



# CREDIT RISK PREDCTION

Project Based Internship: Data Scientist ID/X

Partners X Rakamin Academy

Muhammad Mirza Faiz Rabbani





# Muhammad Mírza Faíz Rabbaní

Saya merupakan mahasiswa Program Studi Informatika di Universitas Diponegoro yang saat ini sedang dalam semester 7. Selama masa studi, saya memiliki minat yang kuat dalam bidang data science, khususnya pada penerapan machine learning, data analytics, dan data visualization untuk mendukung pengambilan keputusan berbasis data.



## Database Lab Assistant

Diponegoro University  
JAugust 2024 - Desember 2024 - 5 Months

## Operation System Lab Assistant

Diponegoro University  
JAugust 2024 - November 2024 - 4 Months

## Software Engineer

Thrive - Internship  
Januari 2025 - Februari 2025 - 2 Months



# Study Case

Proyek ini merupakan kolaborasi antara ID/X Partners dan sebuah perusahaan lending yang menghadapi tantangan dalam menilai risiko kredit calon peminjam. Untuk membantu proses pengambilan keputusan, dikembangkan model prediksi credit risk berbasis machine learning menggunakan data historis pinjaman yang mencakup peminjam yang diterima maupun ditolak.

Model ini bertujuan mengklasifikasikan calon peminjam ke dalam kategori berisiko tinggi atau berisiko rendah, sehingga perusahaan dapat meminimalkan potensi gagal bayar dan meningkatkan efisiensi proses pemberian pinjaman. Hasil akhir proyek meliputi model prediksi terlatih dan visualisasi interaktif untuk mendukung presentasi solusi kepada klien.



# Data Collecting

Dataset yang digunakan dalam proyek ini berasal dari file loan\_data\_2007\_2014.csv, dengan total sebanyak 466.285 data pinjaman. Setiap data merepresentasikan satu aplikasi pinjaman yang diajukan oleh peminjam, lengkap dengan status akhir pinjaman tersebut.

Berdasarkan analisis awal, distribusi nilai pada kolom loan\_status adalah sebagai berikut:

- Current: 224.226 data
- Fully Paid: 184.739 data
- Charged Off: 42.475 data
- Late (31–120 days): 6.900 data
- In Grace Period: 3.146 data
- Does not meet the credit policy – Fully Paid: 1.988 data
- Late (16–30 days): 1.218 data
- Default: 832 data
- Does not meet the credit policy – Charged Off: 761 data

# Data Understanding

Setelah memahami distribusi status pinjaman, langkah selanjutnya adalah melakukan pengelompokan data menjadi dua kategori utama, yaitu:

Good Loan → pinjaman yang menunjukkan performa baik atau telah dilunasi dengan lancar.

