Laporan Kelompok 7
Tugas Group Project (TGP) #1

Disusun oleh:

1. Kayla Putri Maharani | 5026231158
2. Tahiyyah Mufhimah | 5026231170
3. Nicholas Evan Sitanggang | 5026231146
4. Baqhiz Faruq S | 5026231212

DEPARTEMEN SISTEM INFORMASI

FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS

TAHUN AJARAN 2024-2025

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

Kontribusi Anggota :

Nama	NRP	Kontribusi
Kayla Putri Maharani	5026231158	**Jupyter Notebook :** 1. Identifikasi Data Hilang 2. Penghapusan Data 3. Imputasi Data 4. Penghapusan Redundansi 5. Optimasi Data 6. Validasi Akhir **Dokumentasi :** Menuliskan penjelasan code line untuk anggota 2, Penanganan Data yang Hilang dan Pembersihan Data untuk kedua Dataset. **Laporan :** Menyusun alasan singkat tentang alasan memilih metode untuk kedua dataset.
Tahiyyah Mufhimah	5026231170	**Jupyter Notebook** 1. Identifikasi Variabel Target dan Distribusi Kelas 2. Pengukuran Skewness dan Kurtosis 3. Penanganan Ketidakseimbangan Kelas 4. Evaluasi Performa Model Setelah Pra Pemrosesan 5. Membandingkan Dataset yang lebih Siap **Dokumentasi :** Menuliskan penjelasan kode terkait tugas Anggota 4, Analisis Ketidakseimbangan Kelas dan Perbandingan Metode Pra Pemrosesan untuk Kedua Dataset **Laporan :** Menyusun analisis ketidakseimbangan kedua dataset, metode yg digunakan, dan alasan dataset mana yang lebih siap untuk prediksi, kesimpulan dan analisis akhir
Nicholas Evan	5026231146	**Jupyter Notebook** 1. Transformasi Variabel Kategorikal 2. Mengurangi Penggunaan Memori 3. Penciptaan Fitur Baru 4. Evaluasi Performa Dataset setelah Transformasi 5. Mempersiapkan Dataset untuk Pemodelan 6. Validasi Akhir **Dokumentasi :** Menuliskan penjelasan code line untuk anggota 3, Penanganan Data yang Hilang dan Pembersihan Data untuk kedua Dataset. **Laporan :** Menyusun alasan singkat tentang alasan memilih metode untuk kedua dataset.
Baqhiz Faruq S	5026231212	**Jupyter Notebook** 1. Memahami struktur dataset (jumlah baris, kolom, tipe data, nilai yang hilang, dll). 2. Memvisualisasikan missing values dan variabel target 3. Menentukan distribusi fitur numerik dan kategorikal. 4. Visualisasi dan memahami pola dan tren dalam dataset. 5. Distribusi data ada fitur-fitur penting.

		Dokumentasi : Menuliskan penjelasan kode line untuk anggota 1, Eksplorasi Dataset dan Analisis Data Awal (EDA) **Laporan :** Menyusun ringkasan hasil eksplorasi data dari kedua dataset.

Ringkasan Hasil Eksplorasi Data dari Kedua Dataset

- **Lending Club Loan data**

  EDA yang dilakukan pada dataset pinjaman yang diterima (Accepted Loans) dan ditolak (Rejected Loans) memberikan beberapa temuan penting:

  1. **Missing Values:**

  - Accepted Loans**:** Banyak kolom memiliki missing values, beberapa lebih dari 2 juta, menunjukkan perlunya penanganan (imputasi/penghilangan kolom).
  - Rejected Loans**:** Kolom Risk_Score paling banyak missing, dengan Employment Length, Debt-To-Income Ratio, dan Zip Code juga menunjukkan kekurangan data, meskipun lebih sedikit.

  2. **Status Pinjaman (Accepted Loans):**

     Mayoritas pinjaman berstatus "Fully Paid" atau "Current". Terdapat jumlah signifikan pinjaman "Charged Off" sebagai indikator kerugian, serta kategori "In Grace Period", "Late", dan "Default" yang menunjukkan kesulitan pembayaran atau ketidakpatuhan kebijakan.

  3. **Distribusi Fitur Utama (Accepted Loans):**

  - Loan Amount**:** Umumnya antara USD 10.000–20.000, dengan beberapa mencapai USD 40.000.
  - Interest Rate**:** Berkisar antara 10% hingga 15%.
  - Installment**:** Cicilan bulanan kebanyakan antara USD 200–400, menurun dengan meningkatnya cicilan.

  4. **Distribusi Fitur Utama (Rejected Loans):**

  - Amount Requested**:** mayoritas pengajuan kecil hingga menengah.
  - Risk Score**:** kebanyakan antara 500–800, dengan puncak di sekitar 650.
  - Application Date**:** pengajuan yang ditolak berfluktuasi antara 2008–2018, dengan puncak sekitar tahun 2014.

