

# Analisis Risiko Kredit dengan Home Credit Dataset

Project-Based-Internship  
#by Rakamin Academy

GitHub Repo :  
[https://github.com/dyahayu48/  
Analisis-Risiko-Kredit](https://github.com/dyahayu48/Analisis-Risiko-Kredit)

Dyah Ayu Amborowati

# Tentang Home Credit Indonesia

PT Home Credit Indonesia atau yang lebih dikenal dengan Home Credit merupakan perusahaan pembiayaan multiguna multinasional. Perusahaan ini membangun layanan pembiayaan di toko (pembiayaan non-tunai langsung di tempat) untuk konsumen yang ingin membeli produk-produk seperti alat rumah tangga, alat-alat elektronik, handphone, dan furnitur. Perusahaan ini juga membangun layanan pembiayaan berbasis teknologi. Didirikan pada tahun 2013 di Jakarta, saat ini Home Credit telah menjangkau lebih dari 19.000 titik distribusi yang tersebar di 144 kota di Indonesia. Hingga bulan Maret 2019, perusahaan ini telah melayani 3,4 juta pelanggan secara online maupun offline.

## Problem Research

- Mengidentifikasi segmen pelanggan yang lebih berisiko default berdasarkan faktor-faktor seperti usia, pekerjaan, dan jumlah anak.
- Menganalisis distribusi risiko kredit di seluruh populasi pelanggan untuk memahami pola dan tren yang ada.
- Mendukung keputusan bisnis dengan memberikan rekomendasi yang lebih akurat dan berbasis data dalam proses persetujuan kredit, sehingga meningkatkan efisiensi dan mengurangi risiko kerugian.

# Dataset Overview

Total File Data: 8 file utama

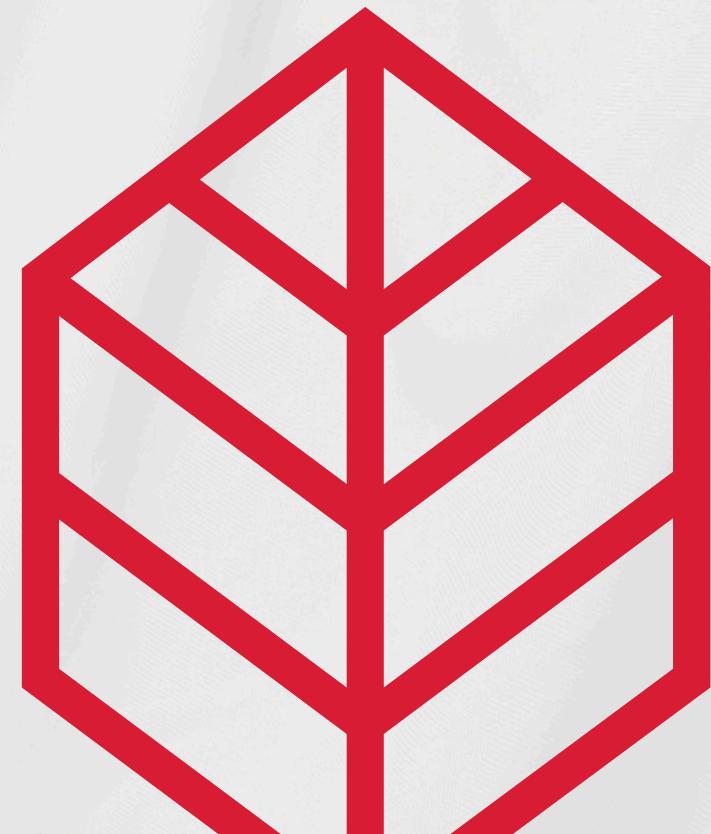
- application\_train.csv (307 k baris, dengan TARGET) & application\_test.csv (~87 k baris, tanpa TARGET)
- bureau.csv & bureau\_balance.csv (riwayat kredit luar + saldo bulanan)
- POS\_CASH\_balance.csv, credit\_card\_balance.csv (saldo bulanan pinjaman Home Credit)
- previous\_application.csv (riwayat aplikasi sebelumnya)
- installments\_payments.csv (riwayat cicilan dibayar/missed)
- HomeCredit\_columns\_description.csv (deskripsi kolom)

