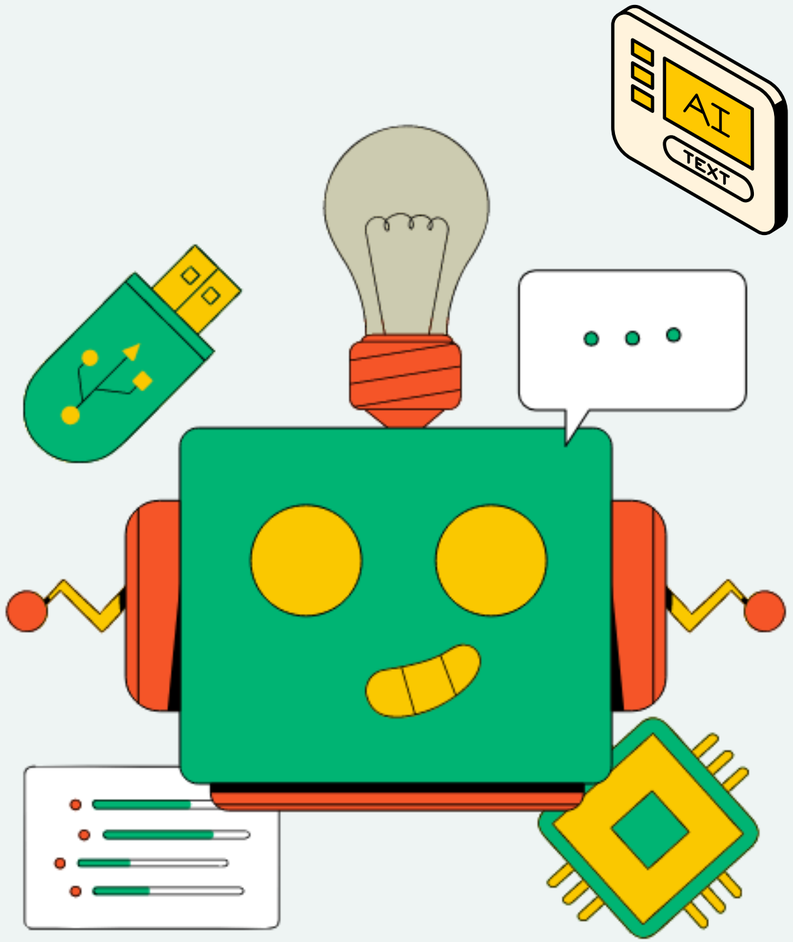
DIBIMBING.ID

WE LEARN FOR THE FUTURE



CREDIT APPROVAL PREDICTION USING CLASSIFICATION MODEL

BATCH 32B | BOOTCAMP DATA SCIENCE AND DATA ANALYST

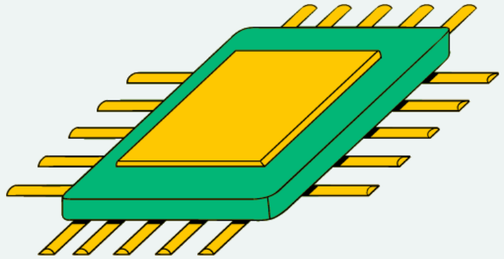


TABLE OF CONTENTS

INTRODUCTION

DIMAS ADI PRASETYO

Student

DATA SCIENTIST & DATA ANALYST

Experience

- Aug 2023 - Dec 2023 — **Game Development**, Infinite Learning Indonesia
- Feb 2023 - Jul 2023 — **Machine Learning**, Bangkit Academy

Education

- Present - Feb 2025 — **Data Science Bootcamp**, dibimbing.id
- Aug 2024 - Jul 2020 — **Bachelor of Science in Informatics Engineering**, Universitas Krisnadwipayana

PREVIOUS PROJECTS

People Analytics (10 – 16 May 2025)

This project investigates employee job satisfaction using survey data collected from various departments within an organization. The goal is to uncover insights into the factors that influence satisfaction and to provide strategic recommendations for improving employee well-being. It aims to identify patterns and insights related to job satisfaction, work-life balance, workload, and training, and to present the findings through interactive visualizations and business recommendations.

https://github.com/Dadipp/People_Analytics

Customer Satisfaction & Sentiment Analysis (3 – 8 May 2025)

This project analyzes customer feedback data using sentiment analysis and CSAT/NPS metrics to evaluate service satisfaction. Key insights include customer loyalty trends, issue categories, and overall service perception over time. All findings are visualized through Power BI dashboards.

