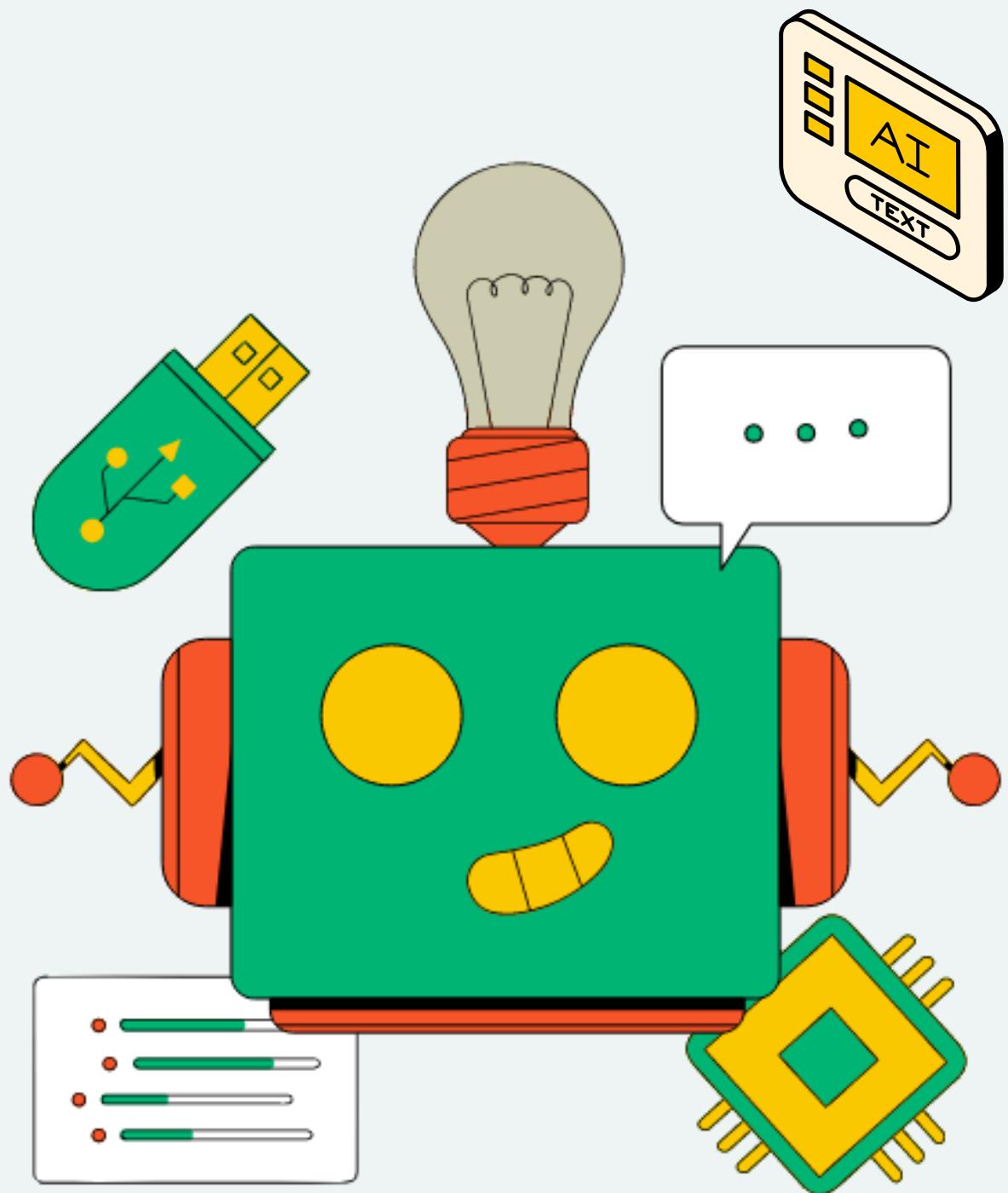


DIBIMBING.ID

WE LEARN FOR THE FUTURE



APPROVAL CREDIT PREDICTION USING CLASSIFICATION MODEL

BATCH 32B | BOOTCAMP DATA
SCIENCE AND DATA ANALYST

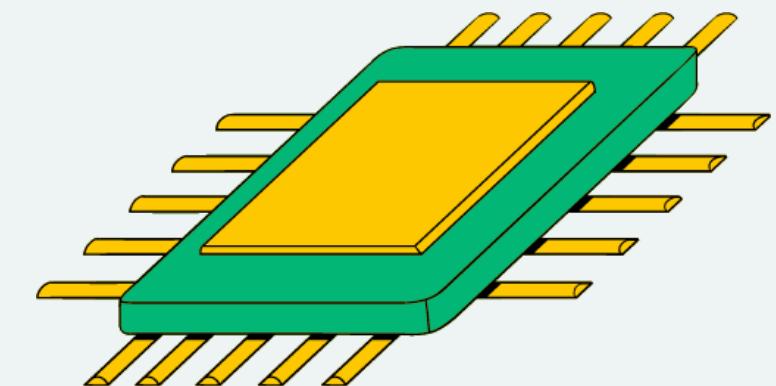


TABLE OF CONTENTS

- A. Introduction
- B. Previous projects
- C. Main Project
 - Project Background
 - Business Objective
 - Data Understanding
 - Exploratory Data Analysis
 - Korelasi Fitur dengan heatmap dan VIF
 - Preprocessing (Encoding, Scaling, SMOTE)
 - Modeling dan Evaluasi Model
- D. Conclusion & Recommendation





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.

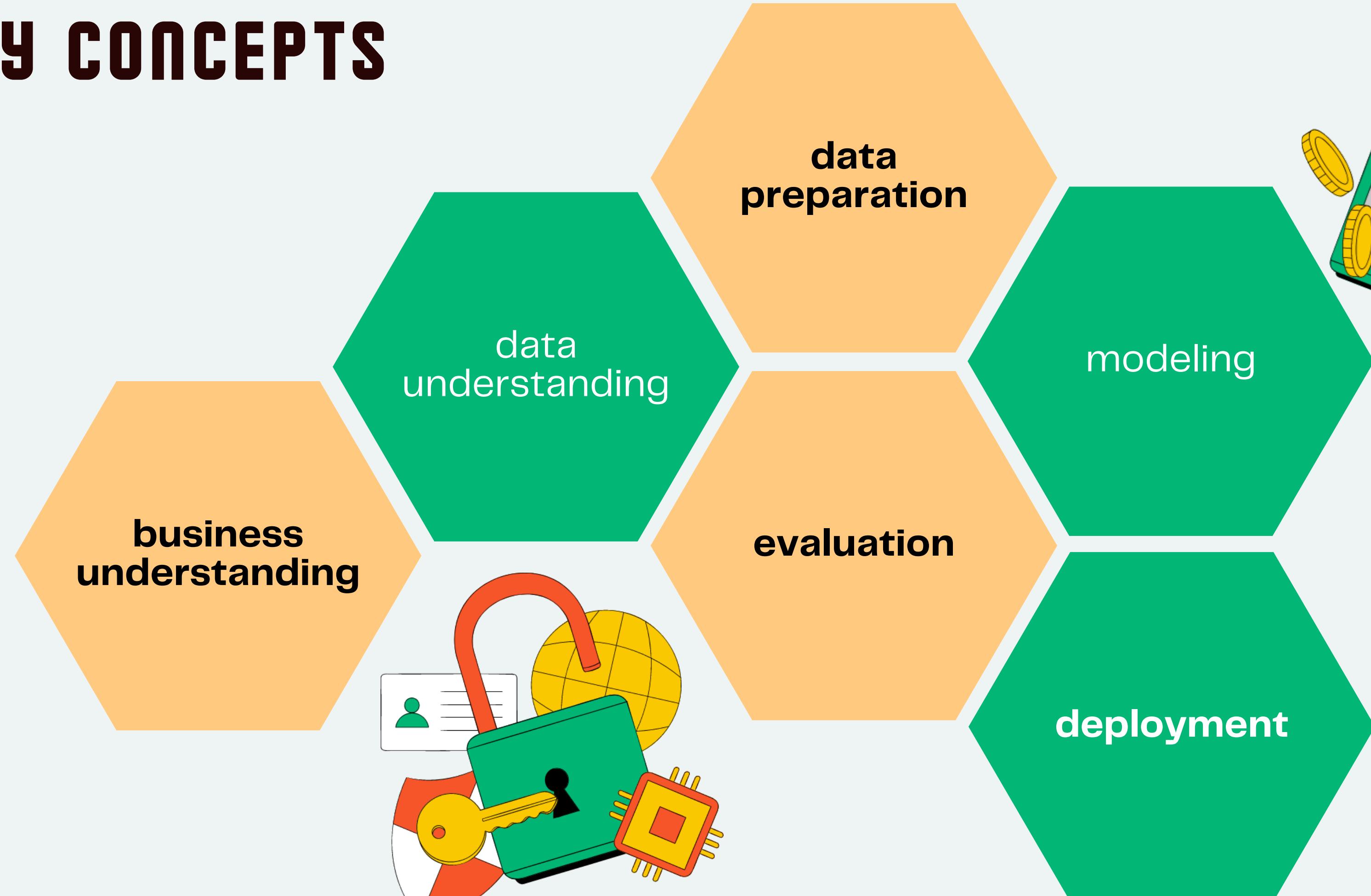
https://github.com/Dadipp/Customer_Satisfaction_and_Sentiment_Analysis



MAIN PROJECT



KEY CONCEPTS



PROJECT BACKGROUND



Menganalisis kelayakan kredit dari pemohon kartu kredit dengan mengevaluasi profil demografis dan finansial mereka, guna mengidentifikasi individu berisiko tinggi dan mempercepat proses persetujuan. Hal ini memungkinkan bank 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.

