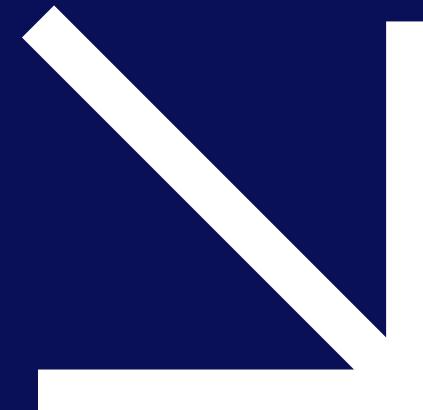


# PROJECT BASED INTERNSHIP



PRESENTED BY

NOVITA CHELSEA



[Link GitHub - Novita Chelsea](#)

# PREDIKSI KEBERHASILAN PEMBAYARAN PINJAMAN MENGGUNAKAN MACHINE LEARNING DI HOME CREDIT

## LATAR BELAKANG

Home Credit merupakan perusahaan fintech global yang menyediakan berbagai produk pembiayaan untuk konsumen, dengan fokus utama pada pinjaman pribadi dan kredit konsumen. Tujuan dari proyek ini adalah untuk membangun sebuah sistem prediksi yang dapat membantu Home Credit meningkatkan akurasi penilaian kredit melalui penerapan machine learning. Proyek ini berfokus pada penerapan tiga model machine learning utama, yaitu Logistic Regression, Random Forest, dan XGBoosting, untuk memprediksi kelayakan kredit nasabah. Logistic Regression digunakan sebagai model dasar untuk pemodelan klasifikasi, sementara Random Forest digunakan sebagai model yang lebih kompleks dengan kemampuan untuk menangani variabel yang lebih banyak dan interaksi yang lebih rumit antar fitur. XGBoost dipilih karena kemampuannya dalam menangani dataset besar dan ketepatannya dalam meningkatkan akurasi prediksi dengan teknik boosting.

Pada proyek kali ini, saya terlibat untuk mengembangkan solusi berbasis data menggunakan machine learning yang dapat meningkatkan kemampuan Home Credit Indonesia dalam memprediksi risiko gagal bayar pinjaman.

Secara khusus, tujuan saya adalah untuk:

1. Meningkatkan akurasi prediksi risiko gagal bayar dengan memanfaatkan data dan algoritma machine learning untuk menghasilkan model yang lebih akurat dalam menilai kelayakan kredit.
2. Mengurangi risiko kerugian finansial yang timbul akibat pemberian pinjaman kepada nasabah dengan potensi gagal bayar yang tinggi.
3. Mengoptimalkan keputusan pemberian pinjaman yang didasarkan pada analisis data yang lebih mendalam.

TUJUAN

# ALUR KERJA

## DATA UNDERSTANDING

mencakup berbagai aktivitas untuk menggali lebih dalam informasi yang terdapat dalam dataset data

## DATA PRE-PROCESSING

Melakukan pembersihan data, transformasi data (encoding), normalisasi data, memastikan kualitas untuk tahap pemodelan.

## DATA VISUALIZATION

Membuat visualisasi untuk memahami distribusi data dan hubungan antara variabel.

## MODELLING

Menggunakan metrik evaluasi seperti accuracy, precision, recall, F1-score, dan ROC AUC score untuk menilai kinerja setiap model.

# DATA UNDERSTANDING

Dataset yang digunakan dalam proyek ini terdiri dari 307,512 baris dan 122 beberapa fitur seperti jenis pekerjaan, pendapatan, tingkat pendidikan, riwayat pinjaman. Target yang ingin diprediksi adalah kolom TARGET dengan 1 sebagai kesulitan membayar kredit dan 0 tidak mengalami kesulutan.

▶ train.shape  
→ (307511, 122)

train.dtypes	
SK_ID_CURR	int64
TARGET	int64
NAME_CONTRACT_TYPE	object
CODE_GENDER	object
FLAG_OWN_CAR	object
...	...
AMT_REQ_CREDIT_BUREAU_DAY	float64
AMT_REQ_CREDIT_BUREAU_WEEK	float64
AMT_REQ_CREDIT_BUREAU_MON	float64
AMT_REQ_CREDIT_BUREAU_QRT	float64
AMT_REQ_CREDIT_BUREAU_YEAR	float64
122 rows × 1 columns	

Numerik = 122, Kategorik = 16

# DATA PRE-PROCESSING

## HANDLING MISSING VALUE

```
[23] train.fillna(train.select_dtypes(include=['float64','int64']).median(), inplace=True)
train.isnull().sum()
```

## ENCODING

```
[23] import pandas as pd
from sklearn.preprocessing import LabelEncoder

# Kolom untuk Label Encoding (kolom biner dan ordinal)
label_columns = ['CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE',
                 'FLAG_PHONE', 'FLAG_EMAIL', 'TARGET']

label_encoder = LabelEncoder()

# Apply Label Encoding
for col in label_columns:
    train[col] = label_encoder.fit_transform(train[col])

[24] # Kolom untuk One-Hot Encoding (kolom nominal)
one_hot_columns = ['NAME_CONTRACT_TYPE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
                   'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE', 'WEEKDAY_APPR_PROCESS_START',
                   'ORGANIZATION_TYPE', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE',
                   'EMERGENCYSTATE_MODE']

# Apply One-Hot Encoding
train = pd.get_dummies(train, columns=one_hot_columns, drop_first=True) # drop_first=True untuk menghindari dummy variable trap
train.head()
```