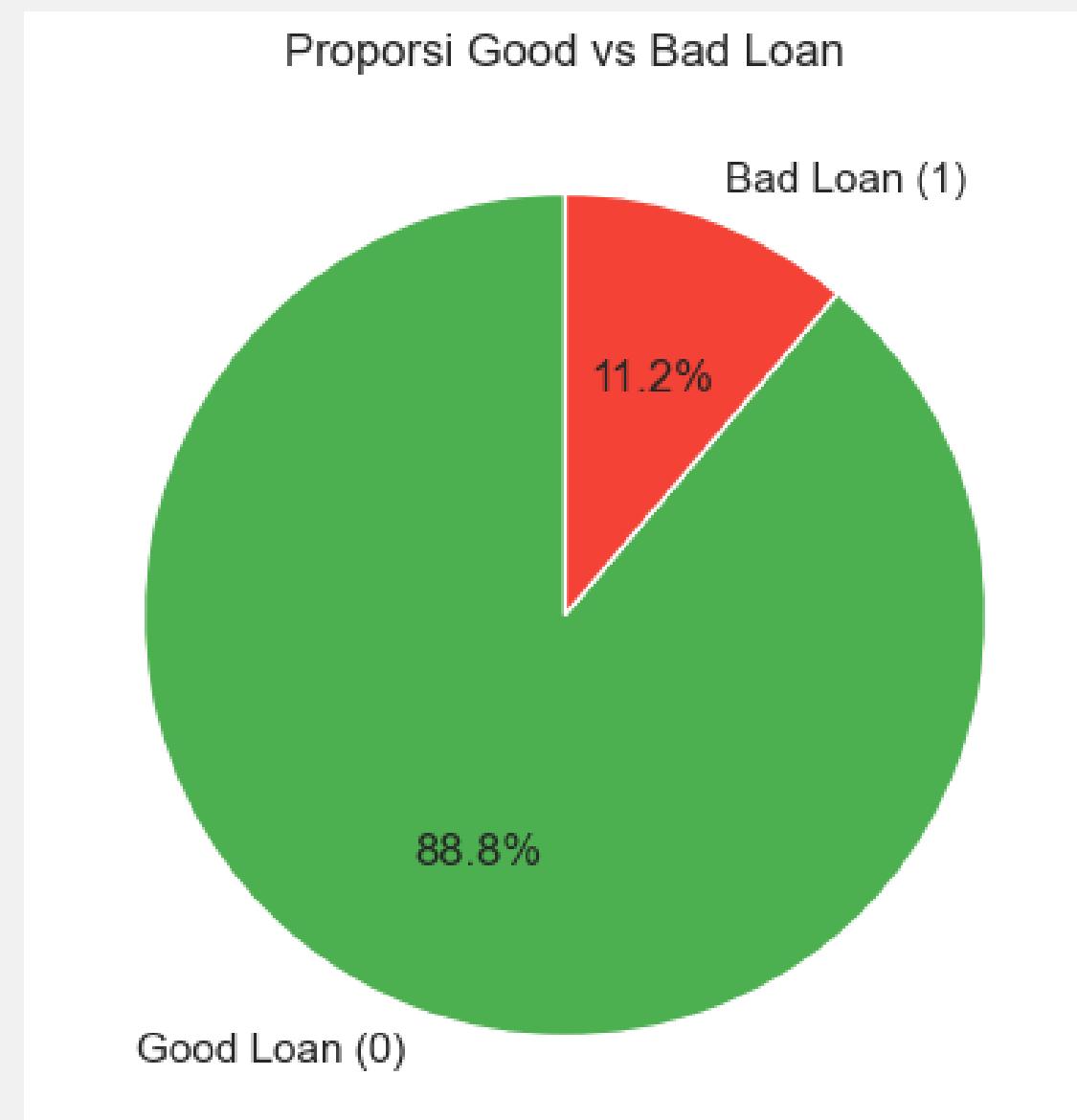
Kategori ini mencakup status

- Current
- Fully Paid
- In Grace Period
- Does not meet the credit policy – Fully Paid

Bad Loan → pinjaman yang berisiko tinggi atau mengalami gagal bayar.

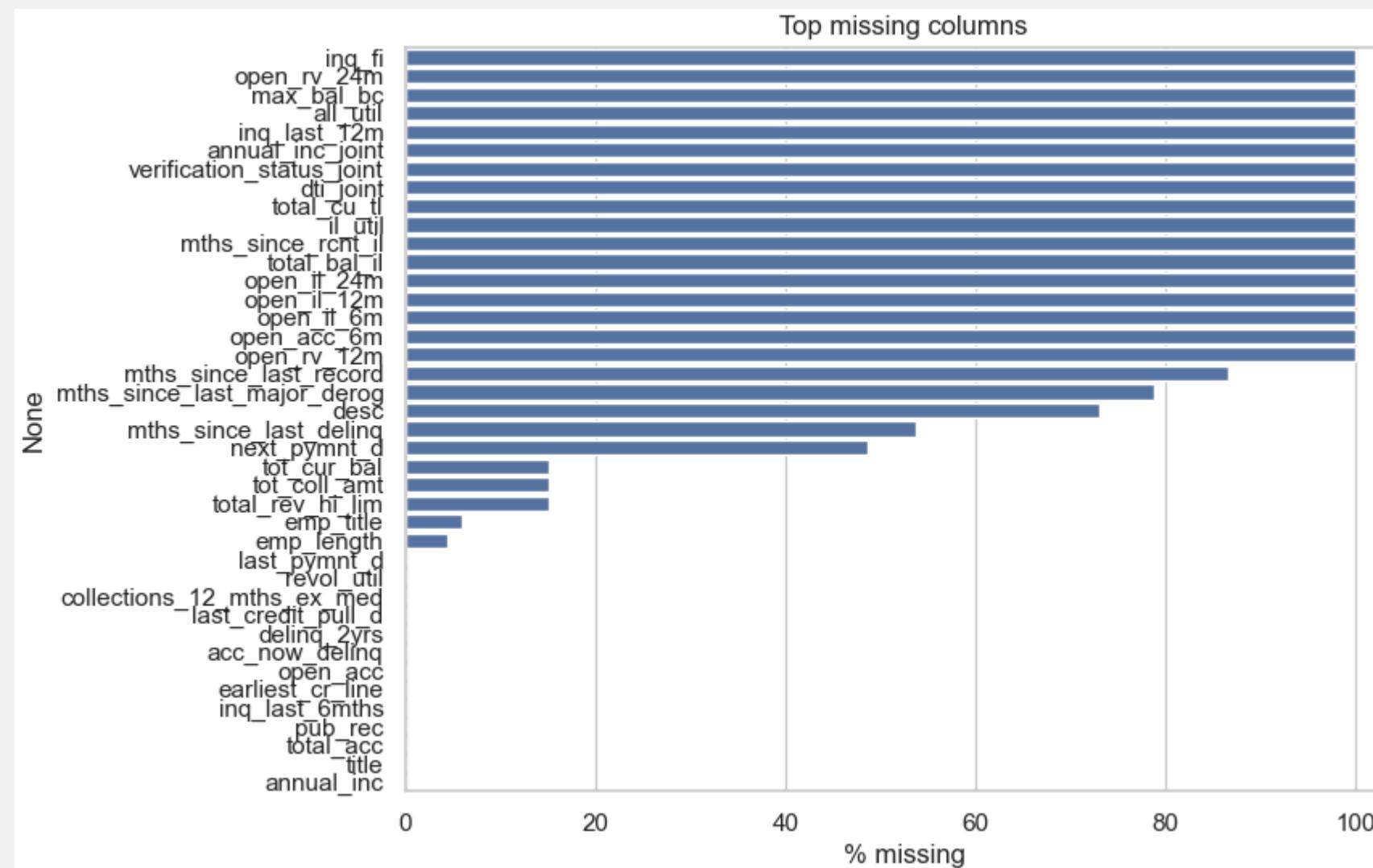
Kategori ini mencakup status:

- Charged Off
- Default
- Late (31–120 days)
- Late (16–30 days)
- Does not meet the credit policy – Charged Off





# Data Understanding



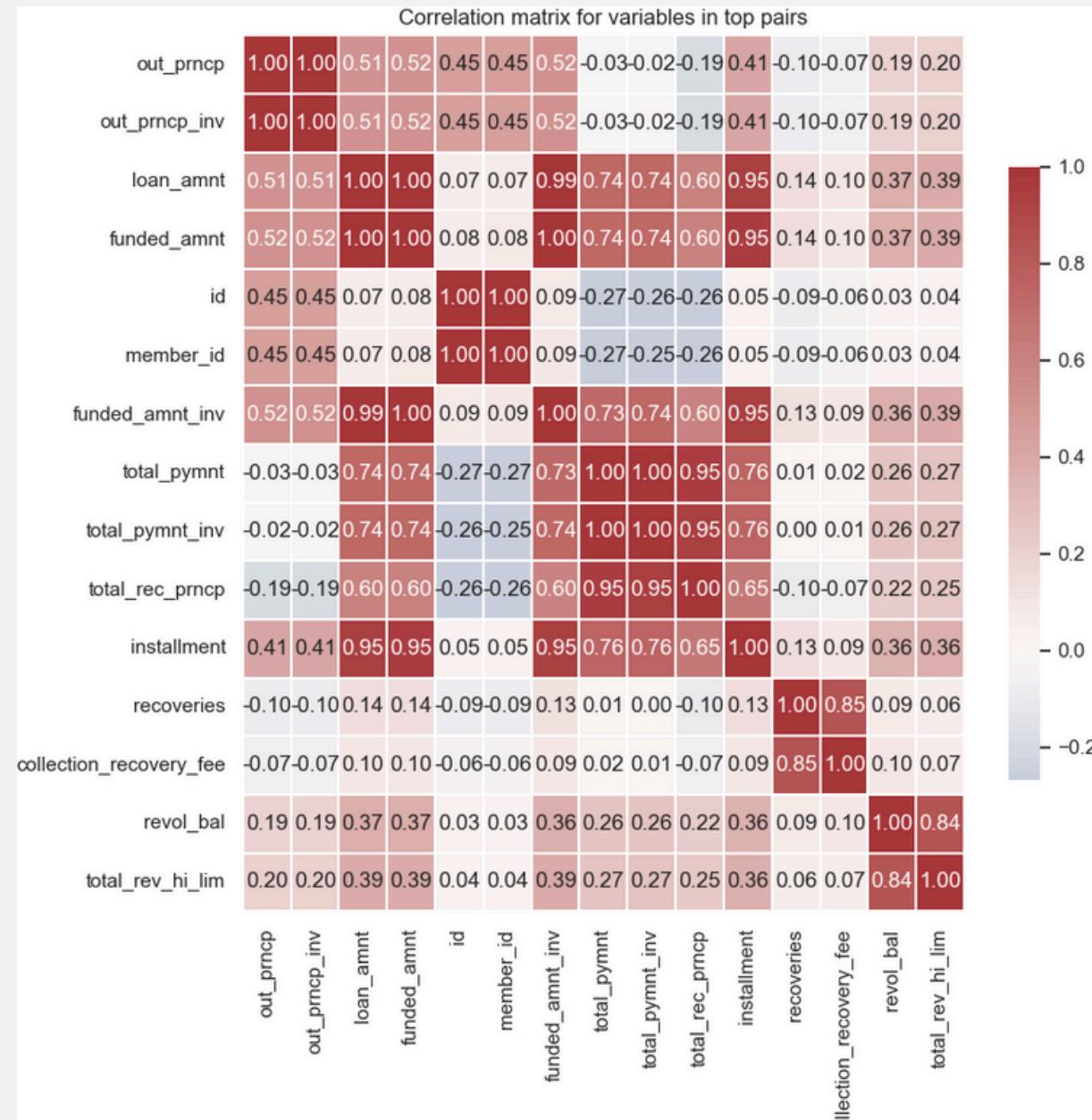
Terlihat bahwa beberapa kolom seperti inq\_fi, open\_rv\_24m, max\_bal\_bc, all\_util, inq\_last\_12m, dan annual\_inc\_joint memiliki tingkat missing hampir 100%, artinya sebagian besar datanya tidak terisi. Kolom seperti ini biasanya tidak relevan untuk model dan lebih baik dihapus karena tidak memberikan informasi yang cukup.

Kolom lain seperti mths\_since\_last\_record, desc, next\_pymnt\_d, dan tot\_coll\_amt memiliki missing values sekitar 40–70%, sehingga perlu dipertimbangkan apakah akan diisi (imputation) atau dihapus, tergantung pentingnya fitur tersebut bagi model.