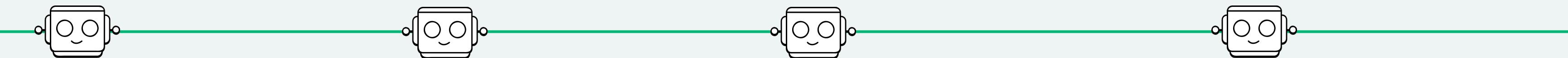
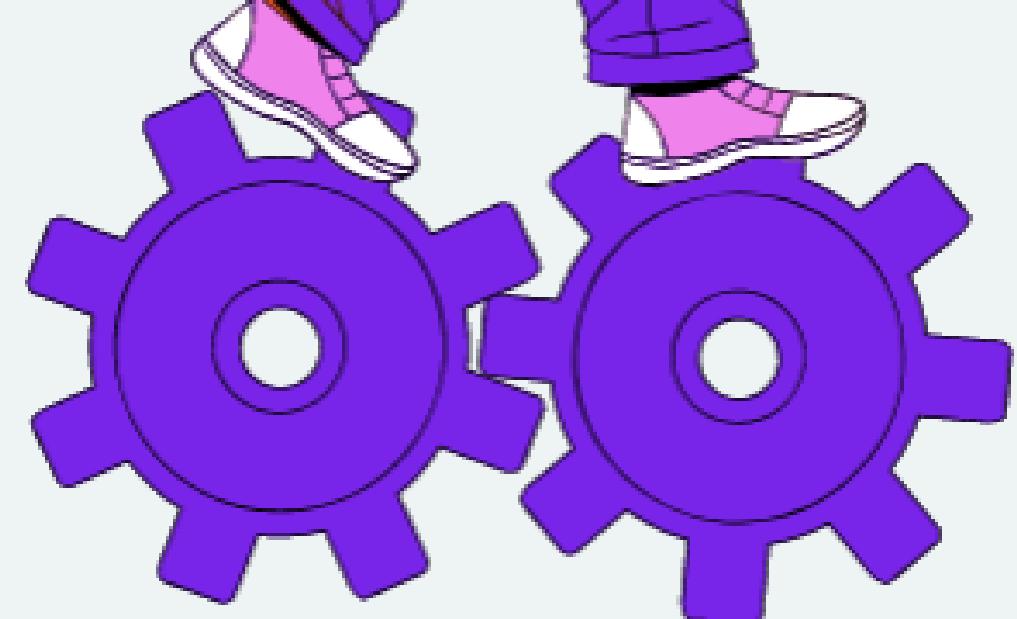


DATA UNDERSTANDING

- **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 waktu, status pembayaran (0–5), dan ID peminjam.
- Dataset akan dihubungkan melalui kolom **ID**, yang merepresentasikan identitas unik pemohon kredit.
- Untuk membangun model prediksi, 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, namun kolom ini tetap mengandung informasi penting terkait profil pekerjaan pemohon. Alih-alih menghapus seluruh baris yang mengandung nilai kosong, saya memilih pendekatan imputasi berbasis modus. Sementara itu, kolom FLAG_MOBIL dihapus karena seluruh nilainya konstan, berarti tidak memberikan informasi atau variasi apapun bagi model, dan hanya menambah dimensi data secara tidak perlu.

REMOVING DUPLICATE

Tidak ditemukan duplicated pada dataset

OUTLIER CHECK

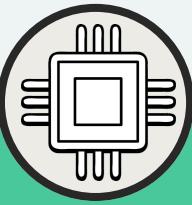
Kolom DAYS_EMPLOYED mencatat durasi masa kerja dalam hari. Observasi menunjukkan adanya nilai ekstrem 365243, yang diinterpretasikan sebagai indikator pemohon yang telah pensiun. Nilai ini dipertahankan karena dianggap sebagai informasi yang relevan dan bukan outlier yang bersifat anomali.

TRANSFORMASI DATA

Beberapa transformasi data dilakukan untuk meningkatkan interpretabilitas dan kualitas data. Kolom DAYS_BIRTH diubah menjadi AGE dalam satuan tahun (nilai negatif dikonversi menjadi usia positif dalam bentuk integer) 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



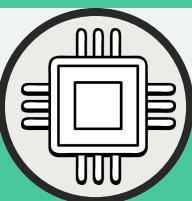
LABELING

Proses labeling untuk status persetujuan (approved) didasarkan pada riwayat kredit pemohon. Kriteria 'disetujui' ditetapkan bagi pemohon yang tidak menunjukkan keterlambatan pembayaran sepanjang periode observasi. Selanjutnya, label yang dihasilkan diintegrasikan ke dalam data aplikasi melalui kolom ID, memastikan setiap entri data memiliki variabel target yang siap untuk diprediksi dalam fase pemodelan.

```
[18] credit_df['STATUS'] = credit_df['STATUS'].replace(['X', 'C'], '0').astype(int)
credit_df['approved'] = credit_df['STATUS'].apply(lambda x: 1 if x >= 1 else 0)
label_df = credit_df.groupby('ID')['approved'].max().reset_index()
label_df['approved'] = label_df['approved'].apply(lambda x: 0 if x == 1 else 1)

[19] # Merge app_df and label_df on 'ID'
df = app_df.merge(label_df, on='ID', how='inner')

# Mapping 1 -> Yes, 0 -> No on the merged dataframe
df['approved'] = df['approved'].map({1: 'Yes', 0: 'No'})
```



FEATURE ENGINEERING

Untuk mempersiapkan data ke tahap analisis lebih lanjut setelah proses labeling, dilakukan serangkaian transformasi fitur. Kolom kategorikal seperti CODE_GENDER, FLAG_OWN_CAR, dan FLAG_OWN_REALTY dikonversi menjadi representasi numerik menggunakan binary mapping. Selanjutnya, untuk memfasilitasi segmentasi dan visualisasi pada tahap EDA, dua fitur baru dibuat: income_bin (berdasarkan total pendapatan) dan AGE (usia pemohon yang dihitung dari DAYS_BIRTH) lalu age_group (pengelompokan berdasarkan usia pemohon), keduanya dikelompokkan ke dalam bins yang relevan.

```
[29] df['CNT_FAM_MEMBERS'] = df['CNT_FAM_MEMBERS'].round().astype(int)
df['CODE_GENDER'] = df['CODE_GENDER'].map({'M': 0, 'F': 1})
df['FLAG_OWN_CAR'] = df['FLAG_OWN_CAR'].map({'N': 0, 'Y': 1})
df['FLAG_OWN_REALTY'] = df['FLAG_OWN_REALTY'].map({'N': 0, 'Y': 1})

[30] bins = [0, 100000, 150000, 200000, 300000, float('inf')]
labels = ['<100K', '100-150K', '150-200K', '200-300K', '300K+']
df['income_bin'] = pd.cut(df['AMT_INCOME_TOTAL'], bins=bins, labels=labels)

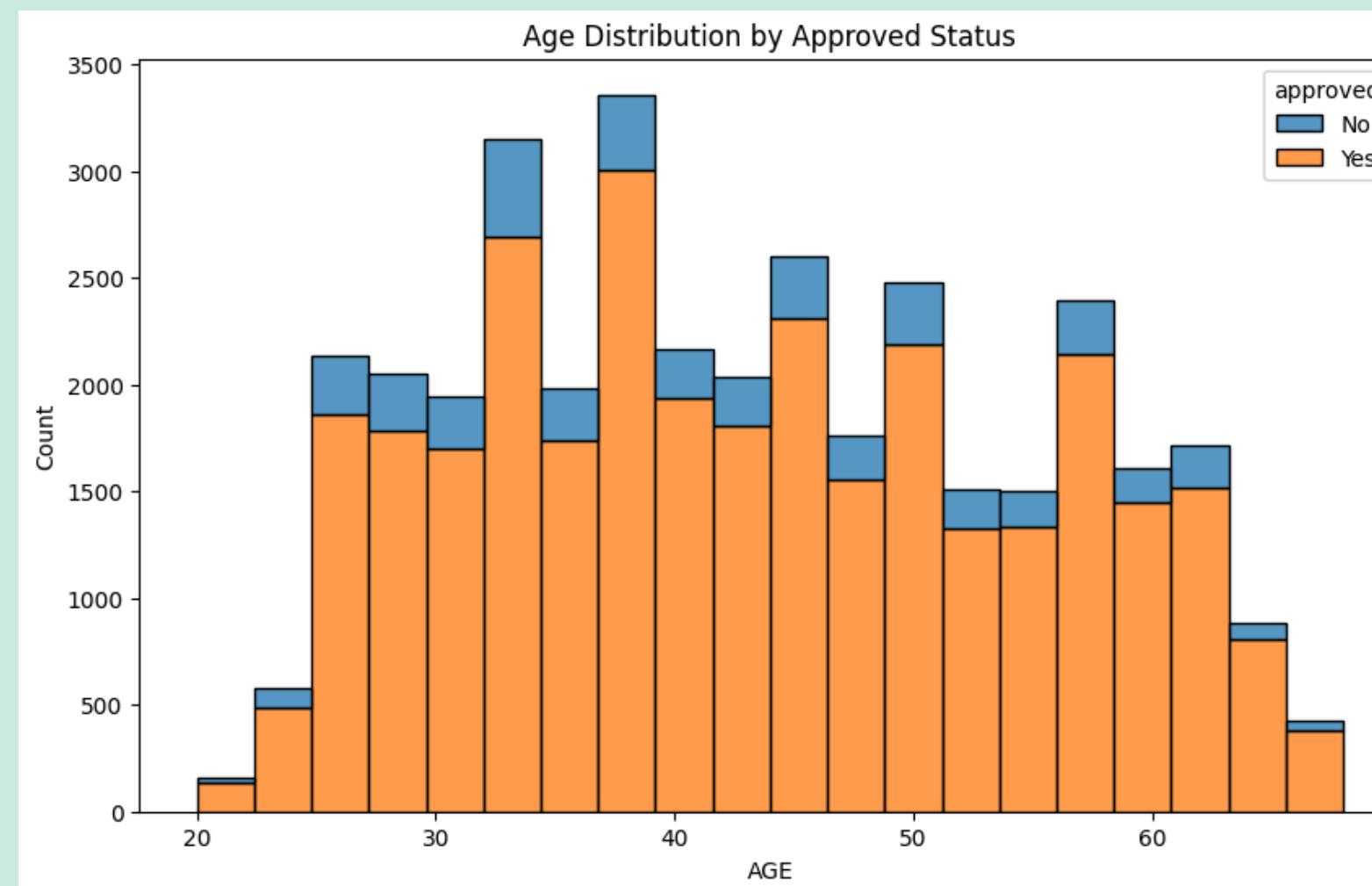
[31] df['age_group'] = pd.cut(df['AGE'],
                           bins=[20, 30, 40, 50, 60, 70],
                           labels=['20-30', '30-40', '40-50', '50-60', '60-70'],
                           right=False)

[32] df['AGE'] = (-df['DAYS_BIRTH'] / 365).astype(int)
df['YEARS_EMPLOYED'] = (-df['DAYS_EMPLOYED'] / 365).astype(int)

[33] df.drop(columns=['DAYS_BIRTH', 'DAYS_EMPLOYED'], inplace=True)
```