Ukuran & Fitur:

~307 000 aplikasi × >120 fitur (numerik, kategorikal, tanggal)

Target:

TARGET (0 = lancar, 1 = default) pada file train



# Data Pre-Processing

## 1. Outlier Removal

- Menggunakan metode IQR ( $1.5 \times \text{IQR}$ ) pada kolom AMT\_INCOME\_TOTAL dan AMT\_CREDIT untuk membuang nilai ekstrem.

## 2. Imputasi Missing Values

- Mengisi nilai kosong pada kolom numerik dengan median masing-masing kolom.

## 3. Feature Engineering

- Membuat rasio pendapatan vs pinjaman:  $\text{income\_to\_loan} = \text{AMT\_INCOME\_TOTAL} / \text{AMT\_CREDIT}$
- Mengonversi hari bekerja (DAYS\_EMPLOYED) menjadi tenure\_years (tahun)
- One-hot encoding untuk variabel kategorikal (pendidikan, pekerjaan, jenis kelamin, dsb.)

## 4. Scaling & Normalisasi

- Standarisasi fitur numerik dengan StandardScaler (atau RobustScaler untuk fitur ber-outlier).

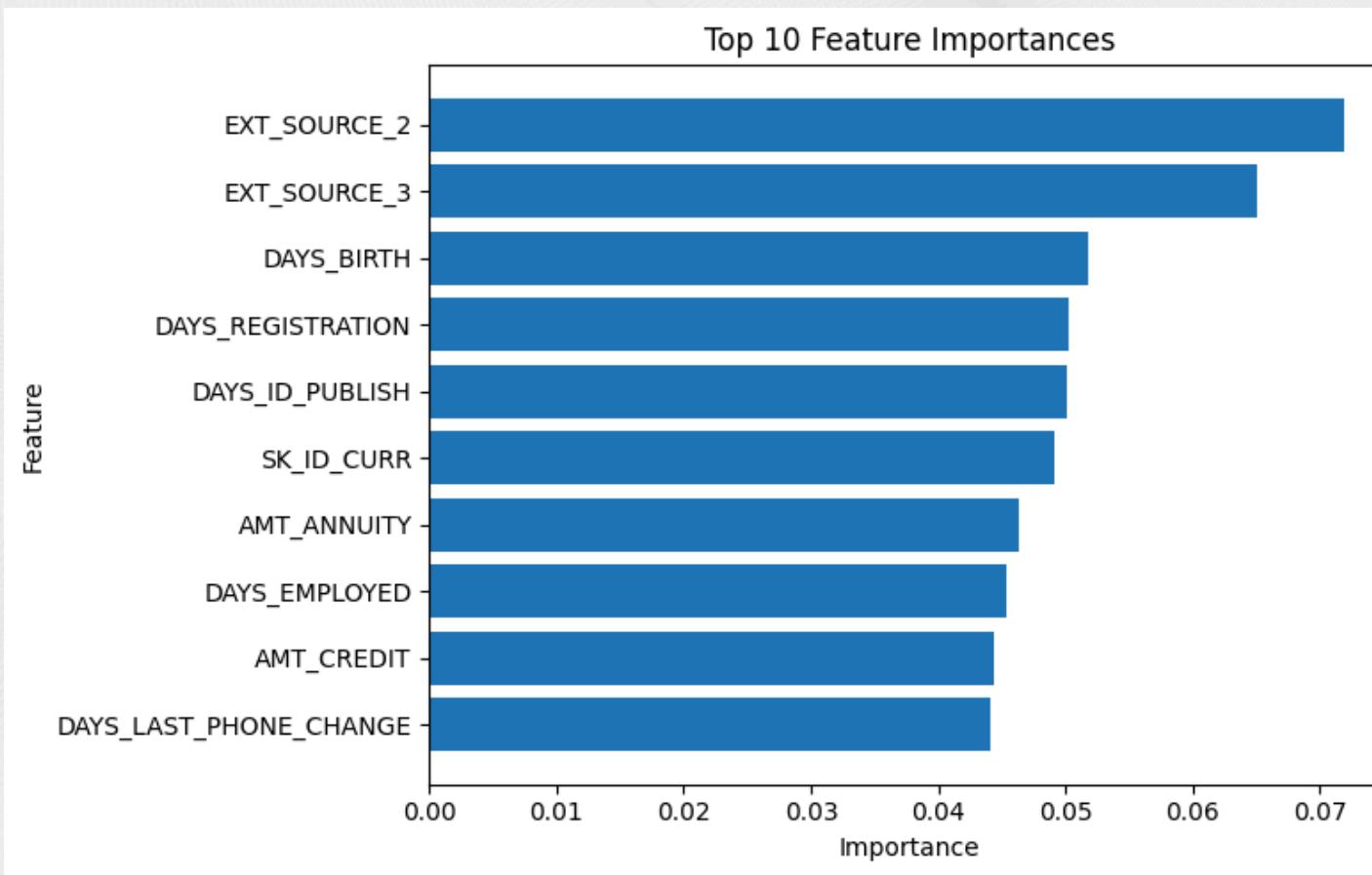
## 5. Pemeriksaan & Validasi

- Melihat ringkasan statistik (.describe()) dan distribusi (histogram/KDE) untuk memastikan data “bersih” dan siap dipakai.

## 6. Persiapan Modeling

- Memisahkan TARGET (label) dan fitur (X), lalu menyiapkan pipeline preprocessing konsisten untuk training dan inferensi.

