**LAPORAN IMPLEMENTASI TAHAP PREPROCESSING DATA**

**STUDI KASUS: CREDIT APPROVAL DATASET**



**Dosen Pengampu:**

**Junta Zeniarja, M.Kom**

**Disusun oleh:**

**Thalita Nadia Azalai**

**A11.2023.15291**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**

**2025**

**DAFTAR ISI**

[1. PENDAHULUAN DAN DATA SELECTION 3](#_Toc210640484)

[1.1 Latar Belakang dan Tujuan 3](#_Toc210640485)

[1.2 Data Selection, Loading, dan Initial Cleaning 3](#_Toc210640486)

[2. CLEANING DAN TRANSFORMATION TERINTEGRASI 4](#_Toc210640487)

[2.1 Perancangan Pipeline dan ColumnTransformer 4](#_Toc210640488)

[2.2 Justifikasi Scaling: Analisis Skewness (Kemiringan) 4](#_Toc210640489)

[2.3 Demonstrasi Komparatif 3 Teknik Scaler 6](#_Toc210640490)

[2.4 Verifikasi Visual Keberhasilan Standardization 7](#_Toc210640491)

[3. KESIMPULAN DAN PERSIAPAN DATA MINING 8](#_Toc210640492)

[3.1 Verifikasi Numerik dan Konsistensi Transformation 8](#_Toc210640493)

[3.2 Data Selection Akhir (Train-Test Split) 8](#_Toc210640494)

[4. KESIMPULAN 9](#_Toc210640495)

[DAFTAR PUSTAKA 10](#_Toc210640496)

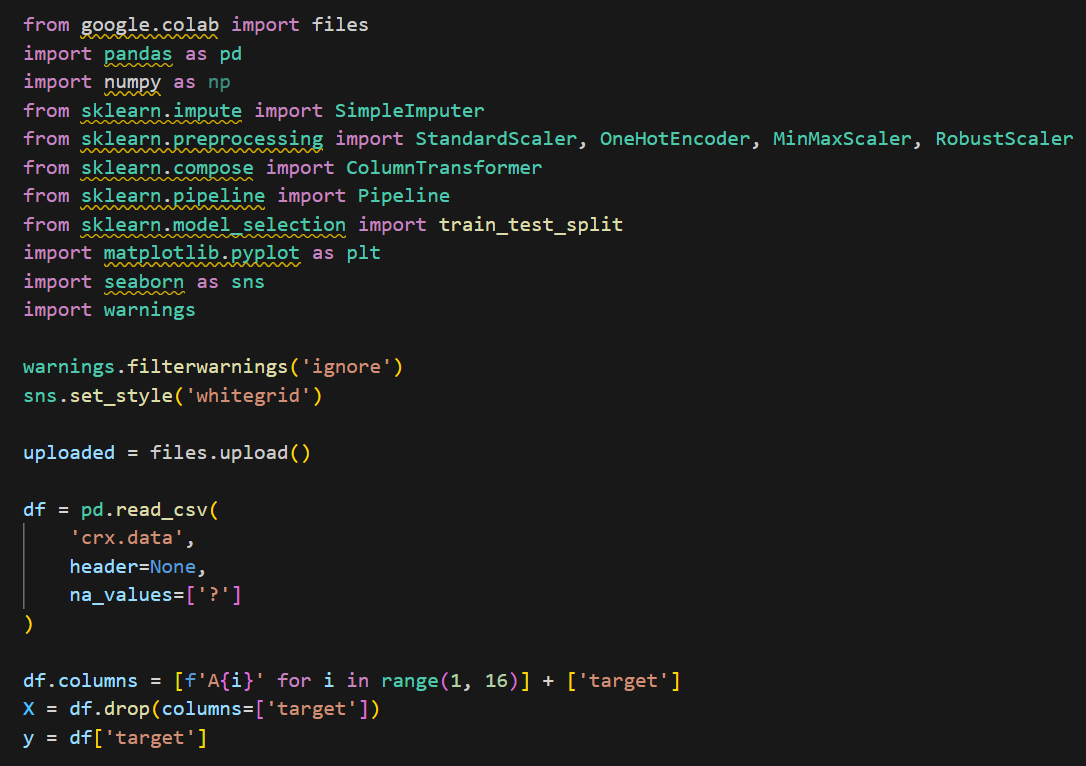
# 1. PENDAHULUAN DAN DATA SELECTION

## 1.1 Latar Belakang dan Tujuan

Laporan ini merupakan implementasi praktis dari tahap **Preprocessing Data**, komponen krusial dalam metodologi **KDD (*Knowledge Discovery in Databases*)**. Studi kasus menggunakan **Credit Approval Dataset** yang diperoleh dari **UCI Machine Learning Repository** [disini](https://archive.ics.uci.edu/dataset/27/credit+approval). Tujuan utamanya adalah mentransformasi data mentah yang memiliki *missing value* dan skala tidak seragam menjadi *input* yang bersih, konsisten, dan terstandardisasi, sehingga optimal untuk proses *Data Mining* selanjutnya.