https://github.com/Dadipp/Customer_Satisfaction_and_Sentiment_Analysis

MAIN PROJECT

CRISP-DM

business understanding

data understanding

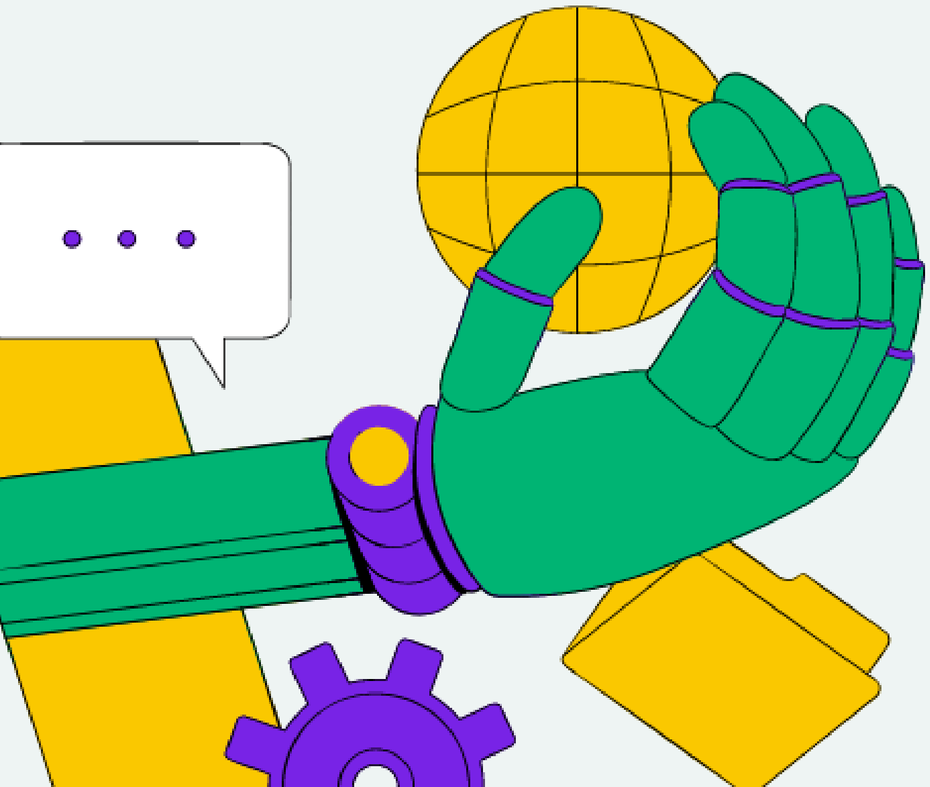
data preparation

modeling

evaluation

deployment

PROJECT BACKGROUND



Menganalisis kelayakan kredit dari pemohon kartu kredit dengan mengevaluasi dari histori pembayaran dan finansial mereka, guna mengidentifikasi individu yang berisiko tinggi dan mempercepat proses persetujuan. Hal ini memungkinkan untuk mengurangi risiko gagal bayar, meningkatkan efisiensi operasional, dan mengoptimalkan kinerja portofolio kredit.

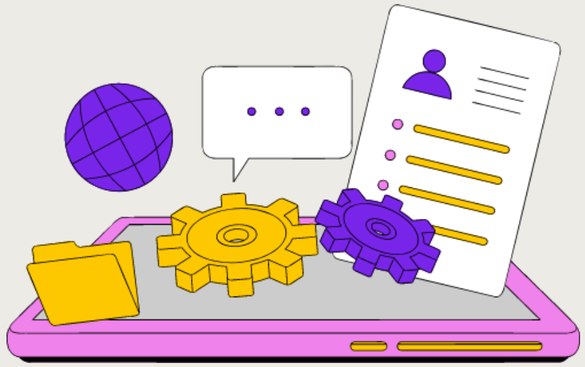
BUSINESS OBJECTIVE

Main Objective

Proyek ini bertujuan untuk membangun model prediksi kelayakan kredit yang membantu lembaga keuangan dalam menilai kelayakan pemohon kartu kredit secara otomatis. Tujuannya adalah untuk mengurangi risiko gagal bayar, mempercepat proses persetujuan, dan mendukung pengambilan keputusan berbasis data.