EDA



```
approved_by_age = df.groupby('age_group')['approved'].value_counts(normalize=True).unstack().fillna(0)
approved_by_age['Approved_Rate'] = approved_by_age['Yes']
print("\nApproved Rate by Age Group:")
print(approved_by_age[['Approved_Rate']])
```

```
Approved Rate by Age Group:
approved    Approved_Rate
age_group
20-30        0.866477
30-40        0.875527
40-50        0.888638
50-60        0.888204
60-70        0.893445
```

01

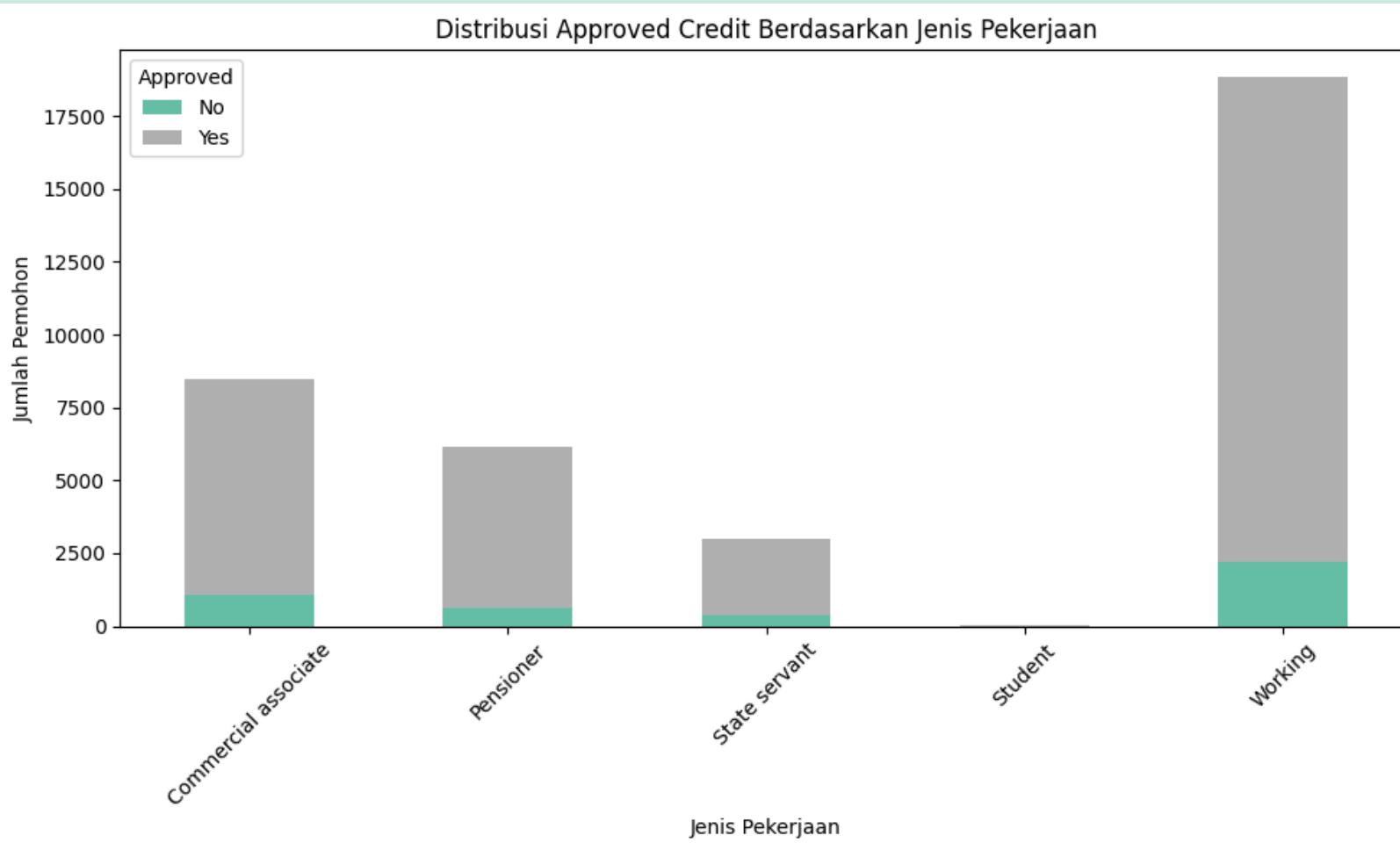
AGE DISTRIBUTION BY APPROVED STATUS



Insight dari Distribusi Usia dan Kelayakan Kredit :

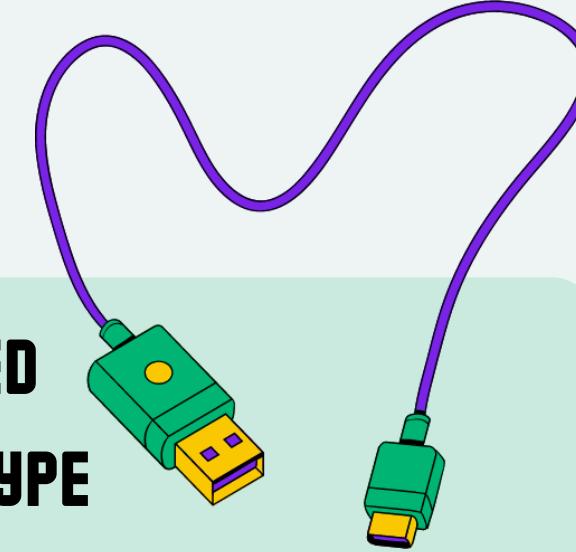
- Mayoritas pemohon kredit berada dalam rentang usia 30 hingga 60 tahun, dengan konsentrasi tertinggi di sekitar usia 35-45 tahun, seperti yang terlihat pada grafik distribusi usia.
- Tingkat persetujuan kredit (Approved_Rate) menunjukkan tren peningkatan seiring bertambahnya usia. Tingkat persetujuan berkisar dari 86.65% pada kelompok usia 20-30 tahun dan terus meningkat hingga mencapai 89.34% pada kelompok usia 60an tahun.
- Pemohon yang berusia di atas 40 tahun secara konsisten menunjukkan tingkat kelayakan persetujuan kredit yang lebih tinggi dibandingkan dengan kelompok usia yang lebih muda. Oleh karena itu, strategi pemasaran dan penawaran produk kredit dapat lebih difokuskan pada segmen usia 40-60an tahun untuk mengoptimalkan tingkat persetujuan aplikasi.





