**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**

Perkembangan teknologi dan digitalisasi dalam bidang pendidikan menghasilkan data akademik yang melimpah dari berbagai aktivitas mahasiswa. Data ini menyimpan potensi besar untuk diolah menggunakan pendekatan *data science* guna memahami faktor-faktor yang mempengaruhi keberhasilan studi, risiko dropout, dan performa akademik mahasiswa.

Melalui proyek ini, dilakukan analisis terhadap dataset *Student Academic Performance* dari UCI Machine Learning Repository. Fokus analisis diarahkan pada hubungan antara nilai masuk mahasiswa (admission grade) dan status akademik akhir (Dropout, Enrolled, Graduate). Pendekatan ini dilakukan dengan pipeline data science yang mencakup *Data Collection, Data Preprocessing*, *Data Visualization, Statistical Analysis*

1. **Tujuan**

Tujuan dari proyek ini adalah:

1. Mengimplementasikan tahapan *data science pipeline* terhadap dataset mahasiswa.
2. Melakukan visualisasi data untuk memahami pola distribusi dan hubungan antar variabel.
3. Mengidentifikasi apakah terdapat perbedaan signifikan nilai *admission grade* antar kelompok *Graduate*, *Dropout*, dan *Enrolled*.
4. Memberikan rekomendasi berbasis data yang dapat mendukung kebijakan akademik kampus.
5. **Rumusan Masalah**
6. Apakah nilai *admission grade* memiliki perbedaan signifikan antar kelompok status mahasiswa (*Graduate*, *Dropout*, *Enrolled*)?
7. Apakah terdapat hubungan antara *admission grade* dan performa akademik semester pertama mahasiswa?
8. Fitur akademik apa yang paling berpengaruh terhadap status akhir mahasiswa berdasarkan hasil analisis statistik?
9. **Metode Penelitian**

penjelasan alur pengerjaan data science: collection, visualization, preprocessing, analysis).

1. **Data Collection**

Dataset yang digunakan berasal dari:

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

**Karakteristik dataset:**

* Jumlah observasi: **4.424 baris**
* Jumlah fitur: **37 kolom**
* Atribut mencakup: umur, status perkawinan, mode pendaftaran, nilai masuk (*admission grade*), nilai per semester, dan status akhir mahasiswa (*target*).

**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.

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**

Beberapa visualisasi utama yang dilakukan:

| **Jenis Visualisasi** | **Alasan Pemilihan** | **Insight Utama** |
| --- | --- | --- |
| **Histogram** | Mengetahui distribusi status perkawinan mahasiswa. | Distribusi sangat miring ke kiri → mayoritas mahasiswa belum menikah. |
| **Boxplot (Admission Grade per Target)** | Membandingkan nilai antar kategori *Target*. | Mahasiswa *Graduate* memiliki median *admission grade* lebih tinggi. |
| **Scatter Plot (Admission Grade vs Application Mode)** | Melihat hubungan antar dua variabel ordinal. | Tidak ada hubungan linear antara status pernikahan dan mode pendaftaran. |
| **Heatmap Korelasi** | Mengetahui kekuatan korelasi antar fitur numerik. | Korelasi tinggi antar nilai akademik semester 1 dan 2. |
| **Violin Plot (Grades vs Target)** | Menganalisis distribusi nilai akademik antar kategori target. | Distribusi nilai *Graduate* lebih rapat di nilai tinggi. |

1. **Statistical Analysis**

#### Uji Parametrik — One-Way ANOVA

#### 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**
2. Mahasiswa dengan *admission grade* tinggi memiliki kecenderungan besar untuk lulus (*Graduate*).
3. Mahasiswa dengan nilai rendah di semester 1 menunjukkan risiko *dropout* lebih besar.
4. Korelasi kuat antara nilai masuk dan nilai akademik semester pertama menunjukkan validitas sistem penerimaan mahasiswa.
5. Faktor non-akademik (seperti *marital status* dan *application mode*) tidak berpengaruh signifikan terhadap hasil akhir studi.
6. **Kesimpulan**
7. Dataset “Predict Students Dropout and Academic Success” berhasil dianalisis sesuai pipeline data science
8. Ditemukan perbedaan signifikan nilai *admission grade* antar kategori *Target* berdasarkan hasil ANOVA dan Mann-Whitney.
9. Korelasi Spearman menunjukkan hubungan kuat antara *admission grade* dan performa akademik awal.
10. Hasil analisis ini dapat membantu pihak kampus dalam:  
    - Mengidentifikasi mahasiswa berisiko dropout lebih dini.  
    - Menyusun intervensi akademik berbasis data (mentoring atau remedial).  
    - Mengevaluasi efektivitas sistem seleksi masuk berdasarkan nilai *admission grade*.