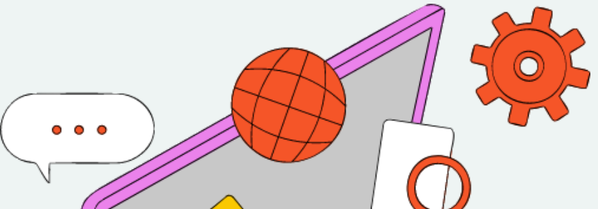
Specific Objective

1. Menganalisis fitur yang mempengaruhi kelayakan kredit.
2. Mengklasifikasikan pemohon ke dalam kategori layak atau tidak layak berdasarkan profil risikonya.
3. Mengidentifikasi pemohon berisiko tinggi sejak awal untuk mengurangi risiko gagal bayar.
4. Mengevaluasi performa model klasifikasi Logistic Regression, Random Forest, dan XGBoost dalam memprediksi risiko kredit.

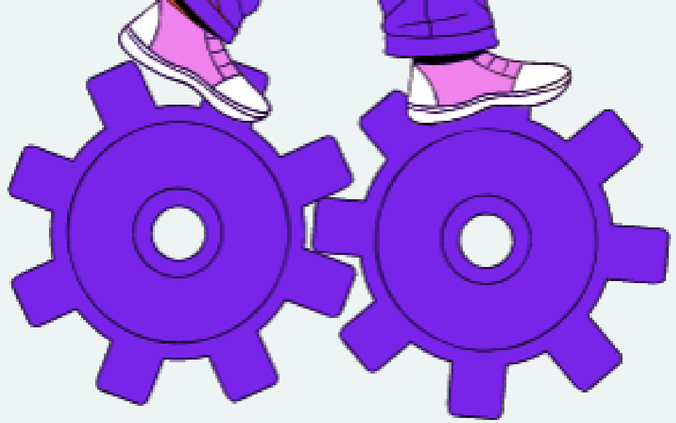


DATA UNDERSTANDING

- application_record.csv (data fitur) : 438,557 index dan 18 kolom. Berisi informasi seperti jenis kelamin, status pernikahan, jumlah anggota keluarga, status pekerjaan, dan pendapatan.
- credit_record.csv (data label) : 1,042,558 index dan 3 kolom. Berisi informasi histori/riwayat, status pembayaran (0–5), dan ID peminjam.
- Kedua dataset digabung berdasarkan ID agar setiap pemohon memiliki fitur dan label (approved atau rejected).



DATA PREPARATION



HANDLING MISSING VALUES

Kolom OCCUPATION_TYPE memiliki sekitar 30% missing values, kolom ini sepertinya mengandung informasi penting terkait profil pekerjaan pemohon. Saya memilih pendekatan imputasi berbasis modus. Sementara itu, kolom FLAG_MOBIL dihapus karena seluruh nilainya konstan, tidak memberikan informasi atau variasi apapun, dan hanya menambah dimensi data secara tidak perlu

REMOVING DUPLICATE

Tidak ditemukan duplicated pada dataset

OUTLIER CHECK

Kolom DAYS_EMPLOYED mencatat durasi masa kerja dalam hari. Observasi menunjukkan adanya nilai ekstrem 365243, yang diinterpretasikan sebagai indikator pemohon yang telah pensiun. Nilai ini dipertahankan karena dianggap sebagai informasi yang relevan dan bukan outlier yang bersifat anomali. Karena DAYS_EMPLOYED akan di konversi menjadi YEARS_EMPLOYED

TRANSFORMASI DATA

Beberapa transformasi data dilakukan untuk meningkatkan interpretabilitas dan kualitas data. Kolom DAYS_BIRTH diubah menjadi AGE dalam satuan tahun agar lebih mudah dipahami dan digunakan dalam analisis. Selain itu, kolom CNT_FAM_MEMBERS yang semula bertipe float dibulatkan dan dikonversi menjadi integer