02

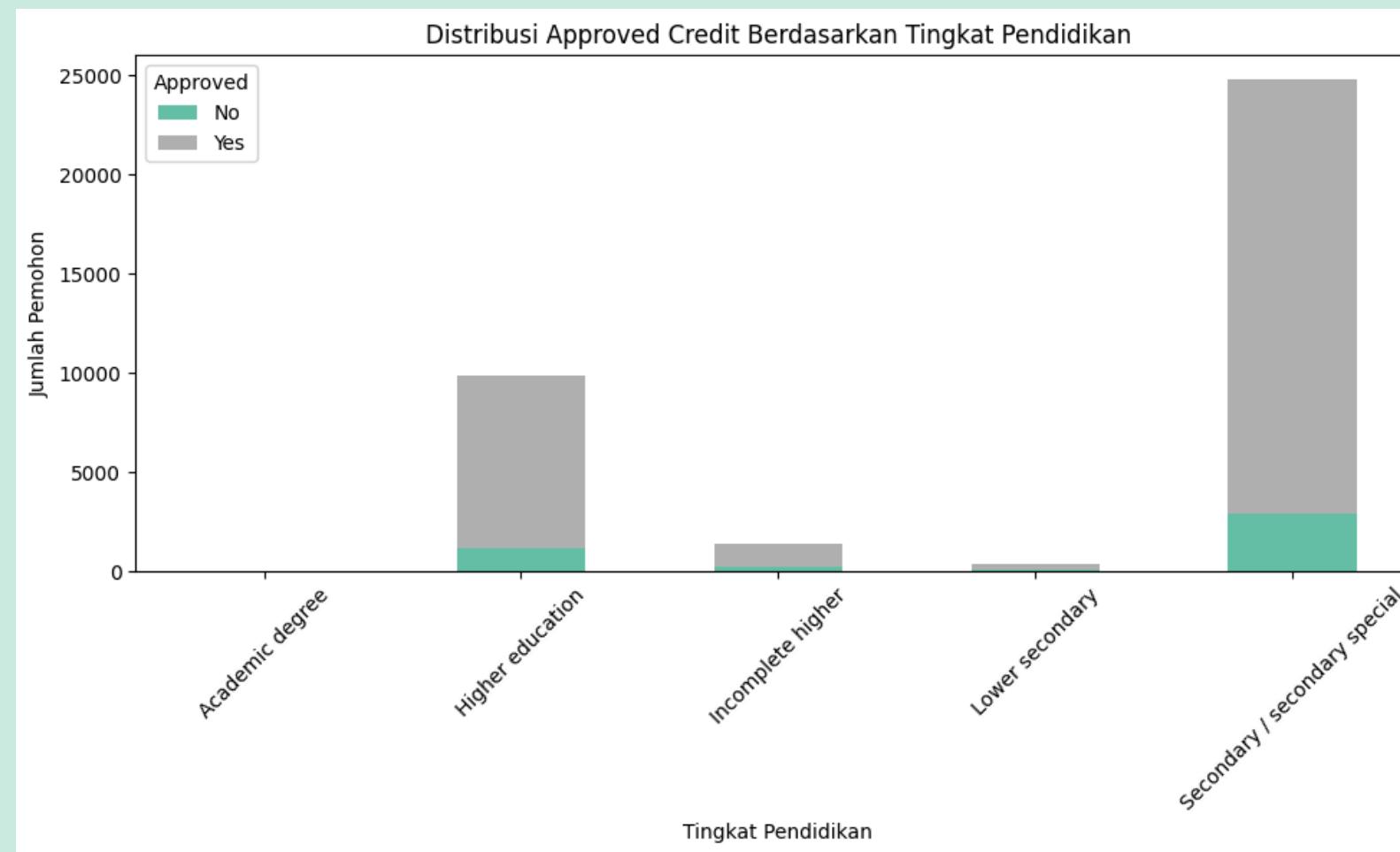
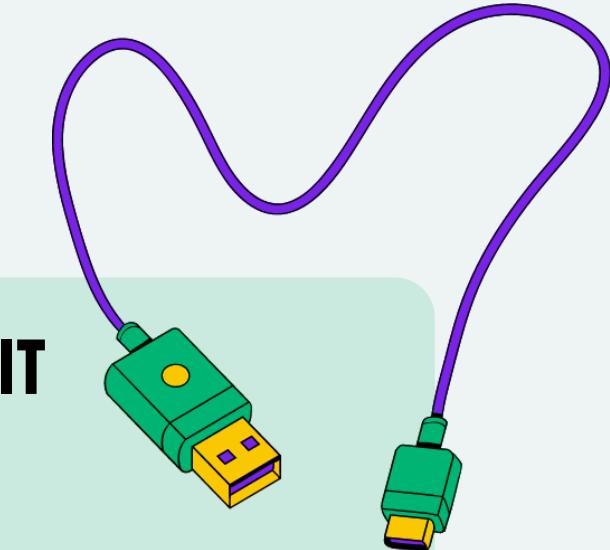
DISTRIBUTION OF APPROVED CREDIT BASED ON INCOME TYPE



Insight dari Distribusi Jenis Pekerjaan dan Kelayakan Kredit:

- Mayoritas pemohon kredit berasal dari kelompok pekerjaan Working, diikuti oleh Commercial associate dan Pensioner.
- Tingkat persetujuan kredit (Approved Rate) secara umum tinggi di hampir semua kategori pekerjaan. Kelompok State servant dan Pensioner menunjukkan stabilitas penghasilan yang cenderung mendukung tingkat persetujuan yang baik.
- Kategori Student memiliki jumlah pemohon yang sangat kecil dan tingkat persetujuan yang rendah. Hal ini kemungkinan besar disebabkan tidak ada nya penghasilan tetap atau riwayat kredit yang minim.
- Meskipun kelompok Working mendominasi jumlah pemohon, tingkat penolakan di kategori ini juga relatif tinggi. Ini menunjukkan perlunya analisis risiko yang lebih mendalam dan spesifik terhadap pemohon dari kategori Working untuk mengidentifikasi faktor-faktor yang berkontribusi pada penolakan.



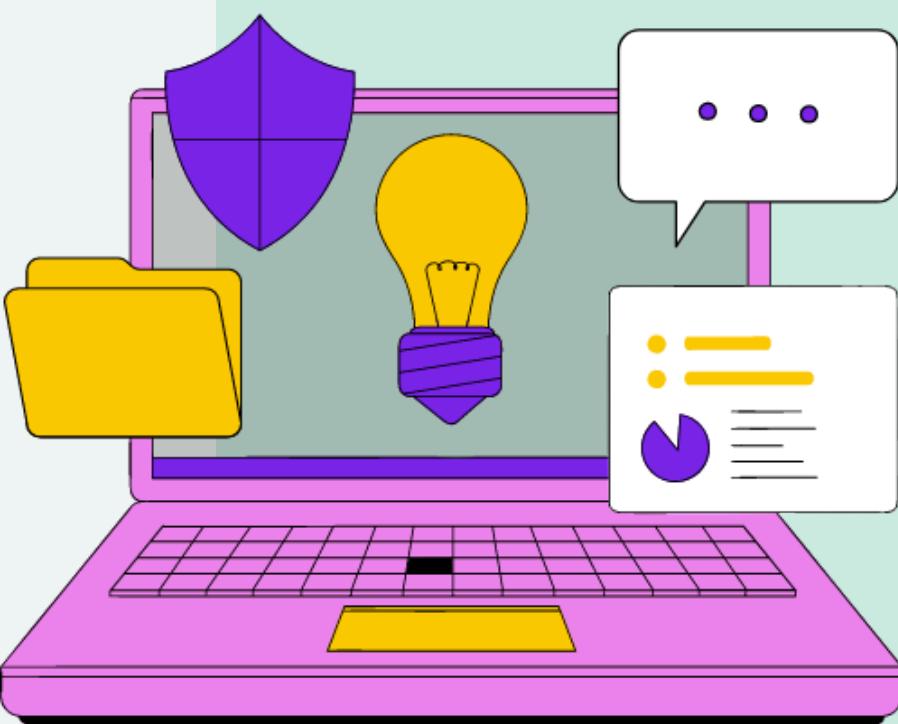


03

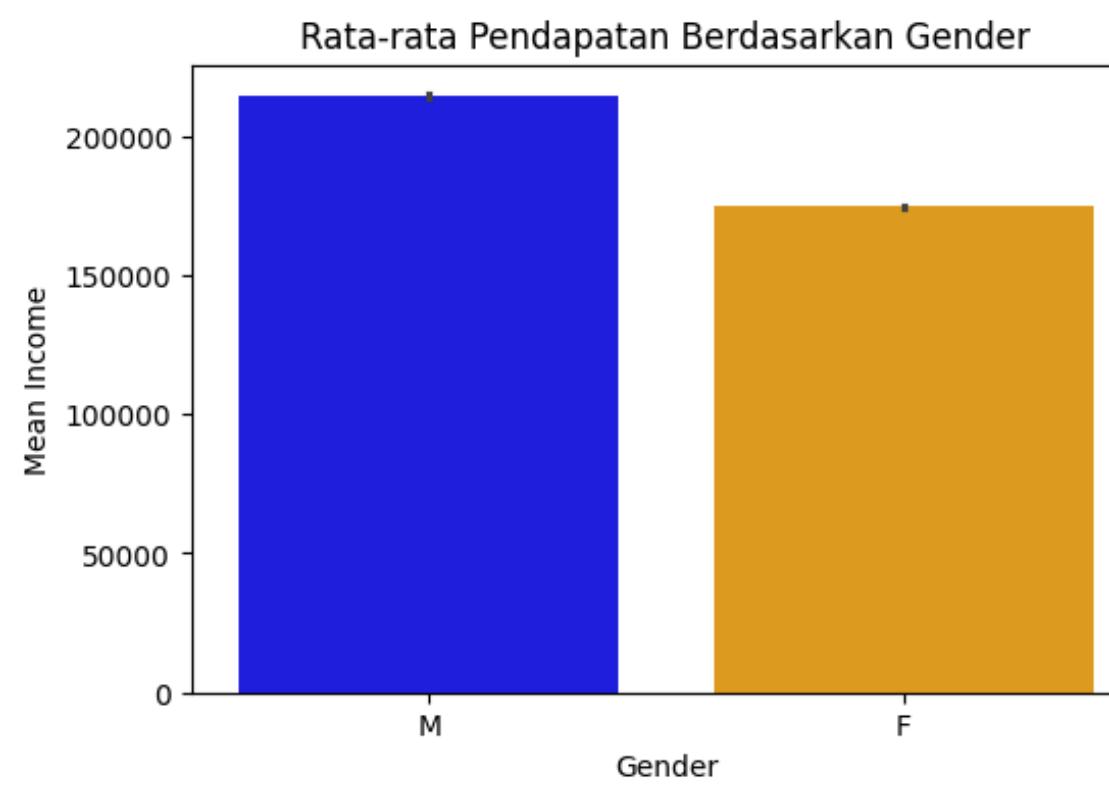
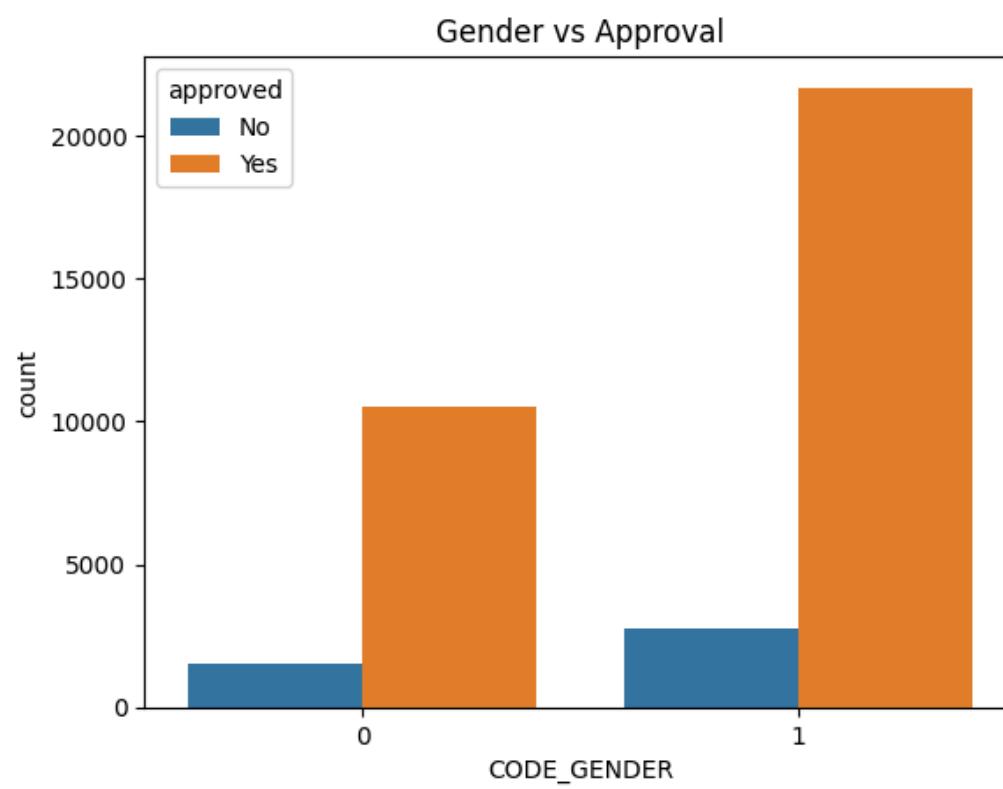
DISTRIBUTION OF APPROVED CREDIT BASED ON EDUCATION LEVEL