## 1.2 Data Selection, Loading, dan Initial Cleaning

Tahap ini mencakup pemilihan data dan penanganan *missing value* pada tingkat *loading*.

****

*Data loading* dilakukan dengan mengidentifikasi karakter **?** sebagai *missing value* (na\_values=['?']), yang merupakan sebuah langkah **Pembersihan Awal**. Data kemudian dipisahkan menjadi matriks X (Fitur) dan vektor y (Target).

# 2. CLEANING DAN TRANSFORMATION TERINTEGRASI

## 2.1 Perancangan Pipeline dan ColumnTransformer

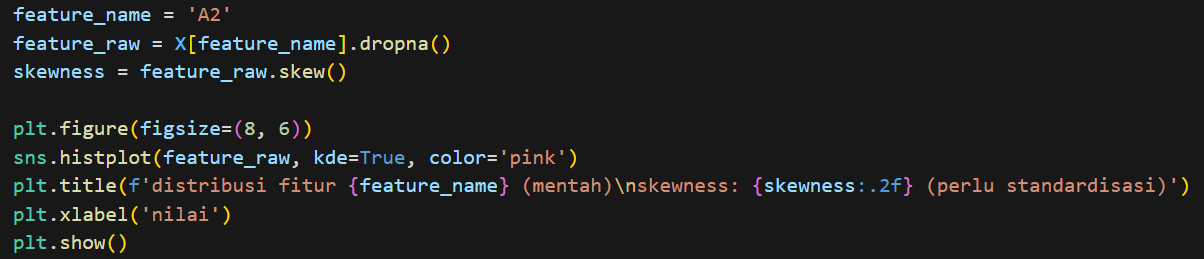
Seluruh proses *cleaning* (*imputation*) dan *transformation* (*scaling* dan *encoding*) diintegrasikan ke dalam Pipeline dan ColumnTransformer.

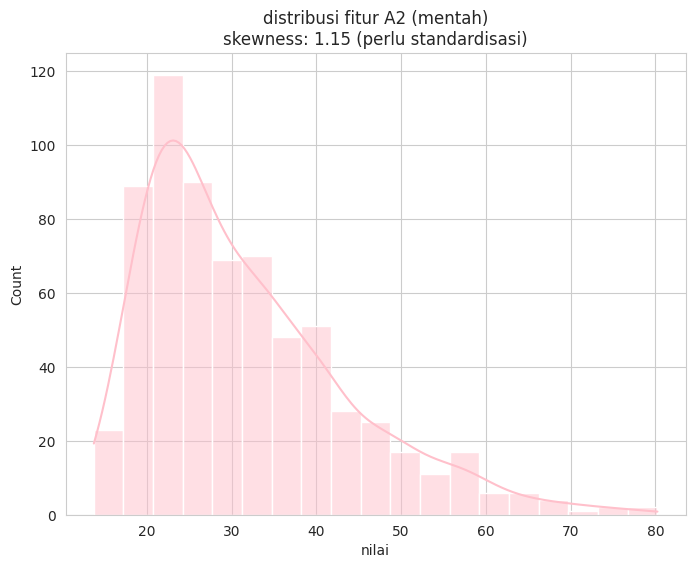


1. **Pipeline Numerik:** Menerapkan SimpleImputer(strategy='median') untuk menangani *missing value*, diikuti oleh **StandardScaler()** yang menjadi fokus utama *transformation*.
2. **Pipeline Kategorikal:** Menerapkan SimpleImputer(strategy='most\_frequent'), diikuti oleh **OneHotEncoder** untuk mengubah variabel nominal menjadi format biner.
3. **Hasil:** **X\_processed\_df** adalah *DataFrame* akhir yang telah melewati kedua proses tersebut secara konsisten.

## 2.2 Justifikasi Scaling: Analisis Skewness (Kemiringan)

Dilakukan *Exploratory Data Analysis* (EDA) untuk memvalidasi perlunya teknik standardisasi.