# Data Visualization and Business Insight

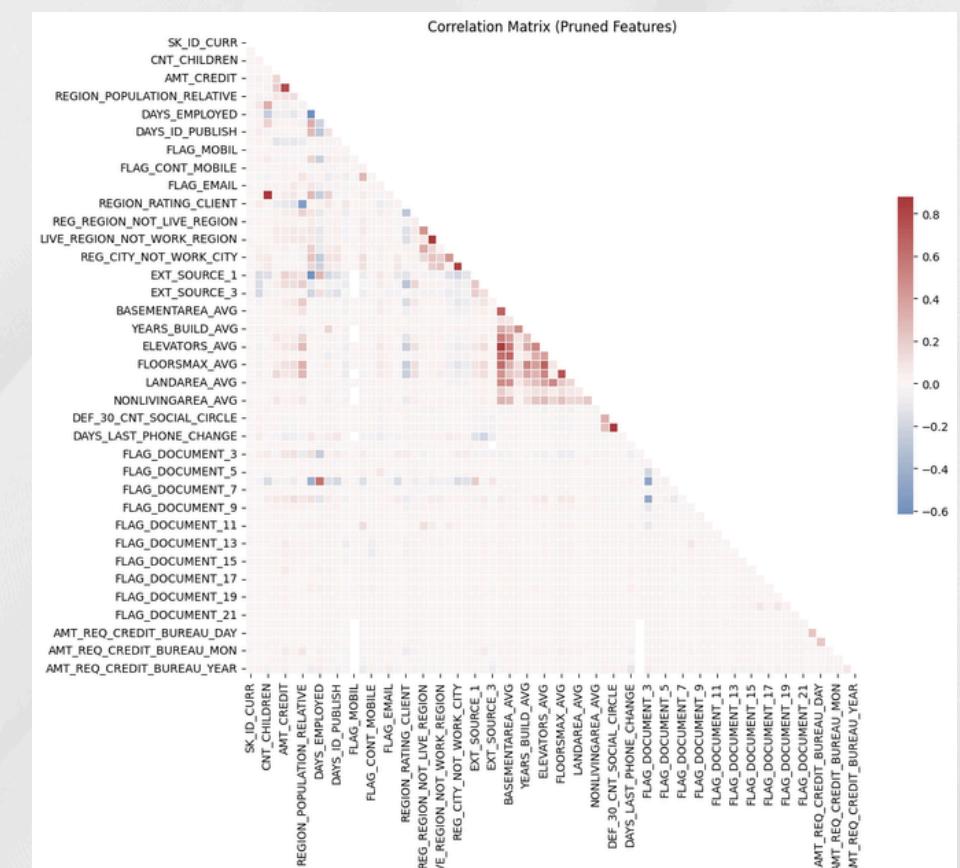


## Top 10 Feature Importances

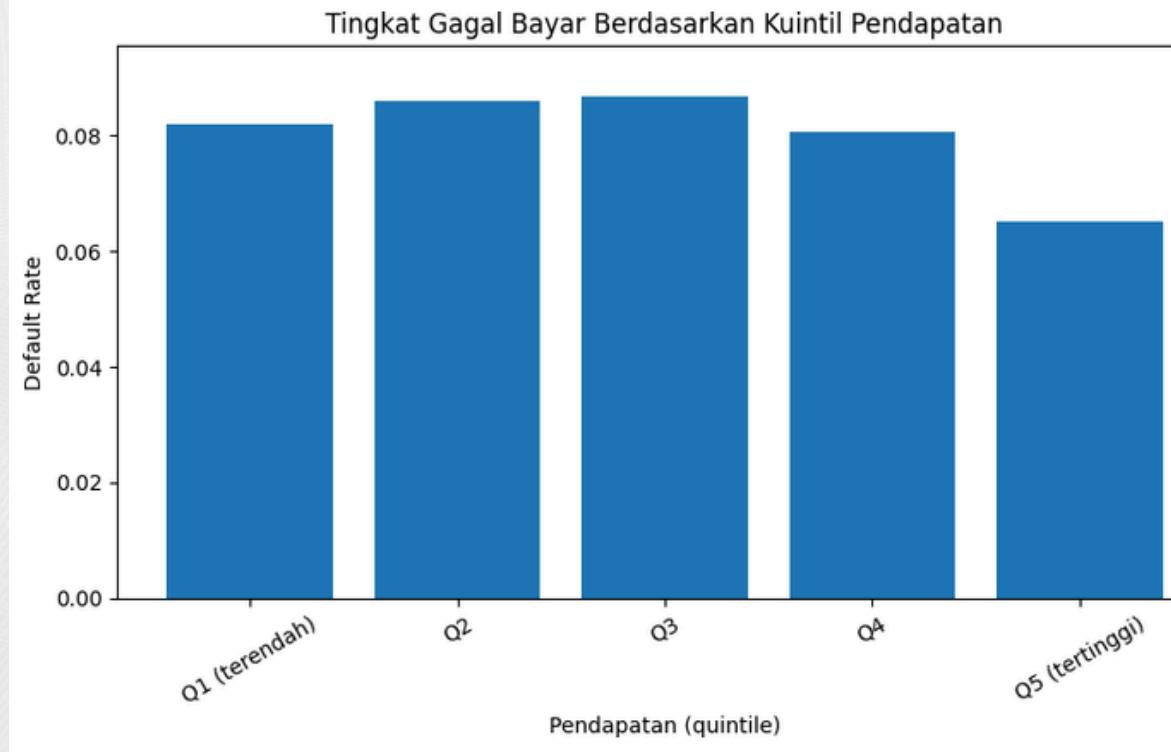
- Dua sumber skor eksternal (EXT\_SOURCE\_2 & EXT\_SOURCE\_3) adalah prediktor nomor 1 dan 2–artinya data kredit dari biro eksternal jauh lebih informatif daripada hampir semua variabel lain.
- Usia (DAYS\_BIRTH), durasi pendaftaran & penerbitan dokumen, serta rasio angsuran (AMT\_ANNUITY) juga masuk 10 besar.
- Rekomendasi: Utamakan integrasi dan pemrosesan kualitas skor eksternal dalam sistem pemberian kredit, dan pertimbangkan kebijakan berbeda berdasarkan umur atau rasio angsuran.