Insight dari Distribusi Jenis Pekerjaan dan Kelayakan Kredit:

- Mayoritas pemohon kredit berasal dari kelompok pendidikan Secondary / secondary special dan Higher education. Kelompok dengan pendidikan menengah menunjukkan jumlah pengajuan tertinggi, disusul oleh pendidikan tinggi.
- Tingkat persetujuan kredit cukup tinggi di hampir semua jenjang pendidikan, terutama pada kelompok Higher education dan Secondary special, yang mencerminkan bahwa tingkat pendidikan turut memengaruhi persepsi kelayakan kredit.
- Meskipun jumlah pemohon dari kategori Incomplete higher, Lower secondary, dan Academic degree relatif kecil, tingkat persetujuannya tetap stabil.
- Dengan demikian, strategi pemasaran kredit dapat difokuskan pada segmen dengan pendidikan menengah dan tinggi karena mencakup mayoritas pemohon serta memiliki potensi persetujuan kredit yang tinggi.



EDA



04

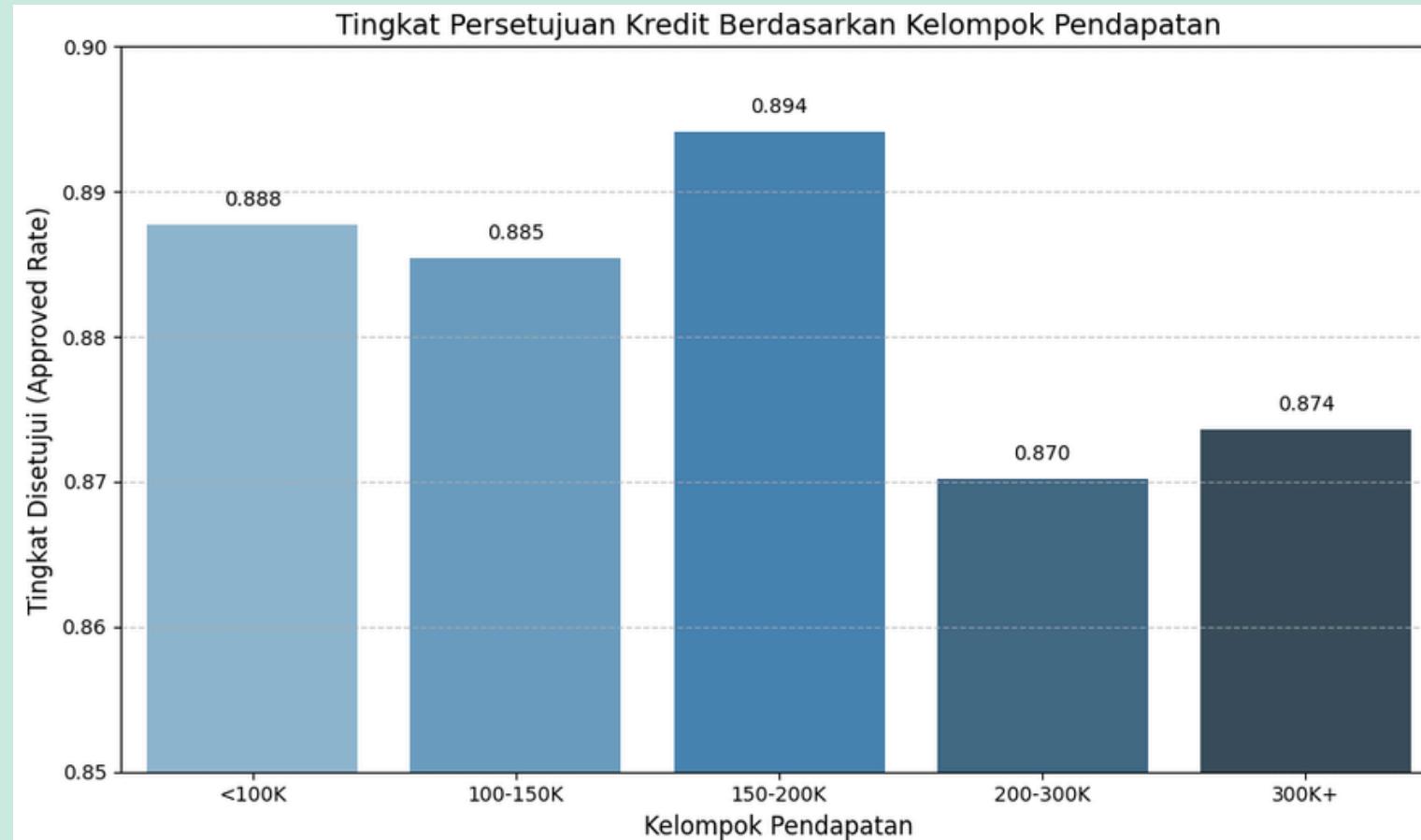
GENDER VS APPROVAL

Insight dari Distribusi Gender dan Kelayakan Kredit :

- Perempuan memiliki tingkat persetujuan aplikasi yang lebih tinggi (88.71%) dibandingkan laki-laki (87.25%).
- Grafik menunjukkan bahwa jumlah pemohon perempuan juga jauh lebih banyak dibandingkan laki-laki.

- Meskipun secara jumlah absolut perempuan mendominasi persetujuan, secara proporsional pun perempuan tetap unggul dalam tingkat persetujuan.
- Hal ini mengindikasikan bahwa gender berpengaruh terhadap persetujuan aplikasi. Perempuan cenderung memiliki profil risiko yang lebih rendah atau lebih sesuai dengan kriteria penilaian kelayakan kredit dalam dataset ini.
- Menariknya, meskipun pendapatan rata-rata laki-laki lebih tinggi, hal tersebut tidak berbanding lurus dengan tingkat persetujuan. Ini menunjukkan bahwa faktor lain seperti stabilitas pekerjaan, riwayat pembayaran, atau rasio beban utang lebih memengaruhi keputusan persetujuan kredit.





