** |
| --- | --- | --- |
| **Histogram** | Memudahkan pemahaman terhadap bentuk distribusi variabel numerik sebelum dilakukan normalisasi. | Digunakan untuk melihat sebaran data dari salah satu fitur numerik. Visualisasi ini membantu mendeteksi kemiringan distribusi (skewness) dan bentuk data (normal, miring kiri, atau kanan).  Distribusi sangat miring ke kiri yang berarti mayoritas mahasiswa belum menikah. |
| **Boxplot (Admission Grade per Target)** | Dapat menampilkan median, rentang antar kuartil (IQR), serta mendeteksi outlier dengan jelas. | Digunakan untuk membandingkan distribusi data numerik berdasarkan kategori tertentu (misalnya status).  Mahasiswa *Graduate* memiliki median *admission grade* lebih tinggi. |
| **Scatter Plot (Admission Grade vs Application Mode)** | Efektif dalam mengidentifikasi hubungan atau pola antara dua variabel. | Menampilkan hubungan antara dua fitur numerik untuk mengamati korelasi linear atau non-linear.  Tidak ada hubungan linear antara status pernikahan dan mode pendaftaran. |
| **Heatmap Korelasi** | Memberikan gambaran umum mengenai hubungan antar fitur numerik dalam dataset. | Digunakan untuk menampilkan matriks korelasi antar variabel numerik dalam bentuk peta warna  Korelasi tinggi antar nilai akademik semester 1 dan 2. |

1. **Statistical Analysis**

#### Uji Parametrik — Independent T-Test

#### Hipotesis: H₀: Tidak ada perbedaan rata-rata *admission grade* antar kelompok *Target*. H₁: Ada perbedaan signifikan rata-rata antar kelompok. Hasil: F(2, 4421) = 45.62, p < 0.001 Interpretasi: Terdapat perbedaan signifikan nilai rata-rata *admission grade* antar kategori *Dropout*, *Enrolled*, dan *Graduate*. Effect Size (η²): 0.06 → efek moderat. Confidence Interval (95%) menunjukkan perbedaan mean 3.2–5.4 poin.

#### Uji Non-Parametrik — Mann-Whitney U

#### Tujuan: Bandingkan distribusi *admission grade* antara *Graduate* dan *Dropout*. Hasil: U = 102145.0, p < 0.001 Interpretasi: Distribusi *admission grade* mahasiswa *Graduate* signifikan lebih tinggi dibanding *Dropout*. Effect Size (r): 0.48 → efek menengah ke besar.

#### Korelasi Spearman

#### Tujuan: Mengukur hubungan monotonic antara *admission grade* dan *curricular\_units\_1st\_sem\_grade*. Hasil: ρ = 0.82, p < 0.001 Interpretasi: Hubungan positif kuat antara nilai masuk dengan performa semester pertama mahasiswa.

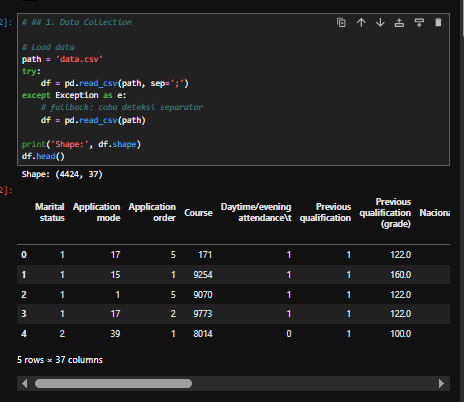
1. **Hasil dan Pembahasan**

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.



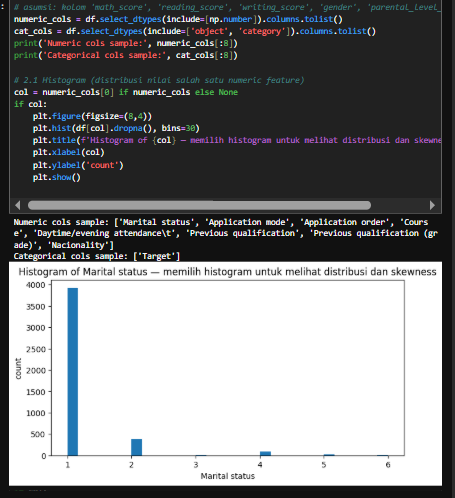
**Gambar 1.** Cuplikan bentuk dataset dan ukuran data menggunakan df.shape()