Sedangkan kolom seperti emp\_title, emp\_length, dan last\_pymnt\_d memiliki missing values di bawah 20%, yang masih dapat ditangani dengan teknik seperti pengisian nilai median, modus, atau kategori “Unknown”.



# Data Understanding

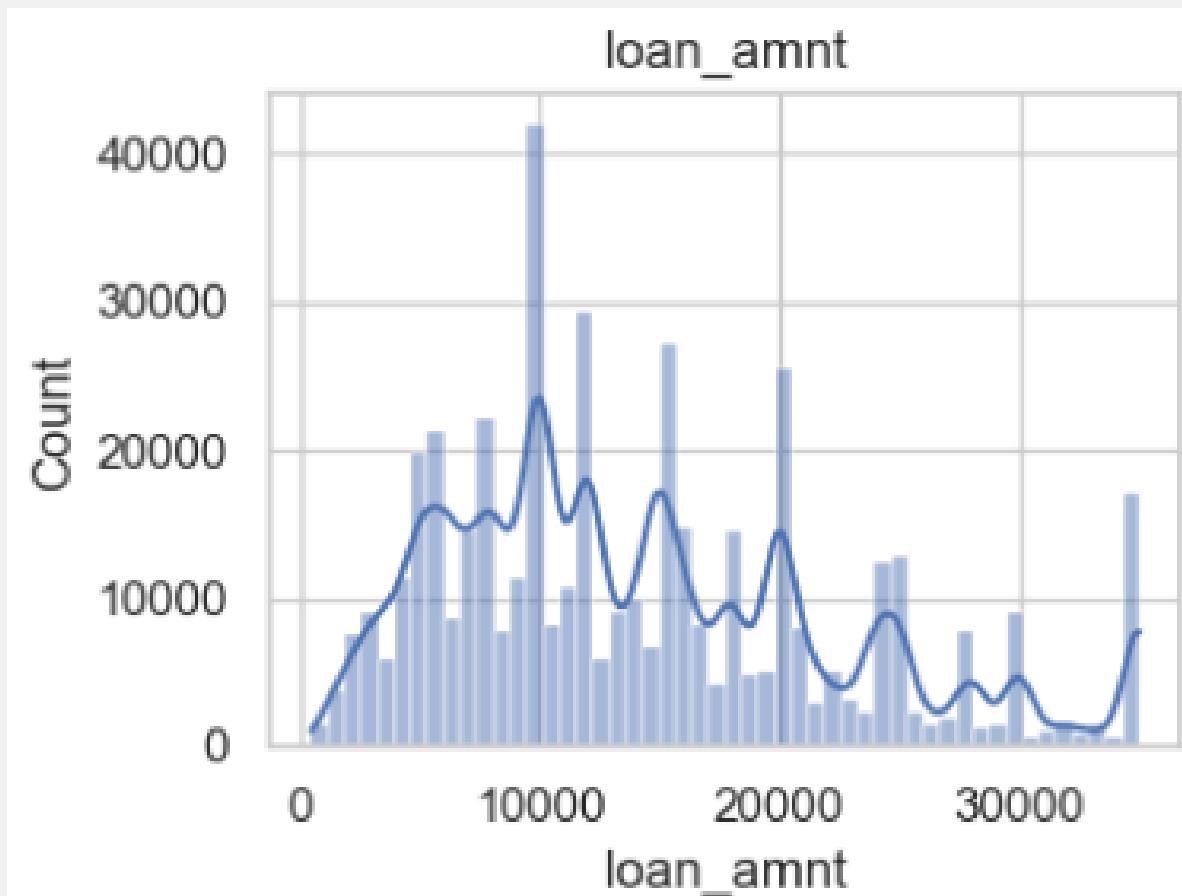


Gambar tersebut adalah heatmap korelasi yang menampilkan hubungan antar variabel numerik pada dataset pinjaman. Visualisasi ini membantu memahami seberapa kuat hubungan (korelasi) antara satu variabel dengan variabel lain, yang sangat berguna dalam tahap analisis fitur (feature analysis) sebelum membangun model machine learning.

Banyak variabel menunjukkan redundansi data (informasi ganda). Sebelum melatih model, penting untuk menghapus fitur dengan korelasi tinggi ( $>0.9$ ) agar model tidak overfitting. Fitur seperti loan\_amnt, installment, dan total\_pymnt bisa menjadi prediktor utama dalam menilai risiko kredit karena berhubungan langsung dengan performa pinjaman.