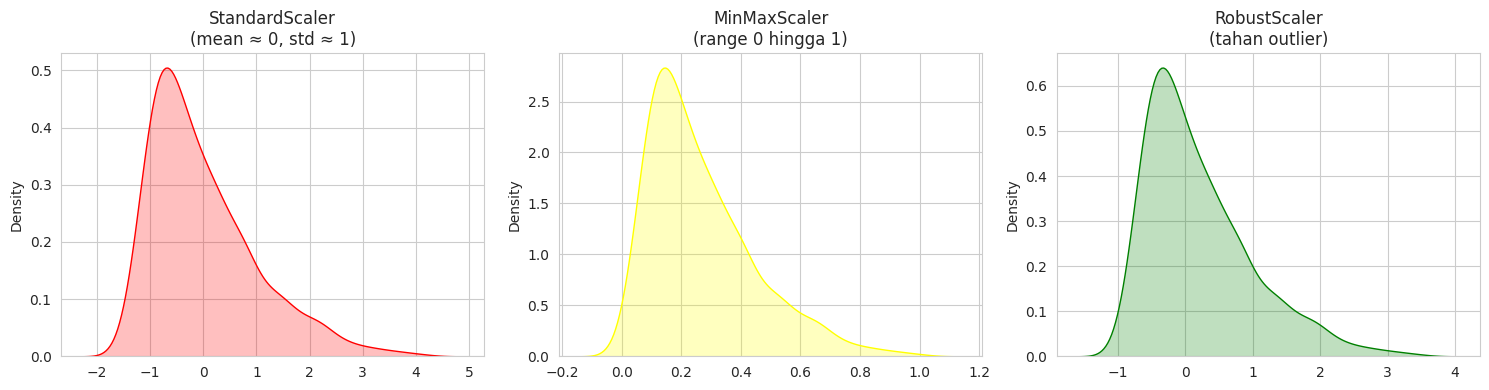


Fitur A2 menunjukkan nilai **Skewness** yang signifikan, yang mengindikasikan **distribusi tidak normal** (*non-Gaussian*). Temuan ini **menjustifikasi** pemilihan StandardScaler (atau RobustScaler) yang relatif lebih baik pada data miring dibandingkan MinMaxScaler.

## 2.3 Demonstrasi Komparatif 3 Teknik Scaler

Demonstrasi ini bertujuan menunjukkan pemahaman terhadap variasi teknik *scaling* sesuai materi (Slide 27-28).