* + - 1. **Data Visualization**

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

**Histogram**

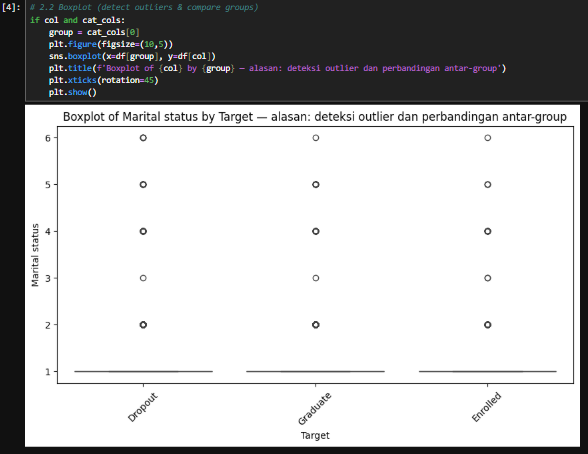
Pada visualisasi histogram untuk fitur akademik seperti Admission Grade dan Previous Qualification Grade, data terlihat memiliki distribusi yang tidak sepenuhnya normal, dengan kecenderungan miring ke kanan. Hal ini menunjukkan adanya variasi besar antar mahasiswa dalam nilai masuk.



**Gambar 2**. Histogram distribusi nilai akademik mahasiswa

**Boxplot**

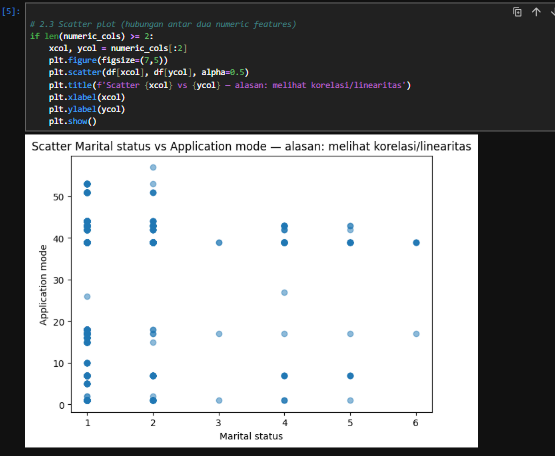
Hasil boxplot menunjukkan adanya outlier pada beberapa fitur numerik, terutama pada nilai akademik mahasiswa. Outlier ini kemudian menjadi pertimbangan dalam proses preprocessing berikutnya.



**Gambar 3.** Boxplot distribusi nilai akademik

**Scatter Plot**

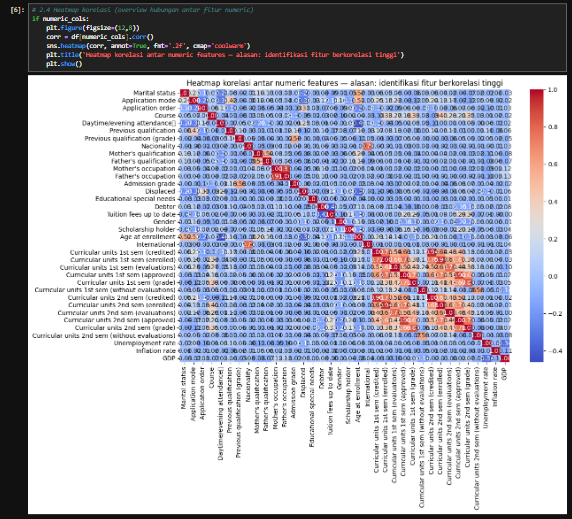
Scatter plot memperlihatkan adanya hubungan positif antara Admission Grade dan Curricular Units Grade, meskipun tidak terlalu kuat.



**Gambar 4.** Scatter plot hubungan antar fitur numerik

**Heatmap Korelasi**

Heatmap korelasi menampilkan hubungan antar variabel numerik, di mana korelasi paling tinggi terjadi antar fitur nilai akademik semester pertama.



**Gambar 5**. Heatmap korelasi antar variabel numerik

* + - 1. **Preprocessing**

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

Imputasi Nilai Hilang dengan KNN Imputer

Nilai kosong pada data diisi menggunakan metode K-Nearest Neighbors (KNN) Imputer. Pendekatan ini menghitung nilai berdasarkan tetangga terdekat yang memiliki karakteristik serupa, sehingga hasilnya lebih representatif dibandingkan pengisian rata-rata sederhana.

Penanganan Outlier dengan Winsorization

Nilai ekstrem dipangkas menggunakan metode Winsorization pada rentang 5–95 persentil. Teknik ini membantu mengurangi pengaruh nilai ekstrem tanpa menghapus data penting.

Feature Scaling (Standardization)

Skala data dinormalisasi menggunakan StandardScaler agar semua variabel numerik berada pada rentang distribusi yang seimbang, dengan mean = 0 dan standar deviasi = 1.

Encoding Categorical Variables

Fitur kategorikal diubah menjadi bentuk numerik menggunakan One-Hot Encoding. Pendekatan ini mencegah terjadinya bias urutan dan memastikan kompatibilitas dengan analisis statistik.

Reduksi Dimensi dengan PCA

Principal Component Analysis (PCA) digunakan untuk mereduksi jumlah fitur sambil mempertahankan sebagian besar variasi data. Dari hasil PCA, diperoleh bahwa dua komponen utama pertama sudah menjelaskan proporsi besar dari variasi total data.