LABELING & FEATURE ENGINEERING

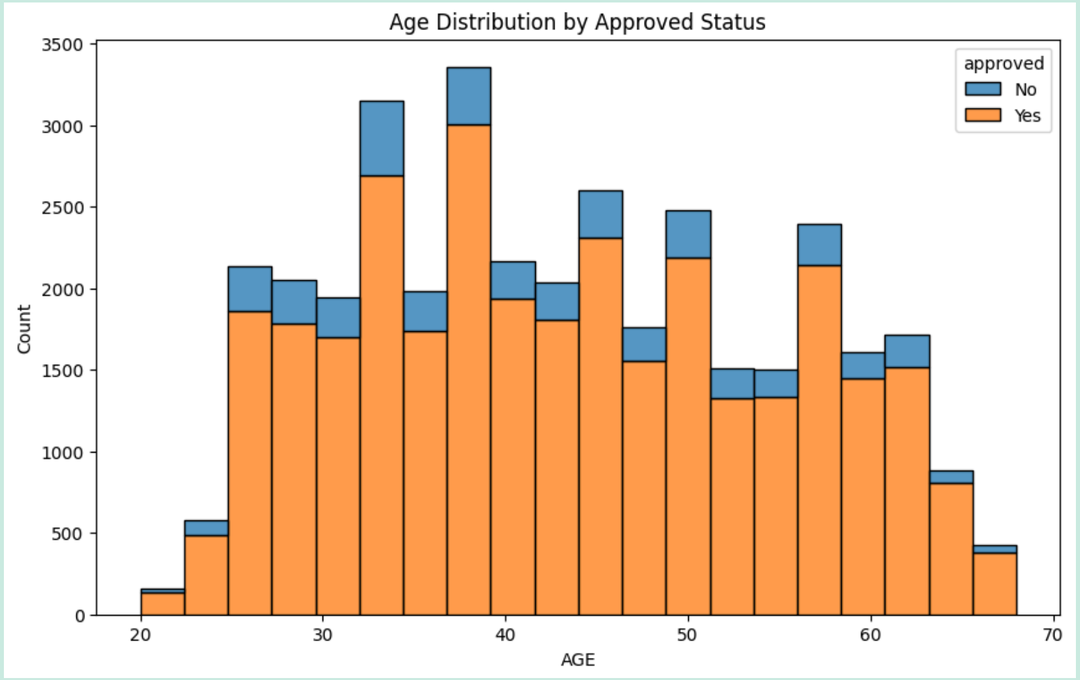
LABELING & MERGE DATA

Label persetujuan kredit (approved) ditentukan berdasarkan riwayat kredit pemohon, di mana pemohon dianggap disetujui jika tidak memiliki keterlambatan lebih dari 1 bulan selama periode observasi. Dari data riwayat kredit, dihitung agregasi statistik seperti MB_min, MB_max, MB_mean, dan MB_count untuk merepresentasikan durasi dan intensitas aktivitas kredit.

FEATURE ENGINEERING

Beberapa kolom kategorikal seperti CODE_GENDER, FLAG_OWN_CAR, dan FLAG_OWN_REALTY diubah menjadi bentuk numerik dengan binary mapping. Membuat fitur tambahan seperti income_bin dari AMT_INCOME_TOTAL, AGE yang dihitung dari DAYS_BIRTH, age_group. Semua transformasi ini bertujuan untuk mempermudah proses eksplorasi data dan pelatihan model prediktif.

EDA



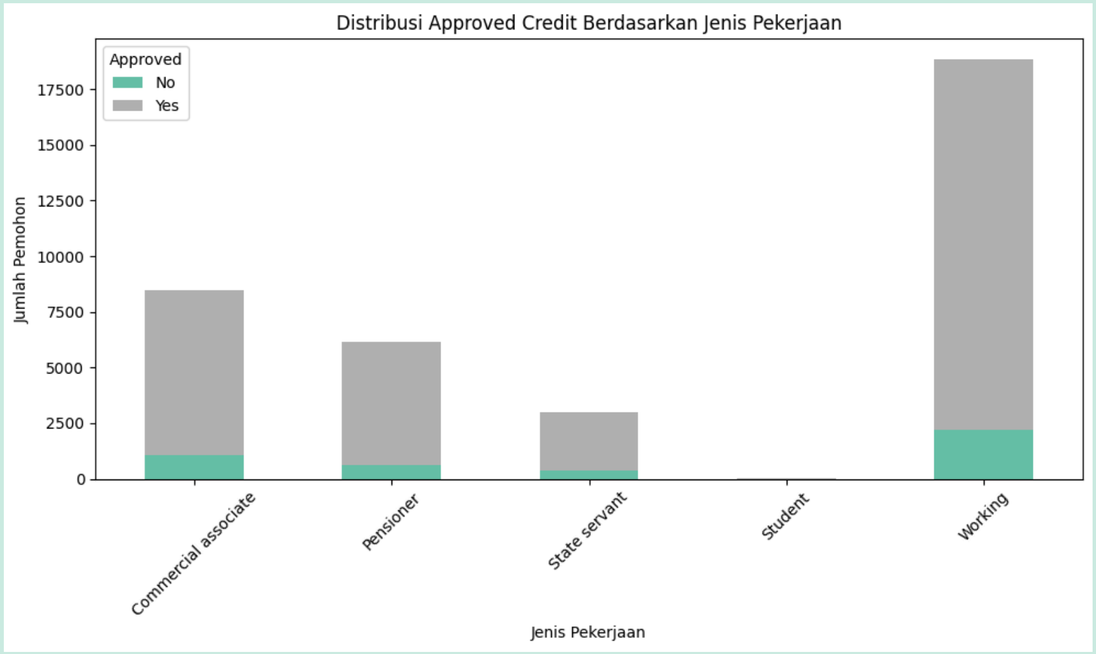
01 AGE DISTRIBUTION BY APPROVED STATUS

Insight dari Distribusi Usia dan Kelayakan Kredit :

- Mayoritas pemohon kredit berusia 30–60 tahun, dengan puncak di usia 35–45 tahun.
- Tingkat persetujuan meningkat seiring usia dari 86.65% (usia 20–30) → 89.34% (usia 60–70).
- Pemohon usia 40+ menunjukkan kelayakan kredit lebih tinggi dan konsisten.