2. Home Credit Default risk

Dataset ini berisi informasi mengenai pengajuan kredit, termasuk data demografis, keuangan, dan status kredit peminjam. Terdapat 122 kolom dengan tipe data numerik dan kategorikal.

1. **Missing Values:**

- Kolom seperti COMMONAREA_AVG, NONLIVINGAPARTMENTS_AVG, FONDKAPREMONT_MODE memiliki >68% missing. Perlu penanganan khusus (imputasi, penghapusan, atau model robust).

2. **Analisis Numerik:**

- TARGET**:** Data tidak seimbang, hanya 8.07% default.
- AMT_INCOME_TOTAL**:** Ada outlier ekstrim yang berpotensi mempengaruhi analisis.
- CNT_CHILDREN**:** Mayoritas tanpa anak, namun ada beberapa entri dengan jumlah anak sangat tinggi.
- AMT_CREDIT**:** Variasi data yang tinggi.

3. **Analisis Kategorikal:**

- Demografi & Sosial Ekonomi**:** Mayoritas peminjam perempuan, tidak memiliki mobil, memiliki properti, berpendidikan menengah, dan berstatus menikah.
- Pekerjaan**:** Laborers adalah jenis pekerjaan paling umum.
- Aplikasi & Jenis Kredit**:** Pengajuan paling sering di hari Selasa dan Cash Loans mendominasi.

4. Visualisasi:

- Missing Values**:** Banyak kolom dengan missing values tinggi yang mempengaruhi analisis.
- Target**:** 91.93% non-default dan 8.07% default.
- Fitur Numerik (AMT_CREDIT, AMT_INCOME_TOTAL, AMT_ANNUITY)**:** Distribusi miring ke kanan (nilai kecil hingga sedang).

5. Fitur Kategorikal Detail:

- NAME_CONTRACT_TYPE**:** Cash Loans (90.48%).
- CODE_GENDER**:** Lebih banyak perempuan (65.83%).
- FLAG_OWN_CAR**:** 65.99% tidak memiliki mobil.
- FLAG_OWN_REALTY**:** 69.37% memiliki properti.

Alasan Memilih Metode pada Langkah Pembersihan

Metode yang digunakan dalam setiap dataset didasarkan pada tingkat missing values, relevansi data, serta dampak terhadap analisis keseluruhan. Alasan pemilihan metode untuk langkah pembersihan, yaitu :

1. Identifikasi dan Penghapusan Kolom dengan Banyak Missing Values

- Kolom dengan lebih dari 51% nilai hilang dihapus karena dianggap memiliki terlalu banyak data yang tidak lengkap, sehingga kurang bermanfaat untuk analisis.
- Teknik ini membantu mengurangi noise dalam data serta meningkatkan efisiensi pemrosesan.

2. Imputasi Missing Values dengan Median atau Mode

- Untuk data numerik, nilai yang hilang diisi menggunakan median, karena median lebih tahan terhadap outlier dibandingkan mean.
- Untuk data kategorikal, nilai yang hilang diisi menggunakan mode (nilai yang paling sering muncul), seperti dalam kolom emp_title pada dataset accepted_2007_to_2018Q4.

3. Penghapusan Baris dengan Missing Values Kecil (<10%)

- Jika suatu kolom memiliki nilai hilang kurang dari 10%, baris yang mengandung missing values tersebut dihapus karena jumlahnya kecil dan dampaknya terhadap analisis minimal.
- Contoh pada dataset rejected_2007_to_2018Q4, kolom loan_title, zip_code, dan state hanya memiliki sedikit missing values sehingga baris yang mengandung nilai hilang tersebut dihapus.

4. Identifikasi dan Penghapusan Kolom Redundan

- Beberapa kolom yang memiliki korelasi tinggi dalam missing values diidentifikasi dan hanya satu kolom yang dipertahankan.
- Contoh pada dataset application_train, ditemukan beberapa kolom yang selalu hilang bersama-sama (APARTMENTS_MEDI, BASEMENTAREA_AVG, dll.), sehingga hanya satu kolom yang dipertahankan dari setiap kelompok.

5. Visualisasi dan Validasi Setelah Pembersihan

- Setelah proses pembersihan, dilakukan visualisasi missing values menggunakan **heatmap** untuk memastikan bahwa tidak ada lagi nilai yang hilang secara signifikan.
- Contoh pada dataset installments_payments, setelah dilakukan pembersihan, diperiksa kembali apakah terdapat missing values menggunakan teknik heatmap dan hasilnya menunjukkan bahwa data sudah bersih.