## STANDARISASI

```
[27] from sklearn.preprocessing import StandardScaler

# Inisialisasi StandardScaler
scaler = StandardScaler()

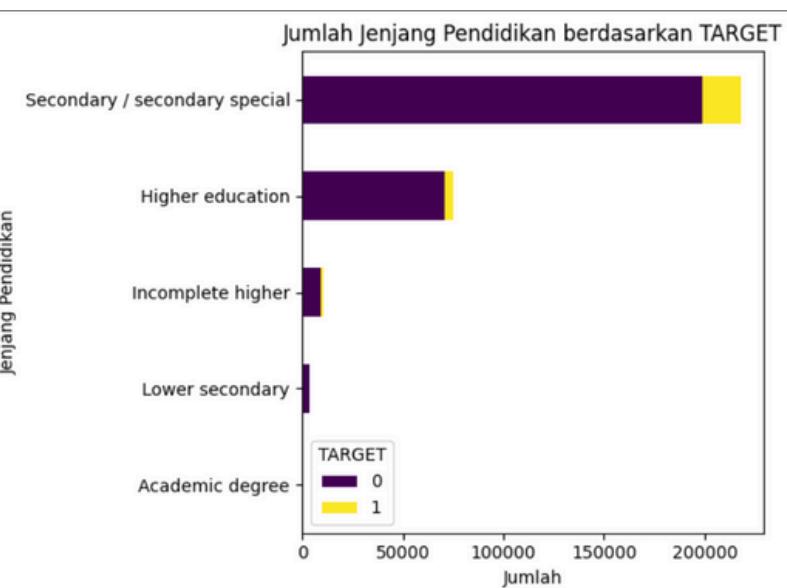
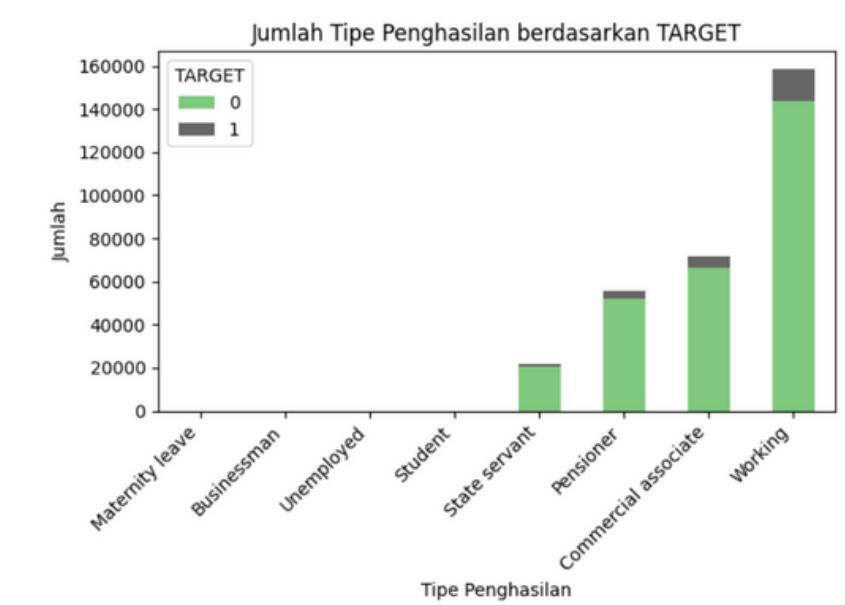
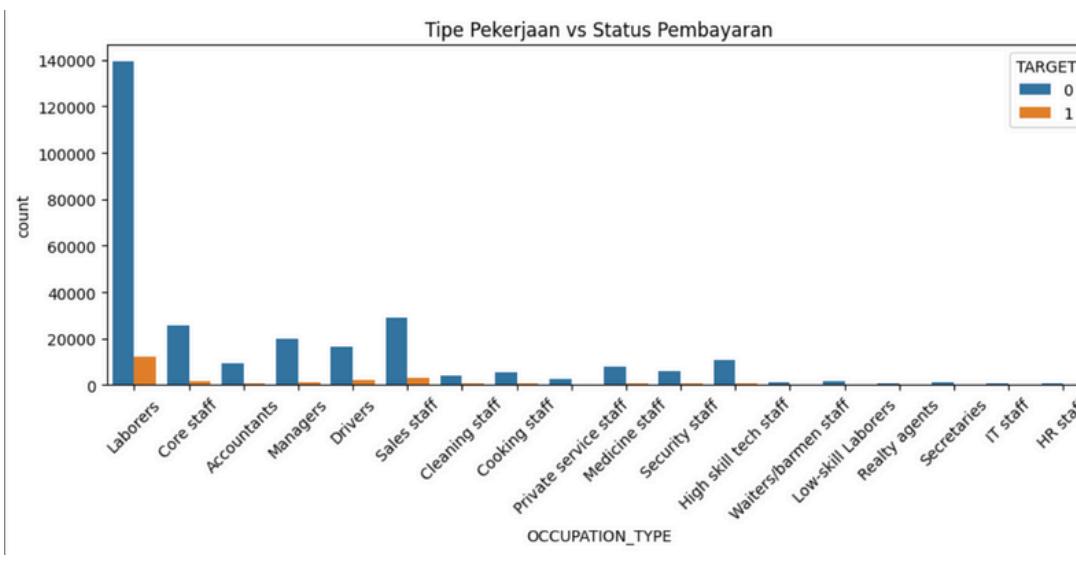
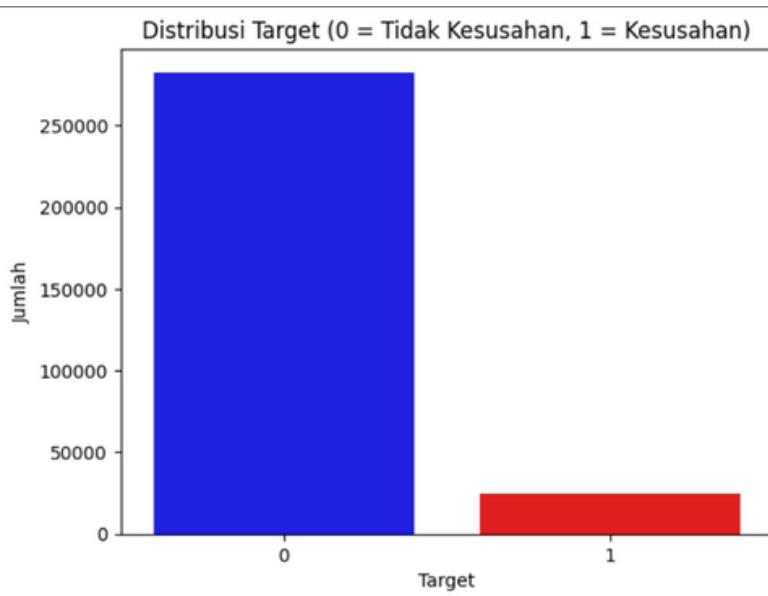
# Fit dan transform data latih
X_train = scaler.fit_transform(X_train)

# Transform data uji dengan scaler yang sama
X_test = scaler.transform(X_test)
```