Rekomendasi: Fokus pemasaran dan penawaran produk kredit pada usia 40–60 tahun untuk memaksimalkan tingkat persetujuan.

EDA



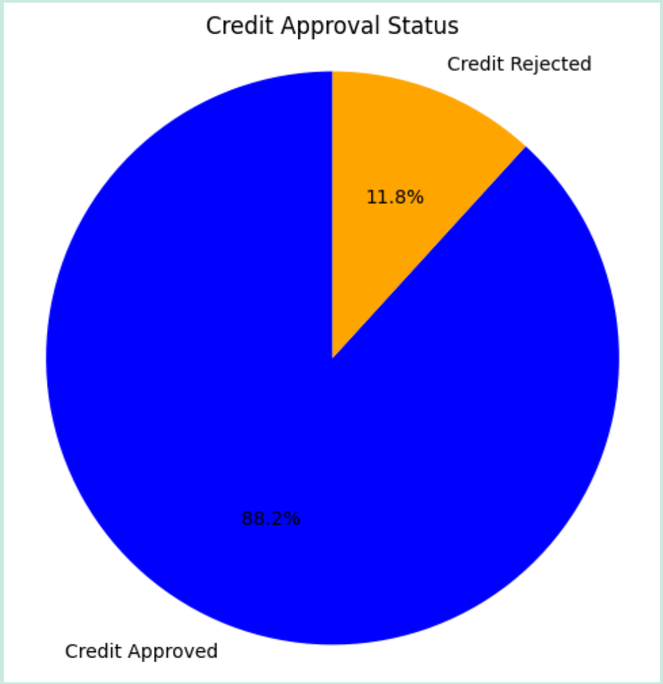
02 DISTRIBUTION OF APPROVED CREDIT BASED ON INCOME TYPE

Insight dari Distribusi Jenis Pekerjaan dan Kelayakan Kredit:

- Mayoritas pemohon berasal dari tipe income Working. Tapi tingkat penolakannya juga cukup tinggi.
- Hampir semua kategori pekerjaan menunjukkan tingkat persetujuan tinggi.
- Jumlah pemohon dan tingkat persetujuan student sangat rendah, kemungkinan karena belum memiliki penghasilan atau riwayat kredit.