# Data Understanding

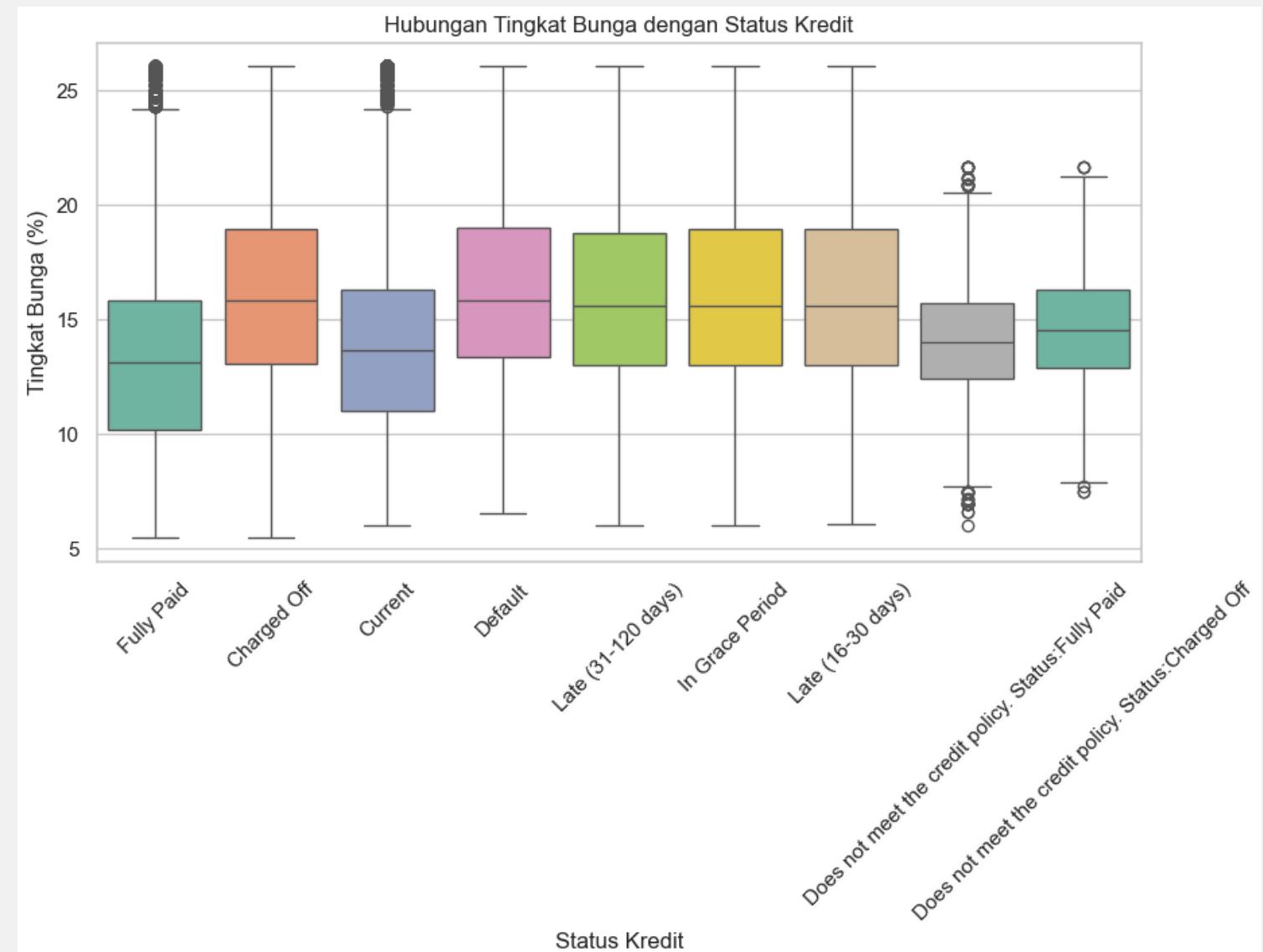


- **Distribusi Berdasarkan jumlah pinjaman**

Histogram ini menggambarkan distribusi jumlah pinjaman, menunjukkan bahwa frekuensi pinjaman tertinggi terkonsentrasi di sekitar angka 10.000, dengan puncak-puncak frekuensi lainnya terlihat pada jumlah pinjaman sekitar 5.000 - 10.000 meningkat, sementara frekuensi cenderung menurun untuk jumlah pinjaman yang lebih besar dari 10.000 meskipun terdapat beberapa lonjakan, dan jumlah pinjaman yang sangat kecil di bawah 5.000 memiliki frekuensi yang relatif rendah.