Dampak Transformasi Fitur terhadap Dataset

A. Fitur baru yang telah dibuat terhadap data accepted, rejected, dan home credit default risk:

1. Preprocessing:

- Downcasting: Mengubah tipe data numerik menjadi tipe data yang lebih hemat memori (int64 ke int32, float64 ke float32).
- Konversi Tanggal: Mengubah kolom tanggal (misalnya 'issue_d', 'last_pymnt_d') menjadi format datetime di dataframe 'df_accepted'.
- Ekstrak Zip Code: Membuat kolom baru 'zip_3digits' yang berisi 3 digit pertama kode pos di dataframe 'df_accepted' dan 'df_rejected'.
- Hapus Kolom: Menghapus kolom yang tidak diperlukan, seperti 'zip_code' di dataframe 'df_accepted' dan 'Zip Code' di dataframe 'df_rejected'.
- Mengubah kolom 'Debt-To-Income Ratio' menjadi float dan menghapus tanda '%' di dataframe 'df_rejected'.

2. Feature Engineering:

- Label Encoding: Mengubah variabel kategori menjadi numerik dengan Label Encoding pada kolom-kolom yang memiliki jumlah kategori yang terbatas. Hal ini dilakukan untuk dataframe 'df_accepted', 'df_rejected', dan 'df_application_train'.
- Drop Columns: Menghapus kolom kategori asli setelah label encoding

B. Contoh Penerapan Fitur Baru:

1. Pada dataframe 'df_accepted', kolom 'grade' yang awalnya berupa kategori ('A', 'B', 'C', ...) diubah menjadi numerik (0, 1, 2, ...).
2. Pada dataframe 'df_rejected', kolom 'Application Date' yang awalnya berupa string tanggal diubah menjadi format datetime.
3. Pada dataframe 'df_application_train', kolom-kolom kategori seperti 'NAME_CONTRACT_TYPE' dan 'CODE_GENDER' diubah menjadi numerik dengan label encoding.

C. Manfaat/Dampak dari Fitur Baru:

1. Mengurangi penggunaan memori.

- Dengan ukuran dataset yang lebih kecil, operasi perhitungan seperti filtering, aggregasi, dan join antar tabel menjadi lebih efisien dan hemat sumber daya komputasi.

2. Memudahkan dalam pemodelan machine learning karena model-model machine learning biasanya bekerja lebih baik dengan data numerik.

- Menghapus kolom kategori asli setelah Label Encoding menghindari data redundan dan memastikan dataset lebih bersih dan ringkas.
- Label Encoding mengubah variabel kategorikal menjadi angka, memungkinkan algoritma machine learning yang tidak dapat menangani data string untuk menggunakannya secara langsung.
- Transformasi fitur grade dan kategori lainnya membantu meningkatkan performa model, terutama untuk model berbasis pohon keputusan dan regresi yang bekerja lebih baik dengan data numerik.

3. Memudahkan dalam analisis dan visualisasi data.

- Dengan data numerik yang lebih bersih dan seragam, proses eksplorasi data menggunakan teknik statistik dan visualisasi seperti histogram, scatter plot, atau heatmap menjadi lebih akurat dan mudah diinterpretasikan.

4. Meningkatkan kualitas dan konsistensi data.

- Menghapus kolom yang tidak relevan menghindari duplikasi informasi dan menyederhanakan dataset agar lebih fokus pada fitur yang benar-benar berguna.

Ringkasan Hasil Analisis Ketidakseimbangan Kelas dan Perbandingan Metode Prapemrosesan

1. Variabel Target dalam Masing-Masing Dataset dan Evaluasi Distribusi Kelas

Dalam **dataset Lending Club Loan (Accepted & Rejected)**, variabel target yang digunakan adalah **loan_status** pada data accepted dan **risk_score** pada data rejected. Pada data accepted, loan_status menunjukkan status pembayaran pinjaman, seperti “Fully Paid” atau “Charged Off,” yang mencerminkan

apakah peminjam berhasil melunasi pinjaman atau mengalami gagal bayar. Sementara itu, dalam data rejected, risk_score digunakan untuk menilai risiko pemohon pinjaman yang tidak diterima.

Pada **dataset Home Credit Default Risk**, variabel target yang digunakan dalam **application_train** adalah **AMT_CREDIT**, yang merepresentasikan jumlah kredit yang diberikan kepada peminjam. Sementara pada **installments_payments**, **tidak ada variabel target** kategorikal yang bisa digunakan untuk analisis keseimbangan kelas, karena semua fitur berupa nilai numerik kontinu, seperti jumlah pembayaran dan keterlambatan. Oleh karena itu, analisis **ketidakseimbangan kelas tidak dapat diterapkan pada installments_payments.**

2. Analisis Visualisasi Distribusi Kelas dan Ketidakseimbangan Data

Pada **loan_status**, terlihat bahwa kelas "Fully Paid" mendominasi dataset, sedangkan kelas "Charged Off" memiliki jumlah yang jauh lebih sedikit, menunjukkan ketidakseimbangan yang cukup besar. Demikian pula, pada **risk_score** dalam dataset rejected, distribusi data sangat tidak merata dengan sebagian besar pemohon memiliki skor risiko tertentu, menyebabkan skews dalam data.