EDA

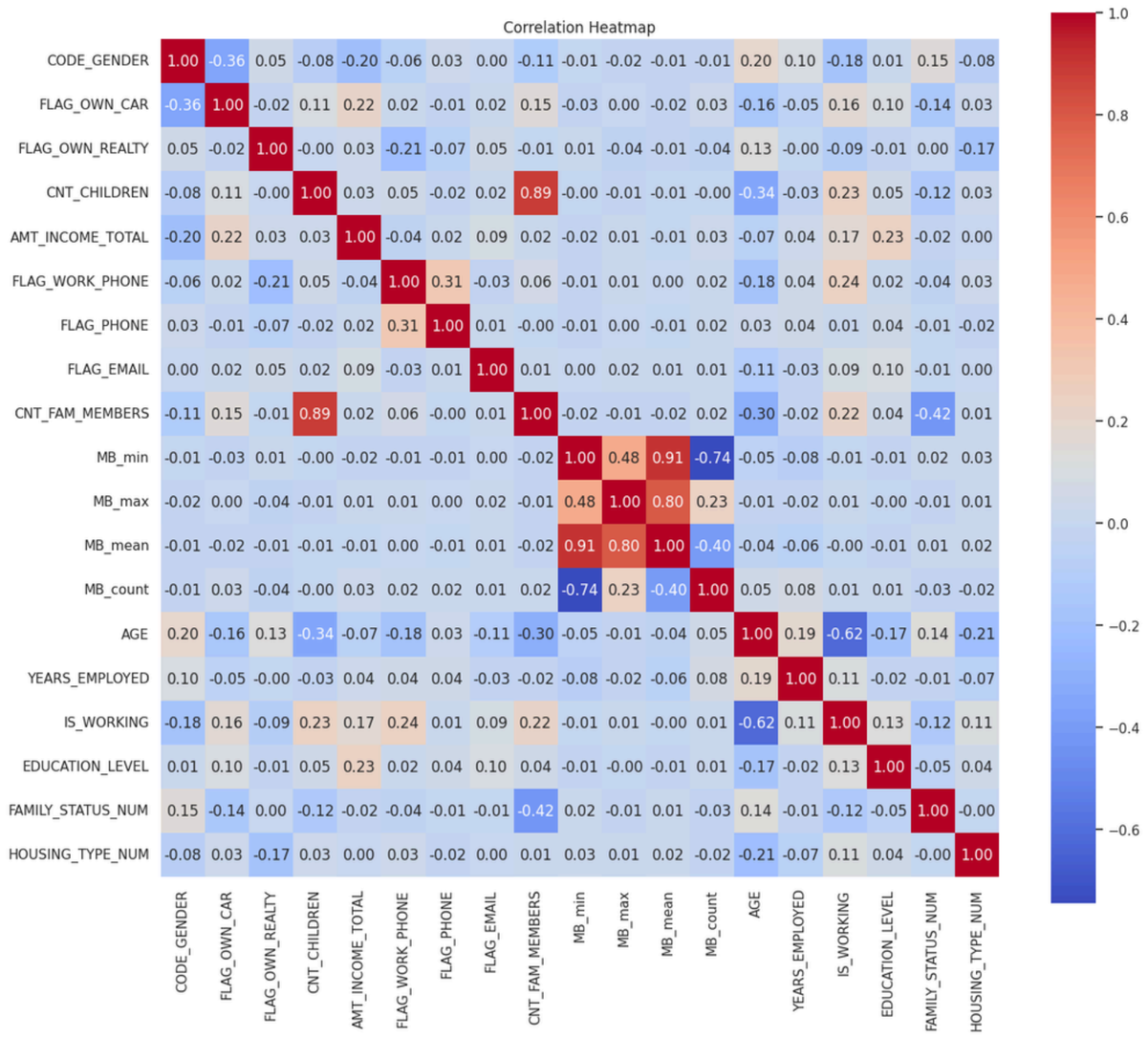
03 CREDIT APPROVAL OVERVIEW



Insight dari Overview Persetujuan Kredit:

- Data menunjukkan ketidakseimbangan kelas: Approved 88.2% vs Rejected 11.8%.
- Ketidakseimbangan ini dapat menyebabkan bias model terhadap kelas mayoritas, menyulitkan prediksi akurat untuk kelas rejected.
- SMOTE digunakan saat preprocessing untuk menyeimbangkan kelas dan meningkatkan kemampuan model dalam mengenali target yang rejected.

FEATURE CORRELATION



Insight dari Analisis Korelasi

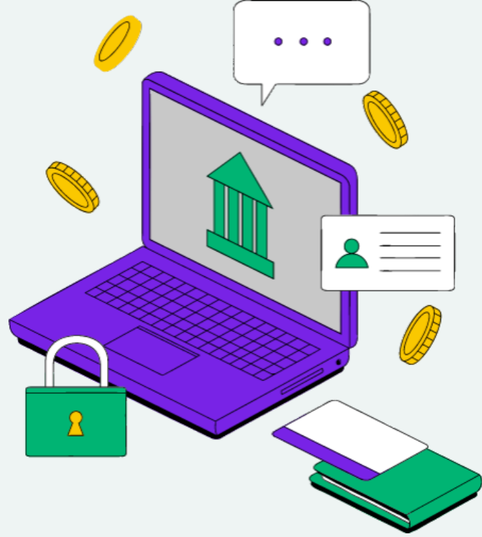
- Beberapa data aktivitas bulanan seperti MB_min, MB_max, dan lainnya saling berkaitan sangat kuat. Untuk menghindari data ganda, dipilih yang mewakili (MB_mean dan MB_count) yang digunakan.
- CNT_CHILDREN memiliki informasi sangat mirip dengan CNT_FAM_MEMBERS. CNT_FAM_MEMBERS dipilih karena lebih informatif.
- Sebagian besar fitur tidak memiliki nilai korelasi yang tinggi. Ini artinya setiap fitur membawa informasi berbeda.
- Terlihat hubungan positif antara usia, lama bekerja, dan pendidikan. Ini bisa jadi tanda bahwa semakin tua dan berpendidikan seseorang, biasanya semakin stabil secara keuangan sehingga lebih mudah disetujui kreditnya.