# Data Understanding

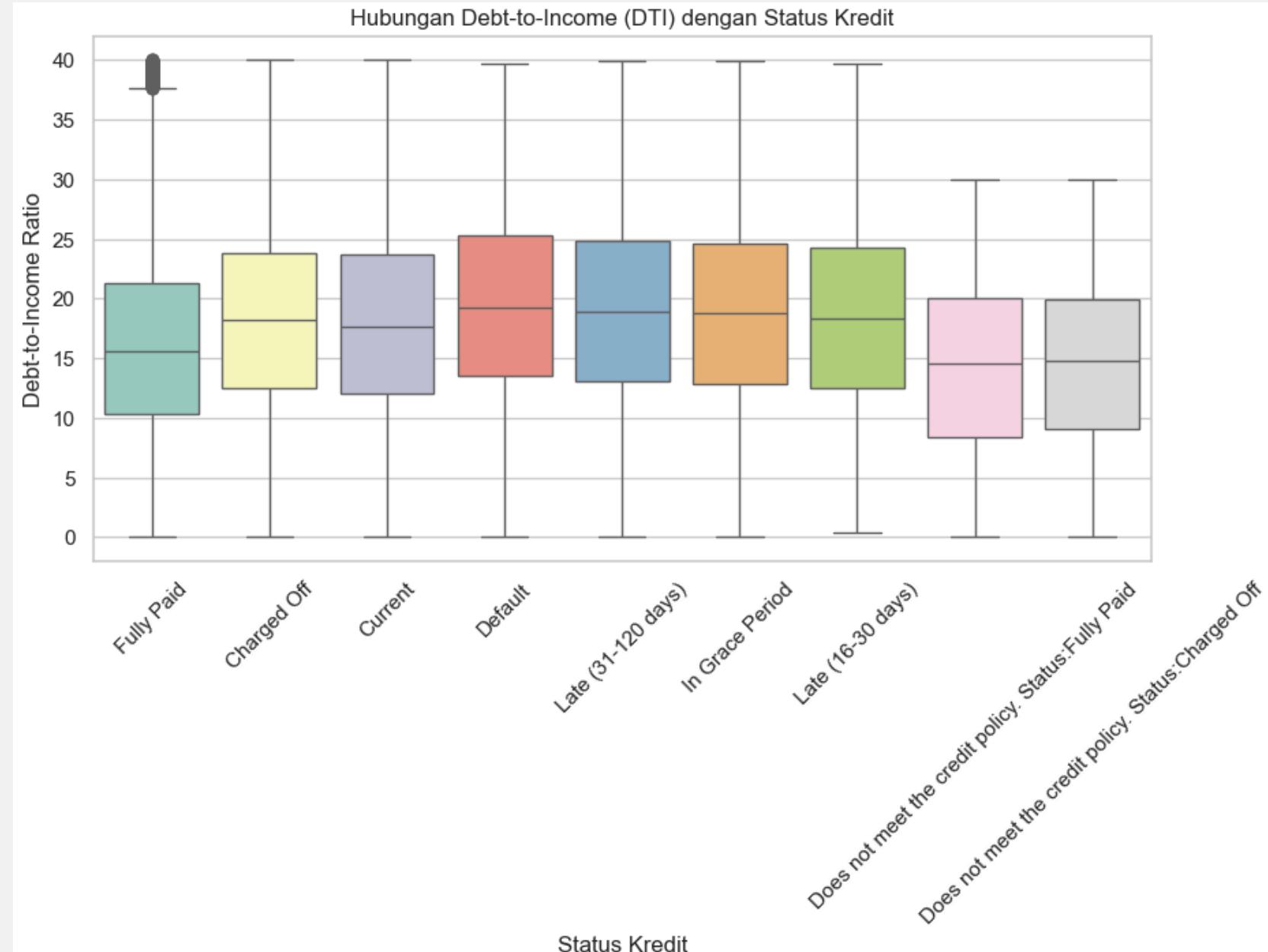


- **Hubungan Tingkat Bunga Dengan Status Kredit**

Grafik ini menegaskan bahwa Ada korelasi antara tingkat bunga dan risiko kredit. Semakin tinggi tingkat bunga, semakin besar peluang gagal bayar. Model prediksi risiko kredit dapat mempertimbangkan tingkat bunga sebagai salah satu fitur penting karena berkaitan langsung dengan kemungkinan default.



# Data Understanding



- **Hubungan DTI Dengan Status Kredit**

Hubungan antara DTI dan status kredit menunjukkan tren bahwa Semakin tinggi DTI, semakin besar risiko gagal bayar atau keterlambatan pinjaman.

Ini berarti DTI bisa menjadi prediktor penting dalam model machine learning untuk menentukan risiko kredit (credit risk prediction).



# Data Pre-processing

Tahap data cleaning dilakukan untuk memastikan dataset siap digunakan dalam analisis dan pemodelan. Berikut tahapan data cleaning :

- Menghapus kolom duplikat agar tidak ada data yang berulang.
- Menghapus kolom dengan missing value > 40% karena dianggap tidak informatif.
- Menghapus baris duplikat untuk menjaga keunikan data.
- Menormalisasi nama kolom dengan huruf kecil dan tanpa spasi atau karakter khusus.
- Membersihkan kolom numerik bertipe string, seperti mengubah “36 months” menjadi 36 dan “10%” menjadi 0.10.
- Mengisi nilai kosong:
- Kolom numerik diisi dengan median.
- Kolom kategorikal diisi dengan modus.
- Menghapus kolom yang tidak relevan seperti id, member\_id, url, title, dan policy\_code.



# Data Encoding

Dilakukan encoding data kategorikal agar bisa digunakan oleh model machine learning.

- Ditemukan 17 kolom kategorikal yang perlu diubah ke bentuk numerik:  
['grade', 'sub\_grade', 'emp\_length', 'home\_ownership', 'verification\_status', 'issue\_d', 'loan\_status',  
'pymnt\_plan', 'purpose', 'zip\_code', 'addr\_state', 'earliest\_cr\_line', 'initial\_list\_status',  
'last\_pymnt\_d', 'last\_credit\_pull\_d', 'application\_type', 'loan\_status\_clean']
- Label Encoding diterapkan pada kolom ordinal:  
grade, sub\_grade.
- Frequency Encoding diterapkan pada kolom dengan banyak kategori:  
issue\_d (90 nilai unik), zip\_code (578), addr\_state (49), earliest\_cr\_line (455), last\_pymnt\_d (91),  
dan last\_credit\_pull\_d (91).
- One-Hot Encoding diterapkan pada kolom kategorikal lain yang jumlah kategorinya sedikit  
(misalnya emp\_length, home\_ownership, purpose, dll).