Dalam **dataset Home Credit Default Risk**, histogram distribusi **AMT_CREDIT** menunjukkan bahwa distribusi cukup normal dengan skewness yang rendah. Oleh karena itu, tidak ada ketidakseimbangan kelas yang perlu ditangani, karena dataset ini lebih cocok untuk regresi daripada klasifikasi.

3. Metode yang Digunakan untuk Menangani Ketidakseimbangan Kelas

Pada dataset Lending Club Loan, beberapa metode diterapkan untuk menangani ketidakseimbangan kelas:

- **SMOTE (Synthetic Minority Over-sampling Technique)** digunakan untuk menambah jumlah sampel kelas minoritas secara sintetis agar lebih seimbang dengan kelas mayoritas.
- **Class Weighting** diterapkan pada model pembelajaran untuk memberikan bobot lebih besar pada kelas yang kurang terwakili, membantu model lebih akurat dalam memprediksi kelas minoritas.
- **Oversampling** dilakukan dengan menambahkan sampel dari kelas minoritas tanpa menghapus data kelas mayoritas, sehingga distribusi kelas menjadi lebih proporsional.

Metode-metode ini tidak diterapkan pada dataset Home Credit Default Risk karena dataset ini tidak mengalami ketidakseimbangan kelas dan lebih cocok untuk prediksi berbasis regresi.

4. Perbandingan Hasil Prapemrosesan dan Dataset yang Lebih Siap untuk Prediksi

Setelah menerapkan metode penyeimbangan kelas pada **Lending Club Loan**, hasil evaluasi menunjukkan peningkatan performa model, terutama dalam recall pada kelas minoritas. Setelah penerapan **SMOTE dan class weighting**, distribusi kelas menjadi lebih seimbang, meningkatkan kemampuan model dalam mengenali pola dari kelas yang sebelumnya kurang terwakili.

Sebaliknya, dalam **Home Credit Default Risk**, tidak ada perubahan signifikan dalam distribusi karena dataset ini tidak memerlukan metode penyeimbangan kelas. Distribusi data tidak terlalu condong ke satu sisi (asimetri rendah) dan tidak memiliki ekor yang terlalu berat. Dengan demikian, tidak diperlukan teknik penyeimbangan data. Kondisi ini mengindikasikan bahwa **dataset Home Loan Credit sudah lebih siap digunakan untuk proses prediksi** tanpa perlu penyesuaian lebih lanjut terhadap distribusi kelas.

Kesimpulan dan Analisis Akhir

Berdasarkan analisis yang telah dilakukan, Home Credit Default Risk lebih siap untuk proses prediksi dibandingkan Lending Club Loan, terutama karena sifat variabel targetnya yang numerik kontinu. Distribusi AMT_CREDIT menunjukkan skewness yang rendah dan tidak memiliki ekor yang berat (low kurtosis), yang mengindikasikan bahwa data tidak terlalu asimetris dan lebih stabil untuk model regresi. Dalam ilmu statistik, distribusi yang mendekati normal lebih memudahkan model machine learning dalam melakukan estimasi karena mengurangi risiko overfitting akibat data yang tidak seimbang. Selain itu,

tidak adanya ketidakseimbangan kelas pada dataset ini berarti tidak perlu dilakukan langkah tambahan seperti SMOTE atau class weighting, sehingga proses pra pemrosesan lebih efisien dan model dapat langsung digunakan untuk prediksi tanpa perlu mengubah distribusi data.

Sebaliknya, Lending Club Loan mengalami ketidakseimbangan kelas yang signifikan, khususnya pada variabel loan_status dan risk_score. Dalam machine learning, ketidakseimbangan kelas dapat menyebabkan bias pada model, di mana model lebih cenderung memprediksi kelas mayoritas karena kelas minoritas kurang terwakili dalam pelatihan. Oleh karena itu, metode seperti SMOTE dan class weighting harus diterapkan agar model dapat mengenali pola dalam kelas minoritas dengan lebih baik. Meskipun langkah-langkah ini dapat meningkatkan performa model, proses ini menambah kompleksitas karena model harus diuji kembali untuk memastikan tidak terjadi overfitting pada data sintetis yang dihasilkan. Dengan demikian, meskipun Lending Club Loan dapat digunakan setelah penyesuaian, Home Credit Default Risk tetap lebih siap untuk prediksi karena tidak memerlukan modifikasi tambahan pada distribusi kelasnya.