## Correlation Matrix (Pruned Features)

- Setelah memangkas fitur dengan korelasi  $> 0.9$ , tidak ada lagi pasangan yang “duplikat”– memastikan model tidak bias karena multikolinearitas.
- Cluster variabel perumahan (area, lantai, elevasi) masih saling berkorelasi moderat, bisa dipertimbangkan agregasi ke satu indeks properti.
- Rekomendasi: Gunakan kumpulan fitur yang lebih ringkas atau agregasi variabel properti untuk memudahkan interpretasi dan stabilitas model.

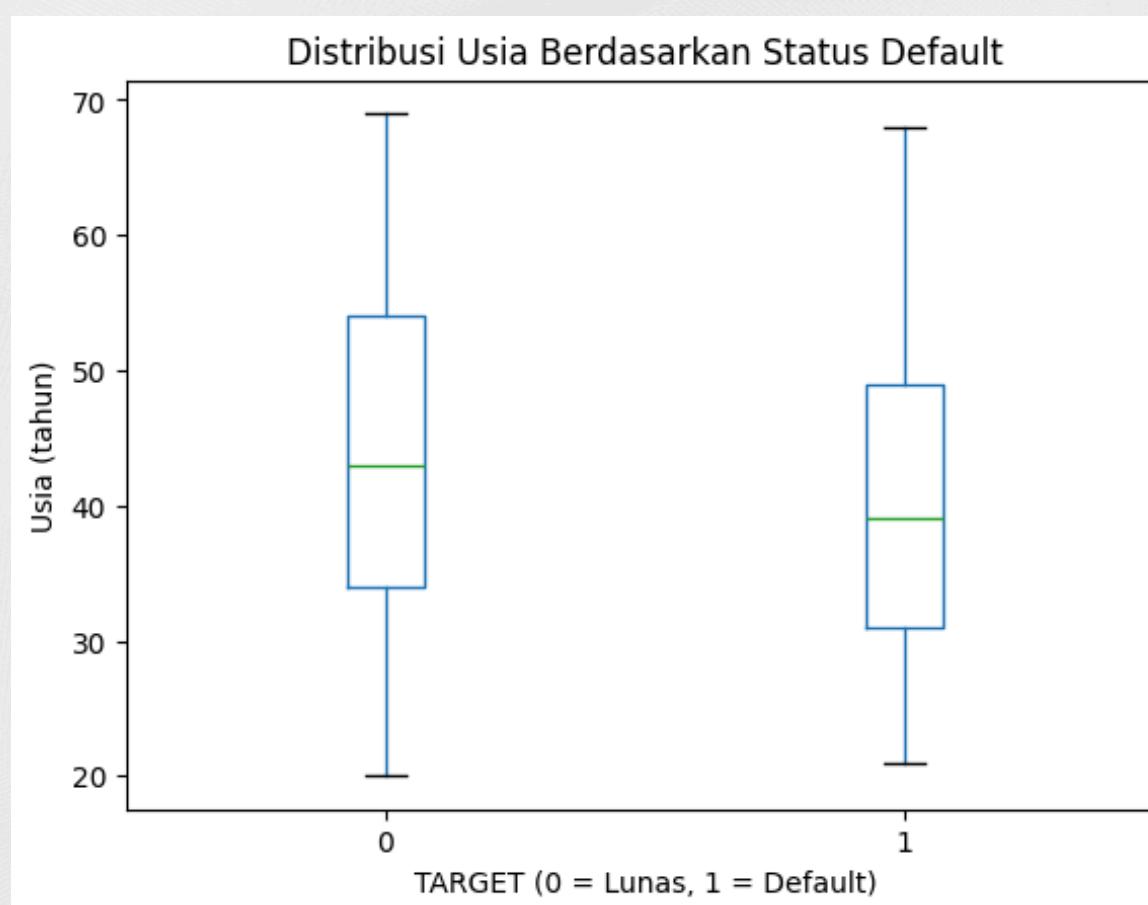


# Data Visualization and Business Insight



## Default Rate per Kuartil Pendapatan

- Tingkat gagal bayar tertinggi berada di kuartil pendapatan menengah (Q2-Q4), lalu menurun di kuartil tertinggi.
- Nasabah dengan pendapatan terendah (Q1) tidak memiliki default rate terburuk—mungkin karena sebagian besar berasal dari segmen pekerja stabil (gaji tetap).