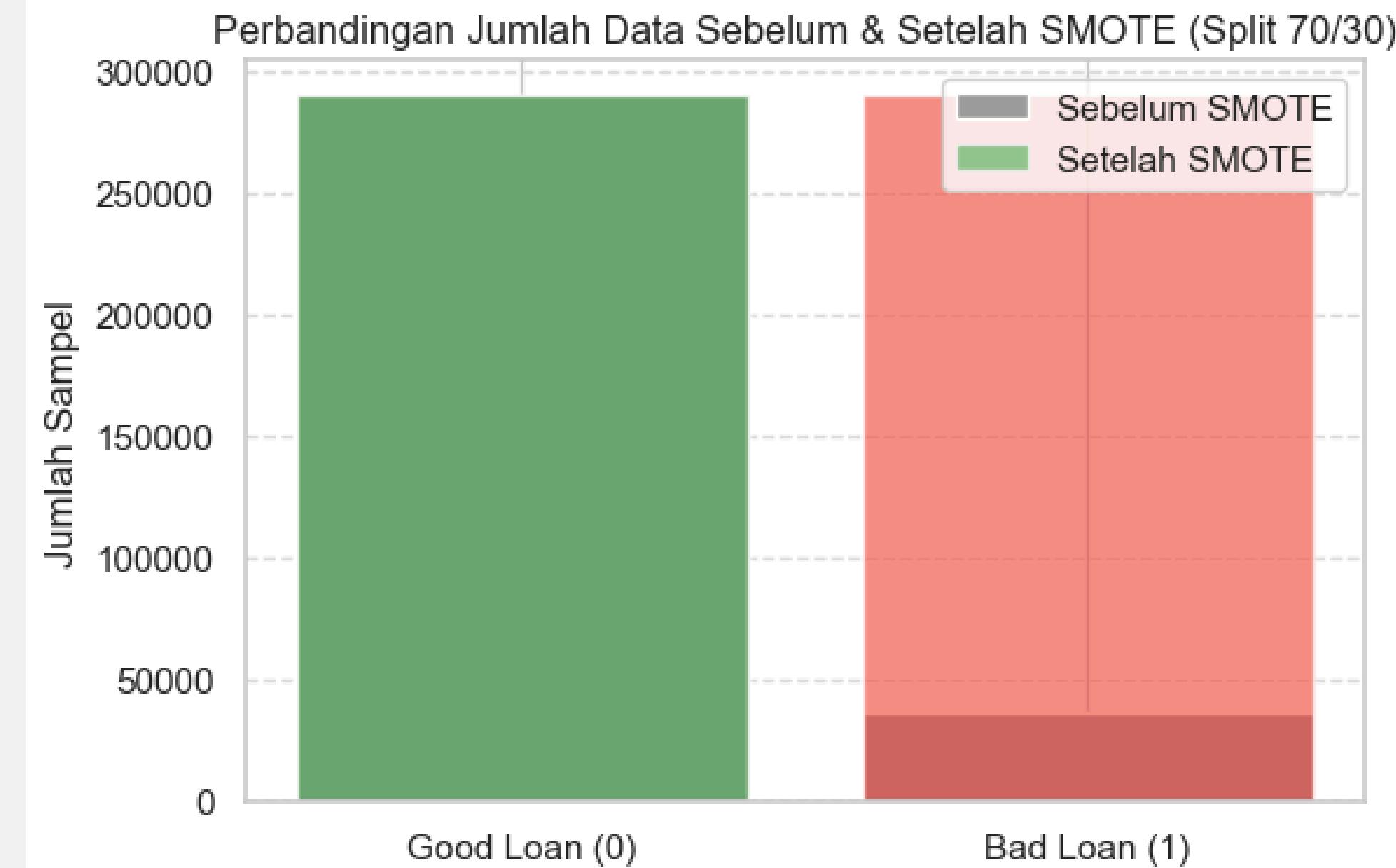
Jumlah kolom meningkat menjadi 89 kolom setelah semua encoding dilakukan, menandakan data sudah siap untuk digunakan pada tahap pemodelan machine learning karena seluruh fitur kini berbentuk numerik.

# Smote dan Data traíníng

- Kelas 0 (Good Loan) berjumlah 289.869 data
- Kelas 1 (Bad Loan) hanya 36.530 data

Sebagian besar data adalah pinjaman yang baik (Good Loan), sedangkan pinjaman bermasalah (Bad Loan) jauh lebih sedikit. SMOTE bekerja dengan membuat data sintetis baru untuk kelas minoritas (Bad Loan) berdasarkan tetangga terdekatnya (nearest neighbors). Data palsu ini bukan duplikasi, melainkan hasil interpolasi antar sampel minoritas.