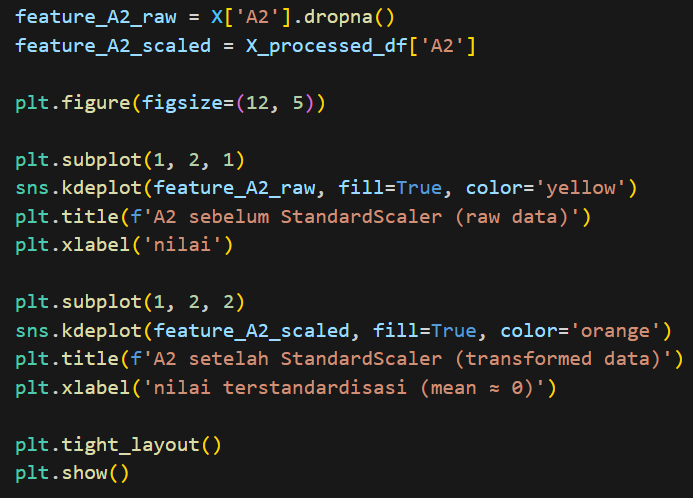


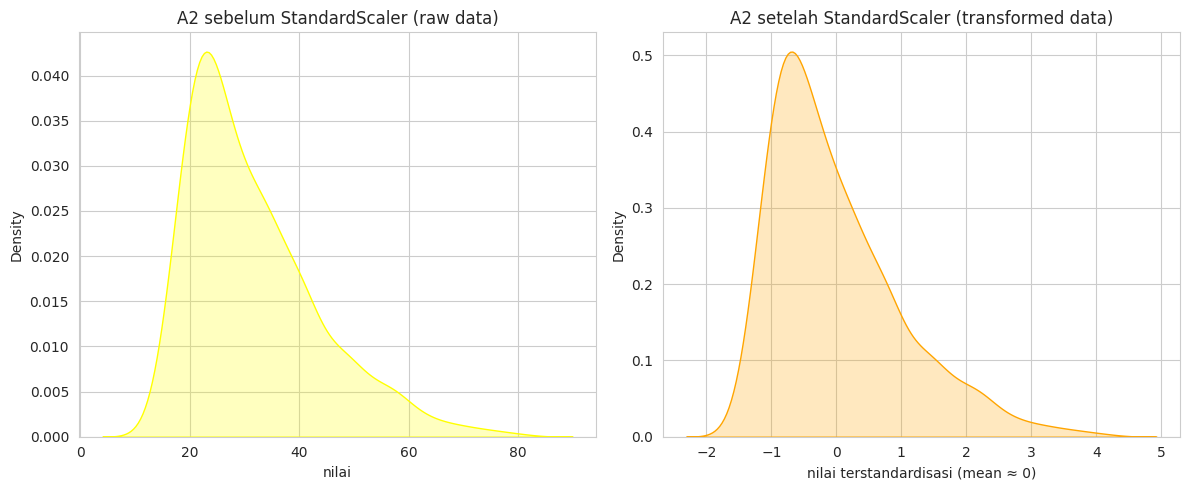


Demonstrasi ini mengonfirmasi prinsip kerja masing-masing *scaler*. **RobustScaler** terbukti efektif terhadap *outlier* (menggunakan IQR), sedangkan **StandardScaler** yang dipilih pada *pipeline* berfungsi sebagai metode *scaling* yang paling umum dan tepat untuk data *machine learning* berbasis jarak.

## 2.4 Verifikasi Visual Keberhasilan Standardization

Verifikasi ini membuktikan secara visual bahwa *StandardScaler* di dalam *pipeline* telah berfungsi dengan benar.





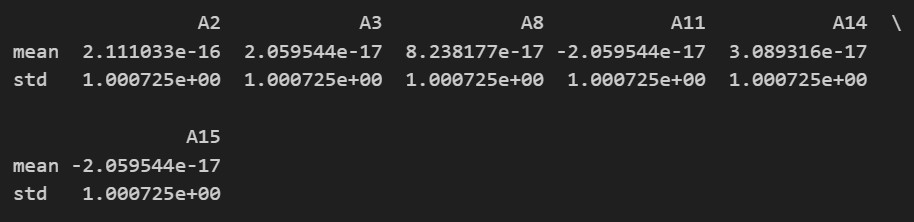
Grafik *transformed data* menunjukkan distribusi yang identik dengan *raw data*, namun seluruh data telah **terpusat di sekitar nilai 0** di sumbu-x. Ini merupakan bukti konklusif bahwa **Standardisasi** telah berhasil dilakukan.

# 3. KESIMPULAN DAN PERSIAPAN DATA MINING

## 3.1 Verifikasi Numerik dan Konsistensi Transformation

Tahap ini memvalidasi hasil *transformation* secara statistik.

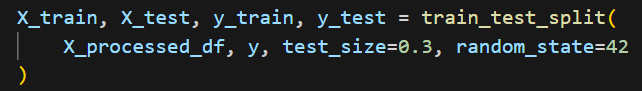




Konsistensi nilai **Mean** yang mendekati **0** dan **Std** yang mendekati **1** di seluruh fitur numerik adalah **bukti numerik** bahwa *StandardScaler* berhasil diterapkan secara menyeluruh dan akurat.

## 3.2 Data Selection Akhir (Train-Test Split)

Langkah terakhir dari *preprocessing* adalah persiapan data untuk *Data Mining*.



Data dibagi menjadi 70% set training dan 30% set testing. Pembagian ini memastikan proses validasi model yang tidak bias, menandai akhir yang sukses dari tahap *Preprocessing* KDD.

# 4. KESIMPULAN

Berdasarkan hasil verifikasi di Bab III, seluruh tahapan *Preprocessing* KDD (**Selection, Cleaning, Transformation**) telah berhasil diimplementasikan secara metodis dan terverifikasi. *Pipeline* yang dirancang telah berhasil:

1. Menangani **Missing Value** (Imputasi Median/Modus).
2. Melakukan **Standardisasi Skala** yang dikonfirmasi oleh Mean≈0 dan Std≈1.
3. **Meng-*encode*** data kategorikal.

Data akhir, X\_processed\_df, kini **konsisten dan optimal** untuk proses selanjutnya.

# DAFTAR PUSTAKA

**Bramer, M.** (2020). *Principles of Data Mining* (4th ed.). Springer.

**Dua, D., & Graff, C.** (2017). **UCI Machine Learning Repository** [Dataset]. Irvine, CA: University of California, School of Information and Computer Science. Diakses dari: <https://archive.ics.uci.edu/dataset/27/credit+approval>

**Han, J., Kamber, M., & Pei, J.** (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

**Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Delehelle, M., Siblini, Y., & Perrot, M.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

**Witten, I. H., Frank, E., & Hall, M. A.** (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Elsevier.