05

APPROVED RATE BY INCOME BIN

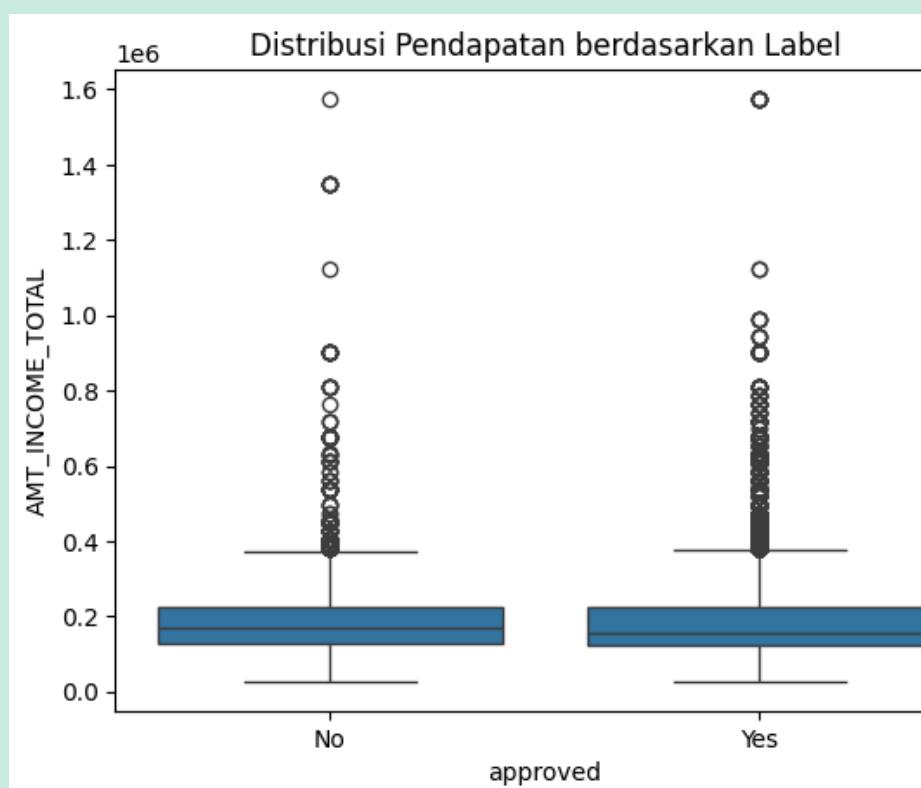
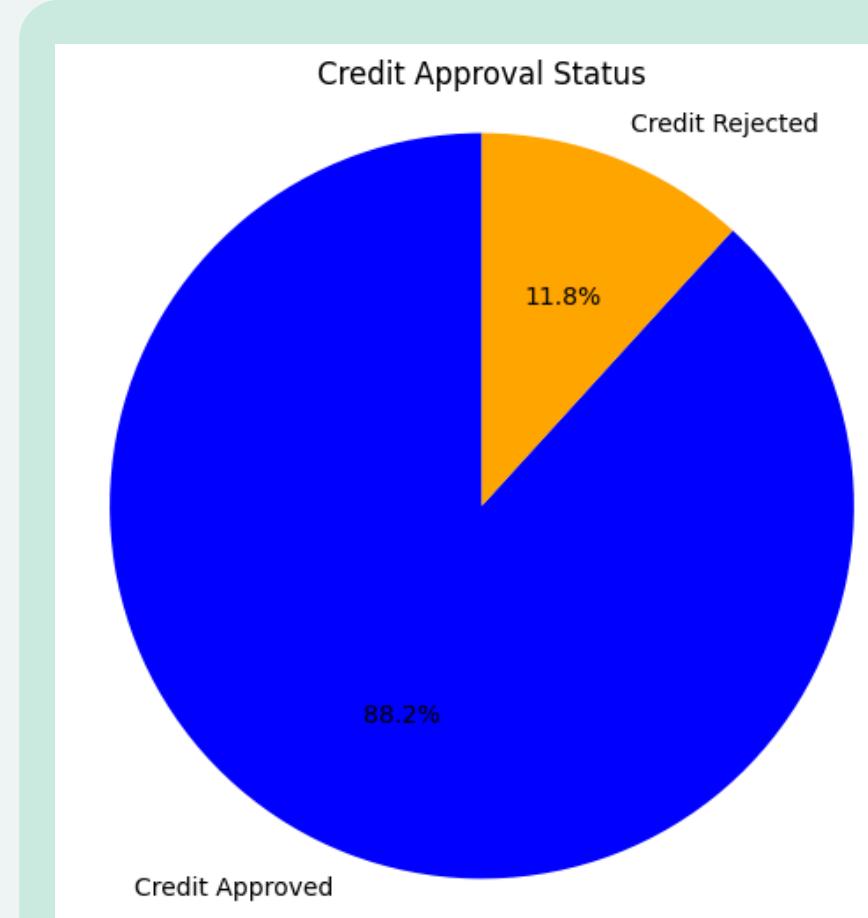
Insight dari Distribusi Pendapatan dan Kelayakan Kredit:

- Tingkat persetujuan kredit cenderung stabil di atas 87% di semua kelompok pendapatan, menunjukkan bahwa faktor pendapatan memiliki pengaruh yang cukup merata terhadap kelayakan kredit.
- Kelompok dengan pendapatan 150-200K memiliki tingkat persetujuan tertinggi yaitu 89.41%.
- Sementara itu, kelompok dengan pendapatan 200-300K dan 300K+ justru mengalami penurunan tingkat persetujuan menjadi sekitar 87%, padahal kelompok berpendapatan tinggi.
- Hal ini menunjukkan bahwa tingkat pendapatan yang lebih tinggi tidak selalu berkorelasi langsung dengan peluang disetujuinya kredit.

Strategi pemasaran dan kebijakan risiko dapat difokuskan pada kelompok pendapatan 100K–200K, karena mereka menunjukkan keseimbangan antara volume pemohon dan tingkat persetujuan yang tinggi.



EDA



06

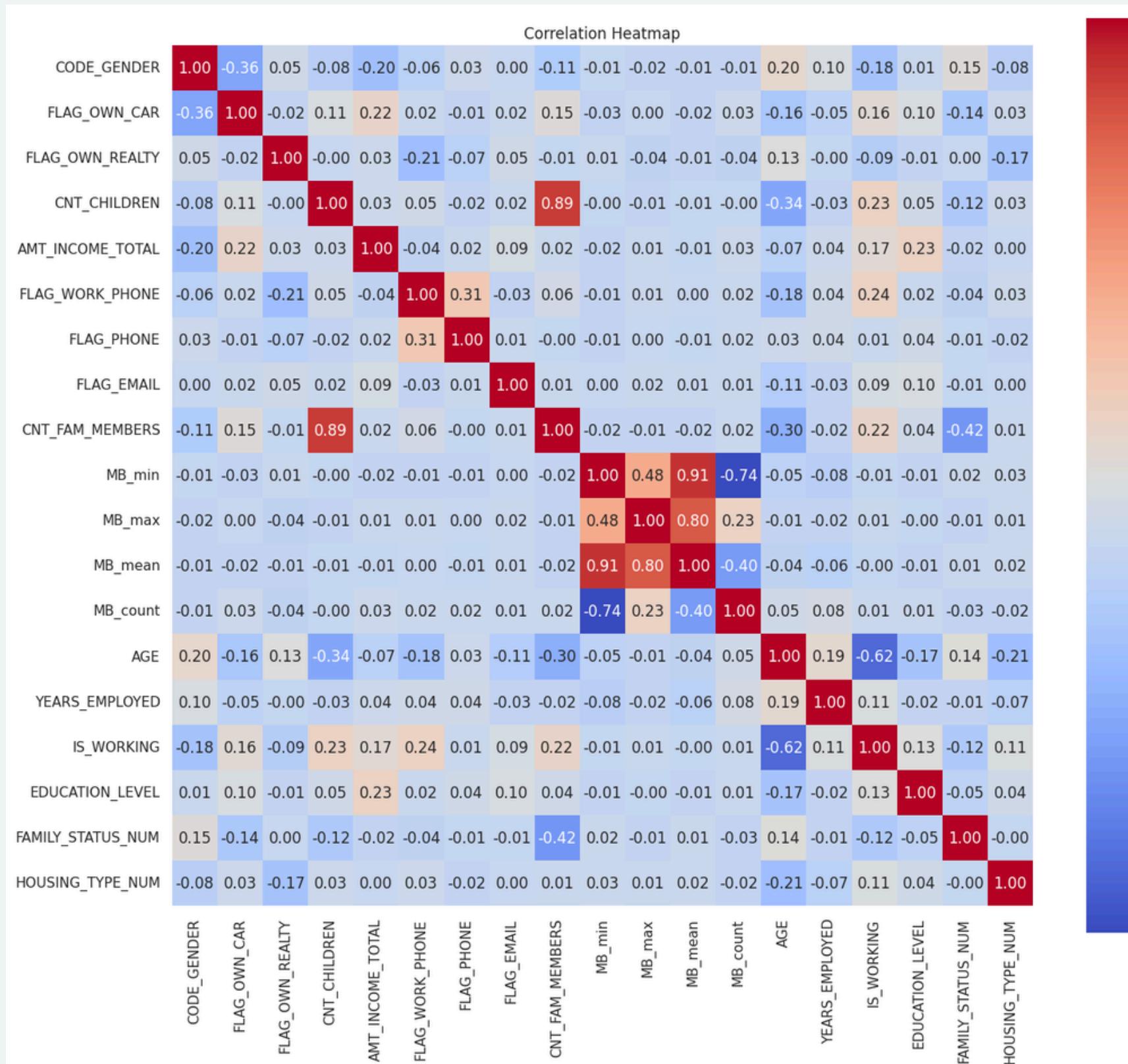
CREDIT APPROVAL OVERVIEW & INCOME DISTRIBUTION

Insight dari Overview Persetujuan Kredit & Distribusi Pendapatan:

- Sebagian besar aplikasi kredit disetujui, dengan tingkat persetujuan mencapai 88.2%, dan hanya 11.8% yang ditolak. Ini menunjukkan mayoritas pemohon memenuhi kriteria kelayakan kredit yang berlaku.
- Distribusi pendapatan antara pemohon yang disetujui dan ditolak tampak serupa, baik dari sisi median maupun sebaran total pendapatannya.
- Hal ini mengindikasikan bahwa pendapatan bukan faktor utama yang membedakan keputusan persetujuan kredit.
- Kemungkinan besar, faktor lain seperti riwayat pembayaran, jenis pekerjaan, atau stabilitas penghasilan lebih berperan dalam menentukan kelayakan kredit.



FEATURE CORRELATION & MULTIKOLINEARITAS



Insight dari Analisis Korelasi

- Fitur Monthly Balance: Terdapat korelasi sangat tinggi antar fitur turunan seperti MB_min, MB_max, MB_mean, dan MB_count. Untuk menghindari informasi yang berulang dan potensi multikolinearitas, hanya MB_mean dan MB_count yang dipertahankan karena dinilai paling mewakili intensitas dan durasi aktivitas kredit.
- Fitur Keluarga: Kolom CNT_CHILDREN memiliki korelasi sangat tinggi dengan CNT_FAM_MEMBERS. Karena CNT_FAM_MEMBERS lebih informatif karena mencakup total anggota keluarga., maka CNT_CHILDREN dihapus agar model tidak terdampak fitur yang tumpang tindih.
- Fitur Lainnya: Sebagian besar fitur lainnya menunjukkan korelasi rendah hingga sedang. Ini berarti setiap fitur cenderung membawa informasi yang berbeda, dan tetap relevan untuk dilibatkan dalam pemodelan.
- Pola Umum: Terlihat hubungan positif antara usia, masa kerja, dan tingkat pendidikan. Ini mengindikasikan bahwa pemohon yang lebih tua dan berpendidikan cenderung memiliki pengalaman kerja yang lebih panjang.