- 70% data digunakan untuk pelatihan (training)
- 30% data digunakan untuk pengujian (testing)





# Modelling

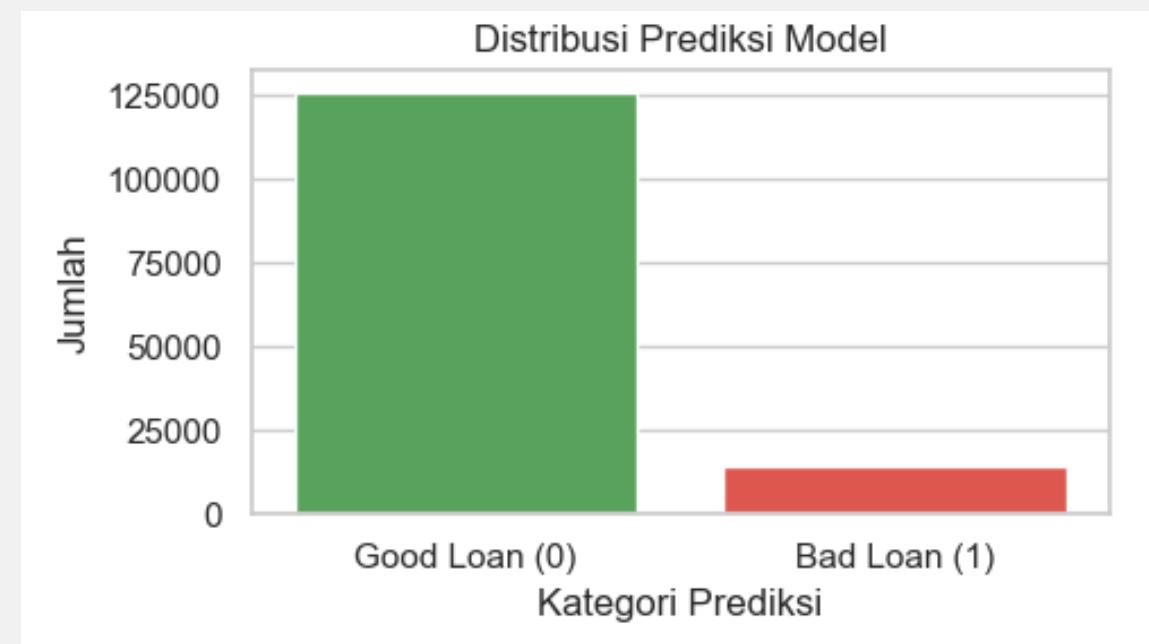
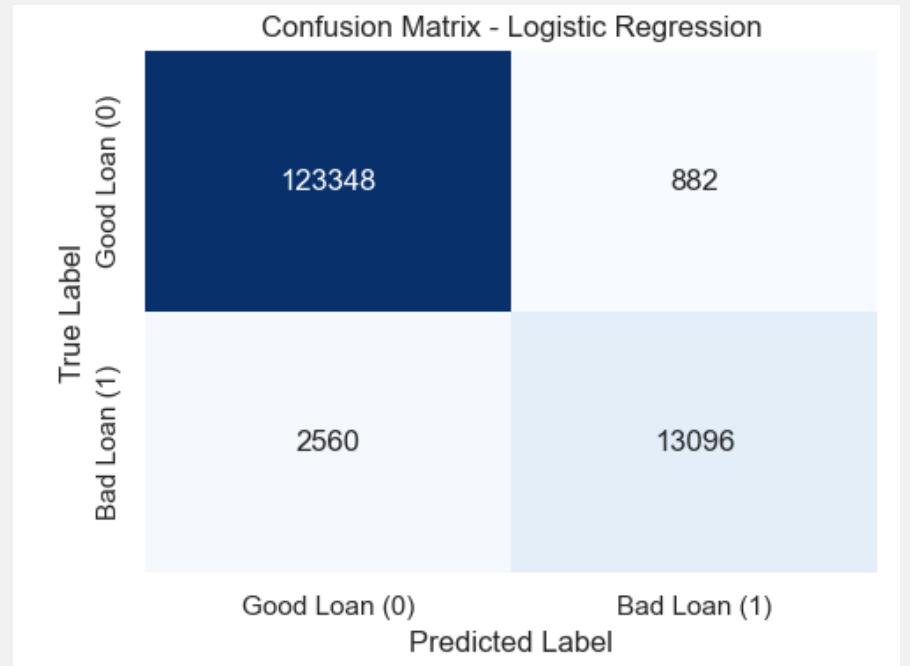
## Logistic Regression

“Setelah penyeimbangan data dengan SMOTE, model Logistic Regression berhasil mencapai akurasi 98%, dengan kemampuan mendeteksi pinjaman bermasalah sebesar 84%. Ini menunjukkan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga cukup andal dalam mengidentifikasi risiko kredit.”

Classification Report:					
	precision	recall	f1-score	support	
0	0.98	0.99	0.99	124230	
1	0.94	0.84	0.88	15656	
accuracy			0.98	139886	
macro avg	0.96	0.91	0.94	139886	
weighted avg	0.97	0.98	0.97	139886	



# Evaluasi



## Confusion Matrix

Model Logistic Regression memberikan hasil yang sangat baik dengan akurasi 98%. Dari confusion matrix terlihat bahwa model sangat akurat dalam memprediksi nasabah baik, dan cukup mampu mengenali nasabah berisiko tinggi. Distribusi prediksi juga menunjukkan keseimbangan yang baik setelah proses SMOTE, menandakan model berhasil belajar pola dari kedua kelas secara efektif.



# TERIMA KASIH

[LINK GITHUB](#)