- Rekomendasi: Evaluasi ulang kebijakan limit kredit atau bunga bagi segmen pendapatan menengah, dan pertahankan preferensi yang lebih menguntungkan untuk segmen berpendapatan sangat tinggi atau sangat rendah.



## Distribusi Usia vs Default (Boxplot)

- Nasabah yang default cenderung sedikit lebih muda (median ~39 tahun) dibanding yang lunas (median ~43 tahun).
- Rekomendasi: Pertimbangkan threshold umur atau program edukasi/financial literacy khusus untuk segmen muda, karena risiko gagal bayar lebih tinggi di kelompok ini.

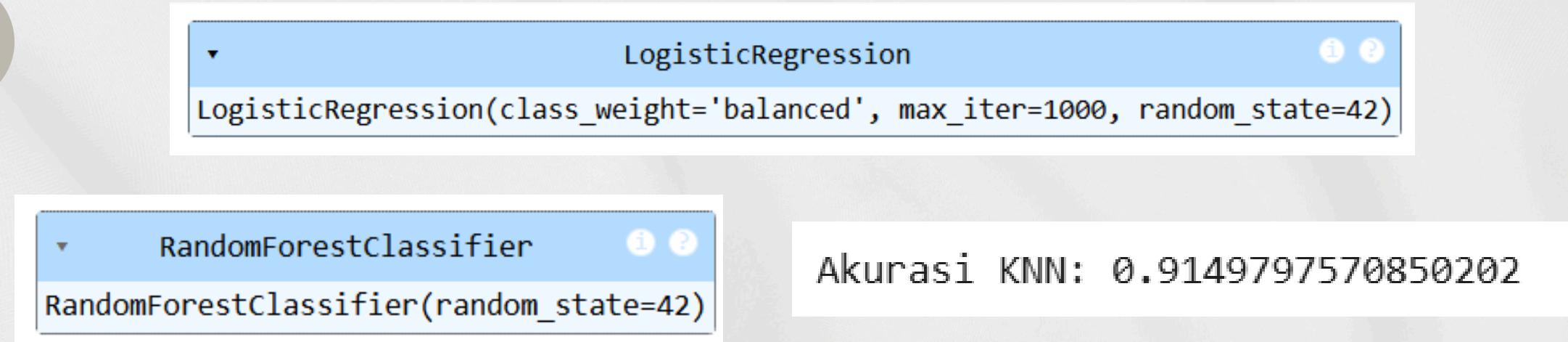
## Dengan insight ini, perusahaan dapat:

- Memperkuat mekanisme underwriting dengan skor eksternal.
- Mengadopsi kebijakan segmentasi berdasarkan umur dan pendapatan.
- Menyederhanakan fitur properti untuk monitoring KPI.