VIF ANALYSIS

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant

features = df.select_dtypes(include=['int64', 'float64']).drop(columns=['ID'])

# Update num_cols to only include columns present in df
num_cols = [col for col in num_cols if col in df.columns]

X_vif = add_constant(df[num_cols])
vif_data = pd.DataFrame()
vif_data["Feature"] = X_vif.columns
vif_data["VIF_score"] = [variance_inflation_factor(X_vif.values, i) for i in range(X_vif.shape[1])]
vif_data
```

	Feature	VIF_score
0	const	79.212302
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 Analisis VIF

- Mengidentifikasi adanya multikolinearitas antar fitur numerik dalam dataset. Nilai VIF mengukur seberapa besar variabilitas suatu fitur dapat dijelaskan oleh fitur lainnya.

• Interpretasi Nilai VIF:

- VIF = 1: Tidak ada korelasi antar fitur.
- VIF 1–5: Masih dapat diterima.
- VIF > 5: Potensi multikolinearitas tinggi, perlu ditinjau ulang.
- VIF > 10: Wajib dipertimbangkan untuk dihapus.

• Hasil Utama:

- Semua fitur memiliki VIF < 5, menunjukkan tidak ada multikolinearitas serius dalam data.
- Fitur AGE dan IS_WORKING memiliki VIF tertinggi, namun tetap dalam ambang aman (masih < 2.1).
- Konstanta (const) memiliki nilai tinggi, namun ini wajar dan tidak menjadi masalah karena bukan fitur prediktor.

• Kesimpulan:

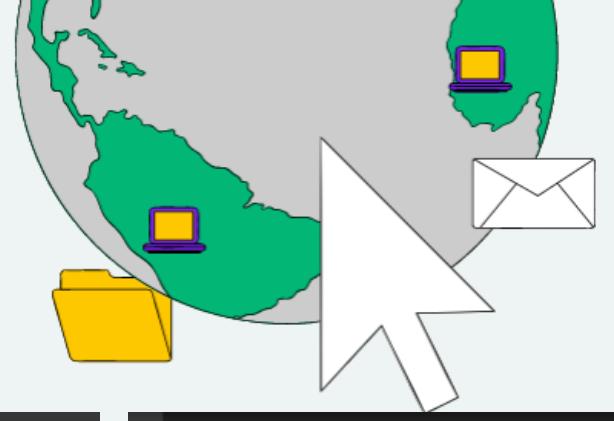
- Semua fitur numerik aman untuk dilanjutkan ke tahap modeling tanpa perlu penghapusan lebih lanjut karena multikolinearitas.



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



```
[72] target_column = 'approved'
y = df[target_column].map({'Yes': 1, 'No': 0})
X = df.drop(columns=[target_column])

[73] # Pisahkan kolom kategorikal dan numerikal
cat_cols = X.select_dtypes(include=['object',
                                      'category']).columns.tolist()
num_cols = X.select_dtypes(include=np.number).columns.tolist()

[76] # One-Hot Encoding
X = pd.get_dummies(X, columns=cat_cols,
                    drop_first=True)

[77] # Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=11,
    stratify=y
)

[78] X_train.shape
[79] X_test.shape

[80] num_cols_final = [col for col in num_cols if col in X_train.columns]

[81] # Scaling numerik
scaler = StandardScaler()
X_train[num_cols_final] = scaler.fit_transform(
    X_train[num_cols_final])
X_test[num_cols_final] = scaler.transform(
    X_test[num_cols_final])
```



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).

XGBOOST

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

Before SMOTE :

==== Model Performance Sebelum SMOTE ====				
Logistic Regression Performance:				
	precision	recall	f1-score	support
0	0.16	0.55	0.24	858
1	0.91	0.61	0.73	6434
accuracy			0.60	7292
macro avg	0.53	0.58	0.49	7292
weighted avg	0.82	0.60	0.67	7292
Random Forest Performance:				
	precision	recall	f1-score	support
0	0.62	0.24	0.35	858
1	0.91	0.98	0.94	6434
accuracy			0.89	7292
macro avg	0.76	0.61	0.65	7292
weighted avg	0.87	0.89	0.87	7292
XGBoost Performance:				
	precision	recall	f1-score	support
0	0.22	0.55	0.32	858
1	0.93	0.74	0.83	6434
accuracy			0.72	7292
macro avg	0.57	0.65	0.57	7292
weighted avg	0.84	0.72	0.77	7292

After SMOTE :

===== Setelah SMOTE =====				
Logistic Regression (SMOTE) Performance:				
	precision	recall	f1-score	support
0	0.15	0.23	0.18	858
1	0.89	0.82	0.85	6434
accuracy			0.75	7292
macro avg	0.52	0.53	0.52	7292
weighted avg	0.80	0.75	0.77	7292
Random Forest (SMOTE) Performance:				
	precision	recall	f1-score	support
0	0.50	0.36	0.42	858
1	0.92	0.95	0.93	6434
accuracy			0.88	7292
macro avg	0.71	0.66	0.68	7292
weighted avg	0.87	0.88	0.87	7292
XGBoost (SMOTE) Performance:				
	precision	recall	f1-score	support
0	0.40	0.14	0.21	858
1	0.89	0.97	0.93	6434
accuracy			0.87	7292
macro avg	0.65	0.56	0.57	7292
weighted avg	0.84	0.87	0.85	7292

INSIGHT & EVALUASI MODEL

Sebelum SMOTE:

- Semua model (Logistic Regression, Random Forest, XGBoost) sangat bias terhadap kelas mayoritas (approved).
- Recall dan f1-score untuk kelas minoritas (rejected) sangat rendah, membuat accuracy tidak representatif.

Setelah SMOTE:

- SMOTE berhasil meningkatkan recall dan f1-score untuk kelas minoritas di sebagian besar model.
- Random Forest (SMOTE) menunjukkan f1-score kelas minoritas tertinggi (0.42), menjadikannya kandidat terbaik.
- XGBoost (SMOTE) menunjukkan penurunan recall dan f1-score untuk kelas minoritas dibandingkan sebelum SMOTE, perlu investigasi lebih lanjut.

Kesimpulan Sementara:

- SMOTE efektif meningkatkan kemampuan model mengenali aplikasi yang ditolak.
- Random Forest (SMOTE) adalah model paling baik berdasarkan keseimbangan performa setelah penanganan imbalance.

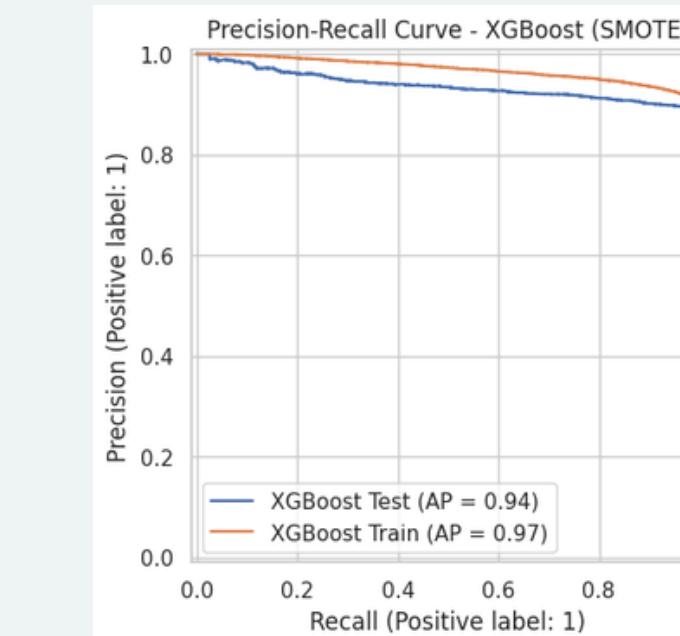
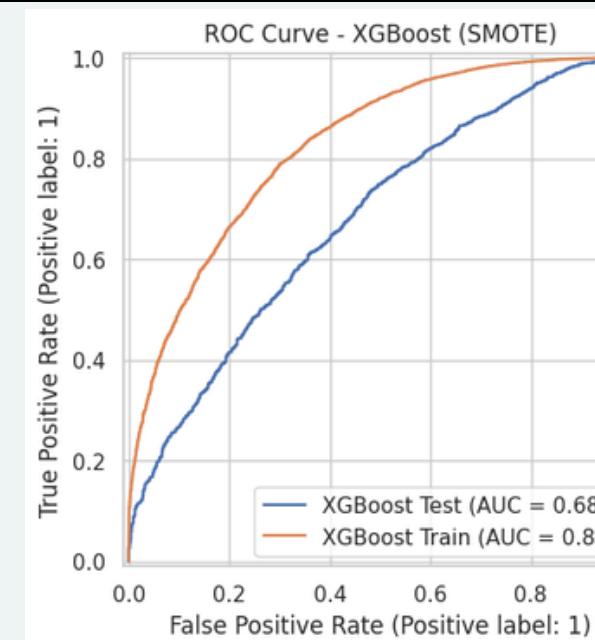


FINAL MODEL ANALYSIS

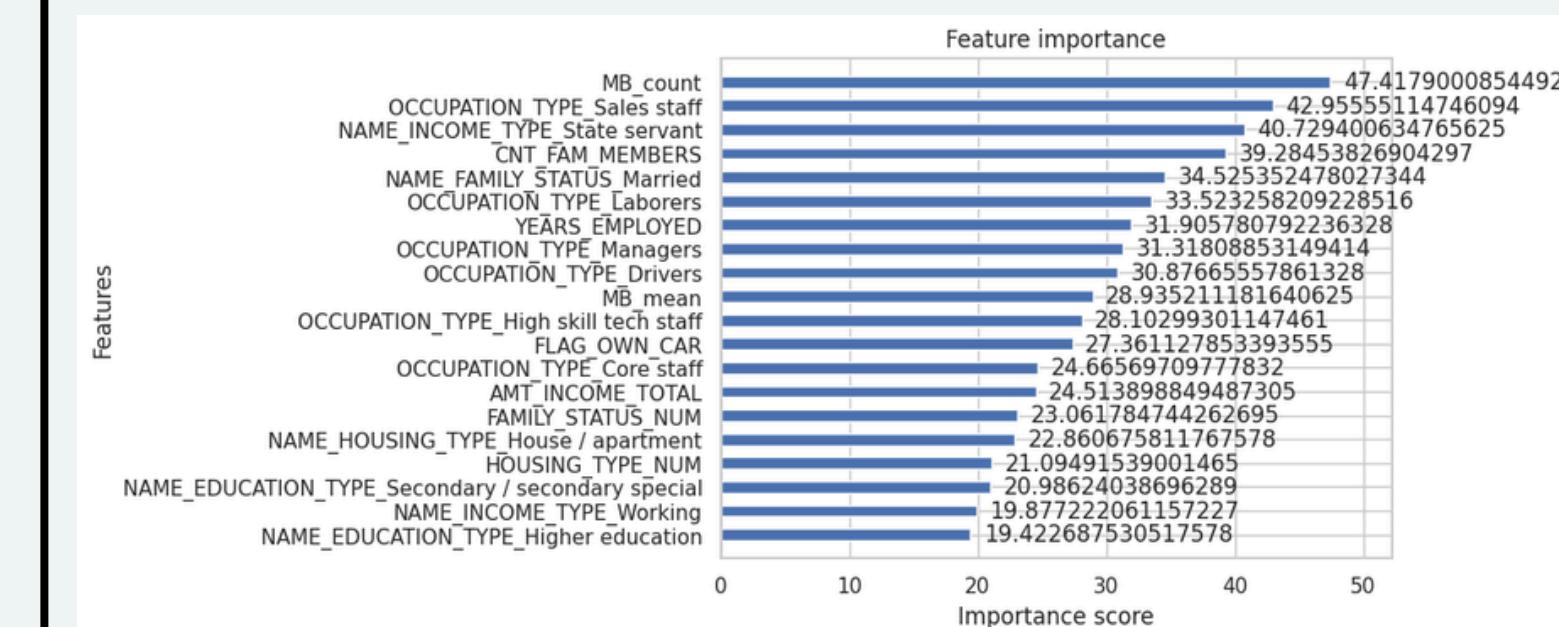
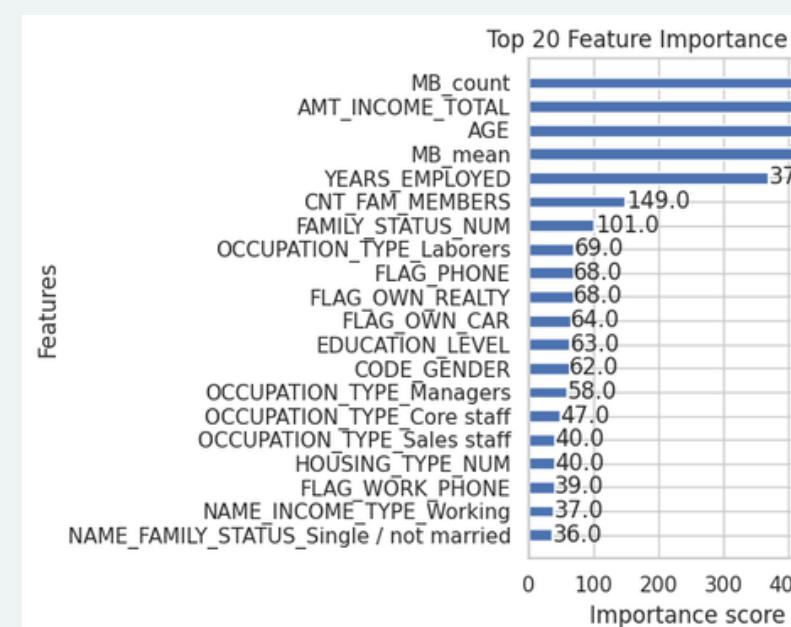
Cross-Validation Evaluation (CV) : Untuk menguji kestabilan model XGBoost (SMOTE), dilakukan validasi silang 5-fold. Hasil rata-rata F1 Score: 0.9238, menunjukkan bahwa model cukup stabil dan konsisten dalam membedakan kelas mayoritas dan minoritas.