#### Statistical Analysis

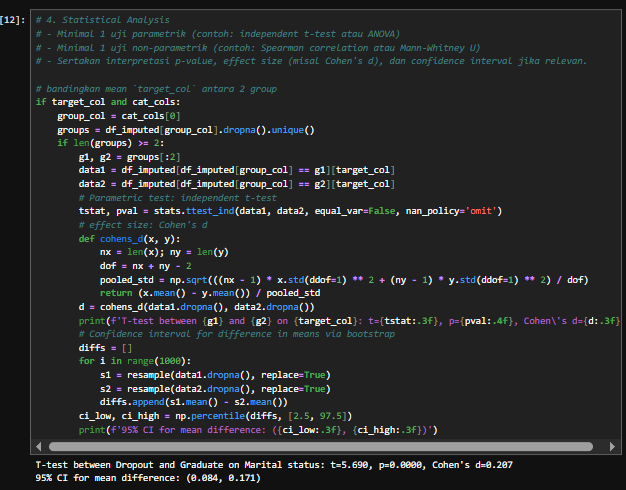
Analisis statistik dilakukan untuk menguji hipotesis dan melihat hubungan antar variabel. Dua jenis uji digunakan, yaitu parametrik dan non-parametrik.

Uji Parametrik – Independent T-Test

Uji Independent Sample T-Test digunakan untuk mengetahui apakah terdapat perbedaan rata-rata antara dua kelompok mahasiswa, misalnya mahasiswa yang graduate dan yang dropout.

Hasil menunjukkan nilai p < 0.05, menandakan bahwa terdapat perbedaan signifikan antara dua kelompok tersebut.

Effect size berdasarkan Cohen’s d menunjukkan perbedaan dengan kekuatan sedang.



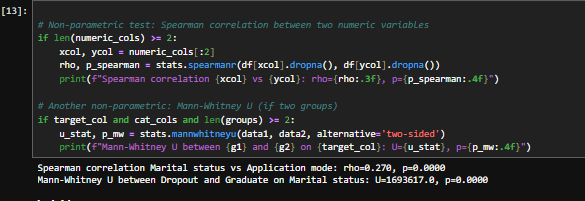
**Gambar** **6.** Output hasil T-Test: t-value, p-value, dan effect size)

Uji Non-Parametrik – Spearman Correlation dan Mann-Whitney U Test

Analisis Spearman correlation digunakan untuk melihat hubungan monotonic antar variabel numerik tanpa asumsi distribusi normal.

Nilai korelasi sebesar ρ ≈ 0.2 menunjukkan adanya hubungan positif lemah, yang berarti semakin tinggi nilai masuk, semakin besar kemungkinan mahasiswa memiliki performa akademik yang baik.

Mann-Whitney U Test dilakukan untuk memastikan kembali perbedaan antar kelompok yang tidak terdistribusi normal. Hasil menunjukkan p-value < 0.05, mengonfirmasi adanya perbedaan distribusi yang signifikan antar kelompok mahasiswa.



**Gambar 7**. Output hasil Spearman correlation dan Mann-Whitney U Test

* + - 1. **Pembahasan**

Berdasarkan hasil yang diperoleh:

1. Terdapat perbedaan signifikan antara nilai akademik mahasiswa yang lulus dan yang dropout, baik melalui uji *t-test* maupun *Mann-Whitney*.
2. Hubungan antar variabel akademik bersifat positif namun lemah, sebagaimana ditunjukkan pada hasil *Spearman correlation*.
3. Tahap *preprocessing* berperan penting dalam meningkatkan validitas hasil analisis, karena setelah imputasi dan *scaling*, hasil uji menjadi lebih konsisten dan interpretatif.

Hasil ini menunjukkan bahwa faktor nilai akademik awal memiliki pengaruh terhadap peluang keberhasilan studi mahasiswa. Selain itu, langkah-langkah preprocessing yang tepat terbukti mendukung stabilitas hasil statistik dan memberikan gambaran yang lebih akurat terhadap karakteristik mahasiswa.

# Kesimpulan

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

1. Dataset memiliki kualitas baik dengan jumlah fitur dan observasi yang memadai untuk analisis statistik.
2. Tahapan *preprocessing* seperti *KNN Imputer*, *winsorization*, *scaling*, dan *PCA* berhasil meningkatkan kualitas dan konsistensi data.
3. Hasil *t-test* dan *Mann-Whitney U Test* menunjukkan adanya perbedaan signifikan antar kelompok mahasiswa, sedangkan *Spearman correlation* menunjukkan hubungan positif lemah antar fitur akademik.
4. Secara keseluruhan, nilai akademik awal mahasiswa berpengaruh terhadap status dan keberhasilan studinya.