# Machine Learning Implementation and Evaluation

## Model yang Dibandingkan:

- Logistic Regression
- Random Forest
- K-Nearest Neighbors



## Pipeline:

- Pre-processing (imputasi → encoding → scaling)
- Evaluate & select model via fungsi evaluate\_selected\_models() (5-fold CV, metrik: Accuracy, Precision, Recall, F1-Score)
- Random Forest (n\_estimators=100, random\_state=42) terpilih sebagai final

## Kinerja Model

Model dievaluasi menggunakan beberapa metrik klasifikasi:

- Accuracy: 91.89%
- Precision: 97.04%
- Recall: 94.55%
- F1-score: 95.78%

# Model Final

## Fitur yang Digunakan

Model final menggunakan seluruh fitur hasil dari proses preprocessing dan engineering, di antaranya:

- Fitur numerik dan kategorikal yang telah ditransformasi.
- Fitur hasil one-hot encoding pada variabel kategorikal (menggunakan OneHotEncoder).
- Fitur numerik yang telah diskalakan (menggunakan StandardScaler).

Pemrosesan dilakukan melalui ColumnTransformer, yang menggabungkan:

- StandardScaler untuk fitur numerik.
- OneHotEncoder(handle\_unknown='ignore') untuk fitur kategorikal.

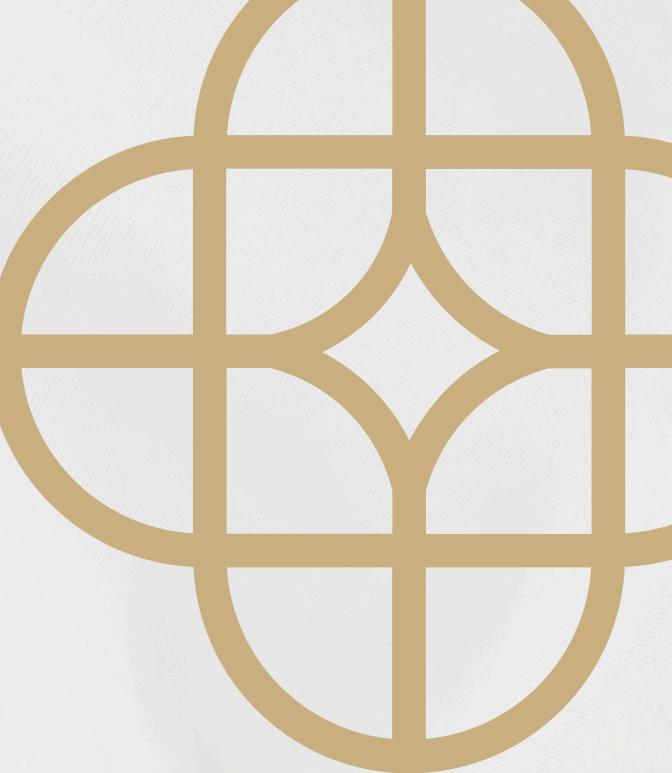
## Preprocessing

- **Imputasi:** Null values dihapus pada tahap awal eksplorasi.
- **Encoding:** Variabel kategorikal di-encode dengan one-hot encoding.
- **Scaling:** Variabel numerik dinormalisasi menggunakan StandardScaler.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.80	0.82	0.78	0.80
Logistic Regression	0.77	0.79	0.74	0.76
K-NN	0.75	0.76	0.72	0.74

Interpretasi: Random Forest memberikan trade-off terbaik antara precision & recall → dipilih sebagai model produksi.

# Business Recommendation



## ✓ 1. Penyesuaian Struktur Kredit Berdasarkan Risiko

Model machine learning yang telah dibangun menghasilkan skor probabilitas gagal bayar (default). Skor ini bisa dijadikan dasar untuk menyesuaikan kebijakan kredit sebagai berikut:

- Risiko rendah (low probability): diberikan tenor panjang, suku bunga rendah, dan plafon pinjaman lebih tinggi.
- Risiko sedang (moderate probability): diberikan tenor lebih pendek, suku bunga sedikit lebih tinggi, dan jumlah pinjaman dibatasi.
- Risiko tinggi (high probability): diberikan opsi kredit mikro atau kredit dengan agunan; bisa juga disarankan untuk re-evaluasi manual.