VIF ANALYSIS

1	CODE_GENDER	1.239939
2	FLAG_OWN_CAR	1.213129
3	FLAG_OWN_REALTY	1.089988
4	AMT_INCOME_TOTAL	1.164929
5	FLAG_WORK_PHONE	1.240083
6	FLAG_PHONE	1.125887
7	FLAG_EMAIL	1.032029
8	CNT_FAM_MEMBERS	1.328319
9	MB_mean	1.196560
10	MB_count	1.205331
11	AGE	2.043106
12	YEARS_EMPLOYED	1.154316
13	IS_WORKING	1.891714
14	EDUCATION_LEVEL	1.104667
15	FAMILY_STATUS_NUM	1.238957
16	HOUSING_TYPE_NUM	1.076453

- Tujuan: Mengecek apakah ada fitur numerik yang terlalu mirip satu sama lain (multikolinearitas).
- Hasil Penting:
  - Semua nilai VIF < 5 → tidak ada multikolinearitas serius.
  - Fitur dengan VIF tertinggi: AGE dan IS_WORKING, tapi masih dalam batas aman.
- Nilai VIF tinggi pada const bisa diabaikan karena bukan fitur prediktor.
- Kesimpulan:
  - Semua fitur numerik aman digunakan dalam pemodelan tanpa perlu penghapusan.

PREPROCESSING: TRAIN-TEST SPLIT, ENCODING & SCALING



Langkah-langkah preprocessing yang dilakukan:

- Target kolom approved diubah menjadi numerik (Yes → 1, No → 0)
- Fitur kategorikal diencoding dengan One-Hot Encoding
- Data dibagi menjadi data training (80%) dan testing (20%)
- Fitur numerik distandarisasi menggunakan StandardScaler
- Fitur yang tidak relevan seperti ID telah dihapus sebelumnya dan fitur age_group dan income_bin yang dibuat untuk keperluan EDA juga dihapus.

MODELING OVERVIEW

Untuk membandingkan performa prediksi kelayakan kredit, saya menggunakan tiga algoritma klasifikasi:

LOGISTIC REGRESSION

Model linier yang sangat interpretatif dan cepat dilatih.

Alasan pemilihan:

- Cocok untuk baseline model klasifikasi.
- Memberikan probabilitas prediksi.
- Sering digunakan untuk interpretasi hubungan antar fitur dan target.

RANDOM FOREST

Model ensemble berbasis decision tree.

Alasan pemilihan:

- Menangani data yang kompleks dan non-linier.
- Tahan terhadap overfitting.
- Dapat mengukur pentingnya fitur (feature importance).

HGBOOST

Model boosting yang sangat kuat dan populer.

Alasan pemilihan:

- Kinerja tinggi dan efisien.
- Baik untuk menangani data tidak seimbang (menggunakan scale_pos_weight).
- Mendukung tuning parameter secara fleksibel.

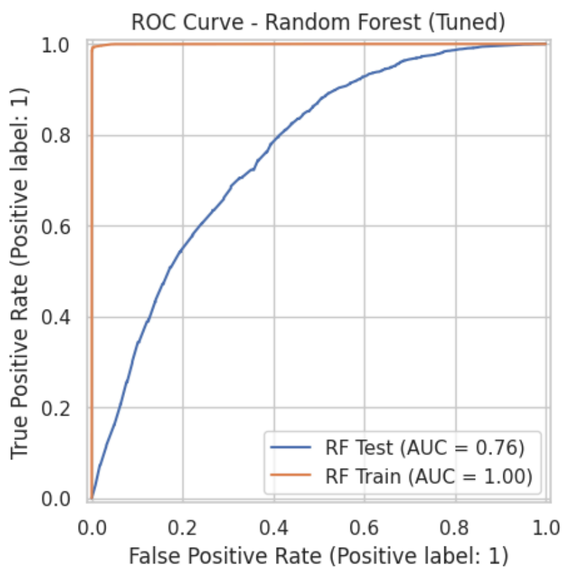
MODEL EVALUATION

Model	Accuracy	Approved (1)			Rejected (0)		
		Precision	Recall	F1 score	Precision	Recall	F1 score
Logistic Regression	0.60	0.91	0.61	0.73	0.16	0.55	0.24
LogReg (Tuned)	0.74	0.89	0.81	0.85	0.15	0.24	0.18
Random Forest	0.89	0.91	0.98	0.94	0.62	0.24	0.35
RF (Tuned)	0.88	0.92	0.95	0.94	0.51	0.36	0.42
XGBoost	0.72	0.93	0.74	0.83	0.22	0.55	0.32
XGB (Tuned)	0.86	0.89	0.95	0.92	0.29	0.15	0.20