```
[94] score = cross_val_score(xgb_sm, X_train_smote, y_train_smote, cv=5, scoring='f1')
print("Average F1 Score (CV):", score.mean())
→ Average F1 Score (CV): 0.9238535469497304
```

ROC & Precision-Recall Curve : Grafik ROC dan Precision-Recall menunjukkan performa model XGBoost sangat baik, dengan kurva mendekati sudut kiri atas (ROC) dan area tinggi (PR), mengindikasikan kemampuan klasifikasi yang kuat pada data tidak seimbang.



Feature Importance : Fitur-fitur seperti **MB_count**, **AMT_INCOME_TOTAL**, **AGE**, dan **YEARS_EMPLOYED** memiliki kontribusi terbesar dalam memengaruhi keputusan model. Hal ini membantu tim bisnis memahami faktor utama yang memengaruhi persetujuan kredit.



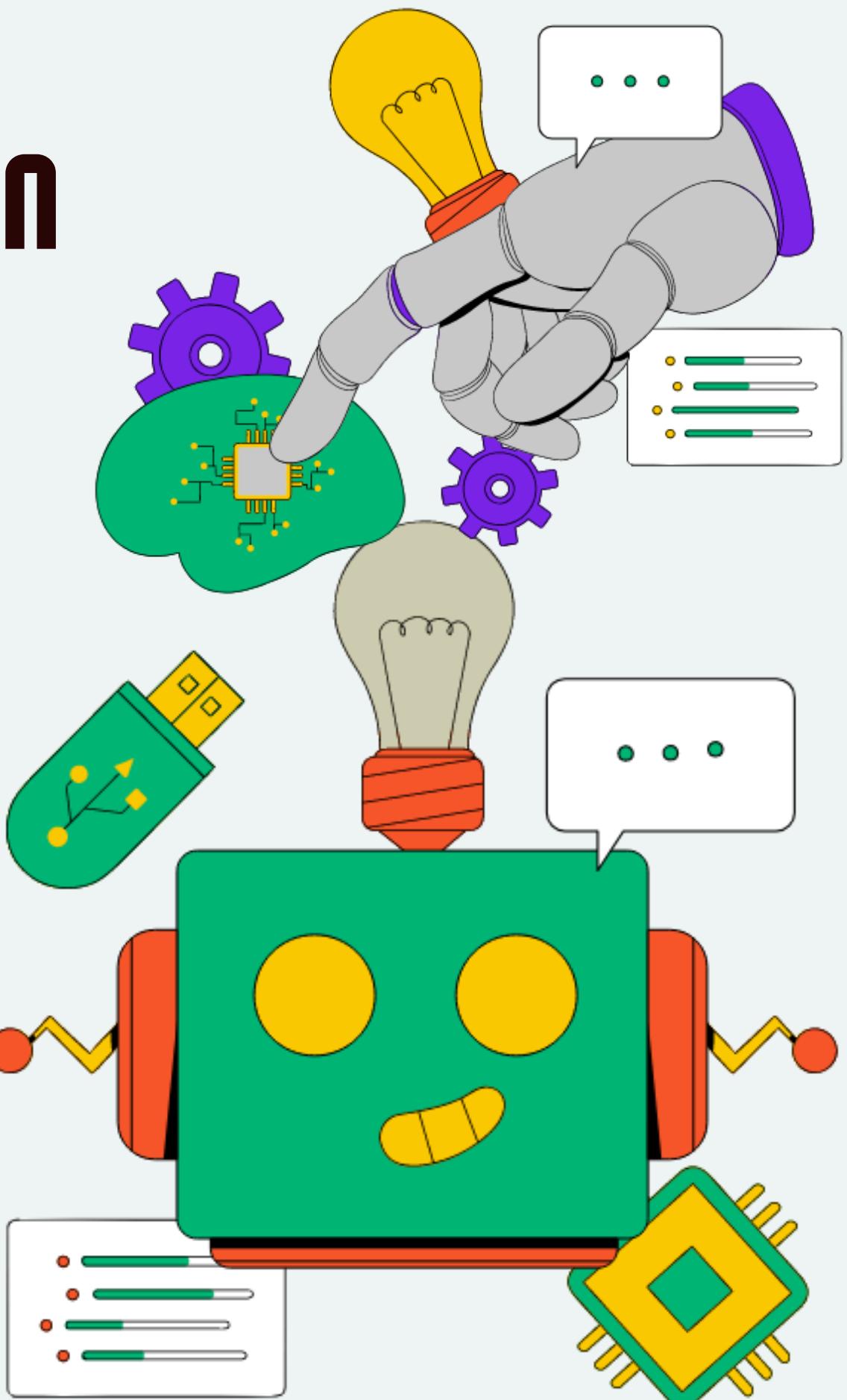
CONCLUSION & RECOMMENDATION

Conclusion:

- Model XGBoost dengan SMOTE memberikan performa terbaik dengan rata-rata F1 Score sebesar 0.92 dari validasi silang.
- Model mampu menangani ketidakseimbangan data dan secara konsisten menunjukkan hasil evaluasi yang kuat pada metrik precision, recall, dan f1-score.
- Fitur-fitur penting yang paling berkontribusi terhadap prediksi persetujuan kredit meliputi:
 - MB_count, AMT_INCOME_TOTAL, AGE, dan YEARS_EMPLOYED.

Recommendation:

- Implementasi Model: Model XGBoost yang telah dikembangkan dapat digunakan sebagai basis sistem screening awal aplikasi kredit untuk mempercepat proses dan mengurangi risiko.
- Pentingnya Fitur: Faktor-faktor seperti aktivitas finansial pemohon (MB_count) dan durasi kerja (YEARS_EMPLOYED) harus menjadi pertimbangan utama dalam evaluasi kredit.
- Kebijakan Kredit: Institusi dapat lebih fokus pada pemohon dengan usia kerja mapan dan riwayat transaksi stabil untuk meningkatkan tingkat persetujuan kredit yang bertanggung jawab.
- Pengembangan Selanjutnya:
 - Integrasikan fitur tambahan seperti skor kredit dari biro kredit (jika tersedia).
 - Lakukan uji coba model pada data real-time untuk evaluasi lebih lanjut.



THANK YOU