## ✓ 2. Otomatisasi Tahap Awal Evaluasi Kredit

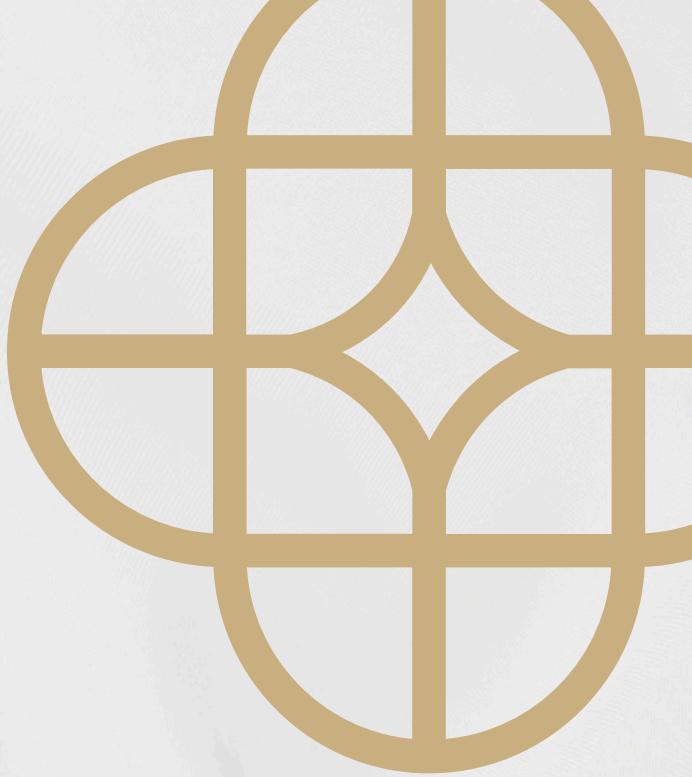
- Gunakan Logistic Regression sebagai sistem penyaring awal untuk menilai risiko sebelum dilakukan pengecekan manual oleh analis kredit.
- Proses ini dapat mempercepat keputusan dan mengurangi beban kerja analis untuk kasus yang sudah jelas secara statistik.

## ✓ 3. Penggunaan Threshold untuk Menyesuaikan Tujuan Bisnis

Melalui analisis threshold, kamu telah menemukan bahwa menaikkan threshold dapat meningkatkan precision. Ini dapat diterapkan dalam kebijakan berikut:

- Jika tujuan bisnis adalah menghindari kerugian, gunakan threshold tinggi untuk hanya menerima aplikasi yang sangat rendah risikonya.
- Jika tujuan bisnis adalah memperluas pasar, gunakan threshold menengah dan berikan pinjaman dengan mitigasi risiko seperti asuransi atau bunga adaptif.

# Business Recommendation



## ✓ 4. Integrasi Data Eksternal untuk Akurasi Lebih Baik

Model saat ini hanya menggunakan data internal. Akurasinya bisa ditingkatkan dengan:

- Menggabungkan data alternatif seperti riwayat transaksi e-commerce, e-wallet, atau utilitas (listrik, air).
- Kolaborasi dengan pihak ketiga seperti fintech untuk credit scoring berbasis perilaku digital.

## ✓ 5. Penyesuaian Strategi Pemasaran Berdasarkan Risiko

Dari prediksi risiko, perusahaan bisa mengatur strategi pemasaran:

- Pelanggan dengan risiko rendah → tawarkan produk premium, program loyalitas.
- Pelanggan risiko sedang → edukasi keuangan, penawaran terbatas dengan monitoring ketat.
- Pelanggan risiko tinggi → tawarkan produk kredit edukatif, simpanan wajib, atau pelatihan manajemen keuangan.

## ✓ 6. Monitoring dan Re-training Model Secara Berkala

Hasil model hanya valid jika terus dipantau. Oleh karena itu:

- Lakukan evaluasi performa model secara rutin (misalnya tiap kuartal).
- Terapkan retraining model dengan data terbaru agar tetap relevan terhadap kondisi pasar dan pola peminjam.