## SMOTE

```
[28] smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

# DATA VISUALIZATION



Terlihat bahwa hampir semua tipe penghasilan didominasi oleh target 0 (tidak kesusahan). Distribusi data sangat tidak seimbang.

Tipe pekerjaan “Laborers” memiliki jumlah tertinggi dan juga menyumbang jumlah TARGET 1 (kesusahan) paling besar.

Tipe penghasilan dengan jumlah tertinggi pada TARGET 1 (kesusahan) adalah “Working”, meskipun proporsinya masih kecil dibandingkan dengan TARGET 0.

Mayoritas orang memiliki pendidikan Secondary / secondary special, diikuti oleh Higher education.

# MODELLING

NO	ALGORITMA	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC
1	LOGISTIC REGRESION	0.6926	0.89	0.69	0.76	0.7358
2	RANDOM FOREST	0.9094	0.87	0.91	0.89	0.6996
3	XGBOOSTING	0.9109	0.88	0.91	0.89	0.7217

XGBoosting direkomendasikan karena akurasi 91.09%, tertinggi di antara semua model, menunjukkan bahwa model ini paling banyak membuat prediksi yang benar secara keseluruhan. Memberikan keseimbangan terbaik antara akurasi tinggi, kemampuan menangkap kasus positif (recall), dan f1-score yang tinggi.

# RECOMMENDATION

## Sesuaikan Penawaran dengan Penghasilan dan Tanggungan Nasabah

Sesuaikan jumlah pinjaman dan tenor berdasarkan penghasilan dan jumlah tanggungan nasabah. Nasabah dengan penghasilan terbatas atau banyak tanggungan sebaiknya diberikan pinjaman dalam jumlah yang lebih kecil dan dengan tenor yang lebih panjang.

## Sistem Reward dan Denda untuk Nasabah

Setiap kali nasabah melakukan pembayaran tepat waktu, mereka mendapatkan poin reward. Untuk nasabah yang sering terlambat atau gagal membayar pinjaman, turunkan limit kredit mereka untuk pinjaman berikutnya atau hentikan akses ke produk pinjaman lain sampai pembayaran tertunda diselesaikan.

## Personalisasi Penawaran Produk Berdasarkan Data

Gunakan analisis data untuk lebih memahami kebutuhan individu nasabah dan menawarkan produk pinjaman yang benar-benar disesuaikan. Misalnya, berdasarkan pengeluaran, kebiasaan pembayaran, dan status pekerjaan, nasabah dapat diberikan rekomendasi pinjaman yang lebih relevan.