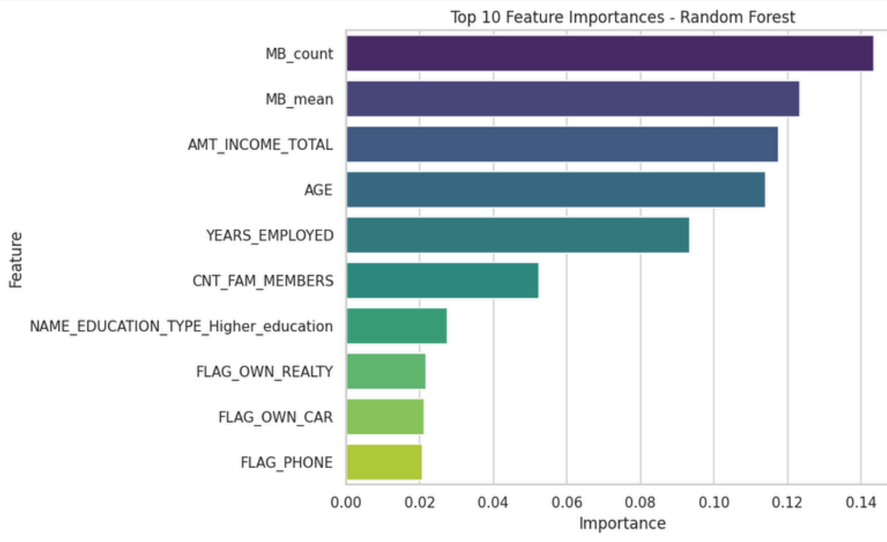
INSIGHT DARI EVALUASI MODEL

Berdasarkan hasil evaluasi, model Random Forest (Tuned) menunjukkan performa paling optimal dan seimbang dalam memprediksi persetujuan kredit. Model ini mampu mengidentifikasi nasabah yang disetujui (approved) dengan sangat baik, terbukti dari nilai recall 0.95 dan F1-score 0.94. Lebih lanjut, model ini juga lebih baik dibanding model lainnya dalam mengenali nasabah yang ditolak (rejected), dengan F1-score tertinggi sebesar 0.42.

FINAL MODEL ANALYSIS



Berdasarkan grafik ROC Curve, model Random Forest (Tuned) memiliki AUC sebesar 0.76 pada data tes. Ini menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam membedakan antara nasabah yang disetujui dan ditolak. Meskipun AUC pada data train mencapai 1.00, perbedaan ini mengindikasikan kemungkinan adanya overfitting. Namun, hal ini masih dalam batas wajar karena performa pada data tes tetap tinggi, menunjukkan bahwa model masih dapat digeneralisasi dengan baik.



Berdasarkan 10 fitur teratas, model mengandalkan fitur MB_count dan MB_mean yang merupakan agregasi dari aktivitas pembayaran sebelumnya, diikuti oleh total pendapatan, usia, dan lama bekerja. Ini menunjukkan bahwa faktor stabilitas keuangan dan pengalaman kerja sangat berpengaruh dalam prediksi persetujuan kredit.

CONCLUSION & RECOMMENDATION

Conclusion:

- Model Random Forest (Tuned) menunjukkan performa terbaik dan paling seimbang dalam memprediksi persetujuan kredit. Hal ini dibuktikan dengan nilai Recall (0.95) dan F1-score (0.94) tertinggi untuk kelas approved (1), serta F1-score tertinggi (0.42) untuk kelas rejected (0) dibandingkan model lainnya.
- Berdasarkan grafik ROC Curve, model memiliki kemampuan generalisasi yang baik dengan nilai AUC sebesar 0.76 pada data tes, meskipun AUC pada data latih mencapai 1.00 yang mengindikasikan sedikit kemungkinan overfitting namun masih dalam batas wajar.

Recommendation:

- Model Random Forest (Tuned) dapat digunakan sebagai sistem screening awal untuk membantu pengambilan keputusan kredit secara cepat.
- Fitur penting seperti intensitas pembayaran, riwayat pembayaran, penghasilan total, usia pemohon, dan lama bekerja harus menjadi perhatian utama dalam evaluasi nasabah.
- Perusahaan dapat mengutamakan pemohon dengan riwayat pembayaran stabil, usia produktif, dan penghasilan tetap untuk meningkatkan rasio persetujuan kredit yang bertanggung jawab.
- Langkah Pengembangan Lanjutan:
  - Lakukan pengujian model pada data real-time untuk mengamati performa aktual.
  - Integrasikan fitur tambahan jika ada.
  - Lakukan pemantauan berkala untuk memastikan performa model tetap stabil dari waktu ke waktu.

THANK YOU