

zenius

**Kampus
Merdeka**
INDONESIA JAYA

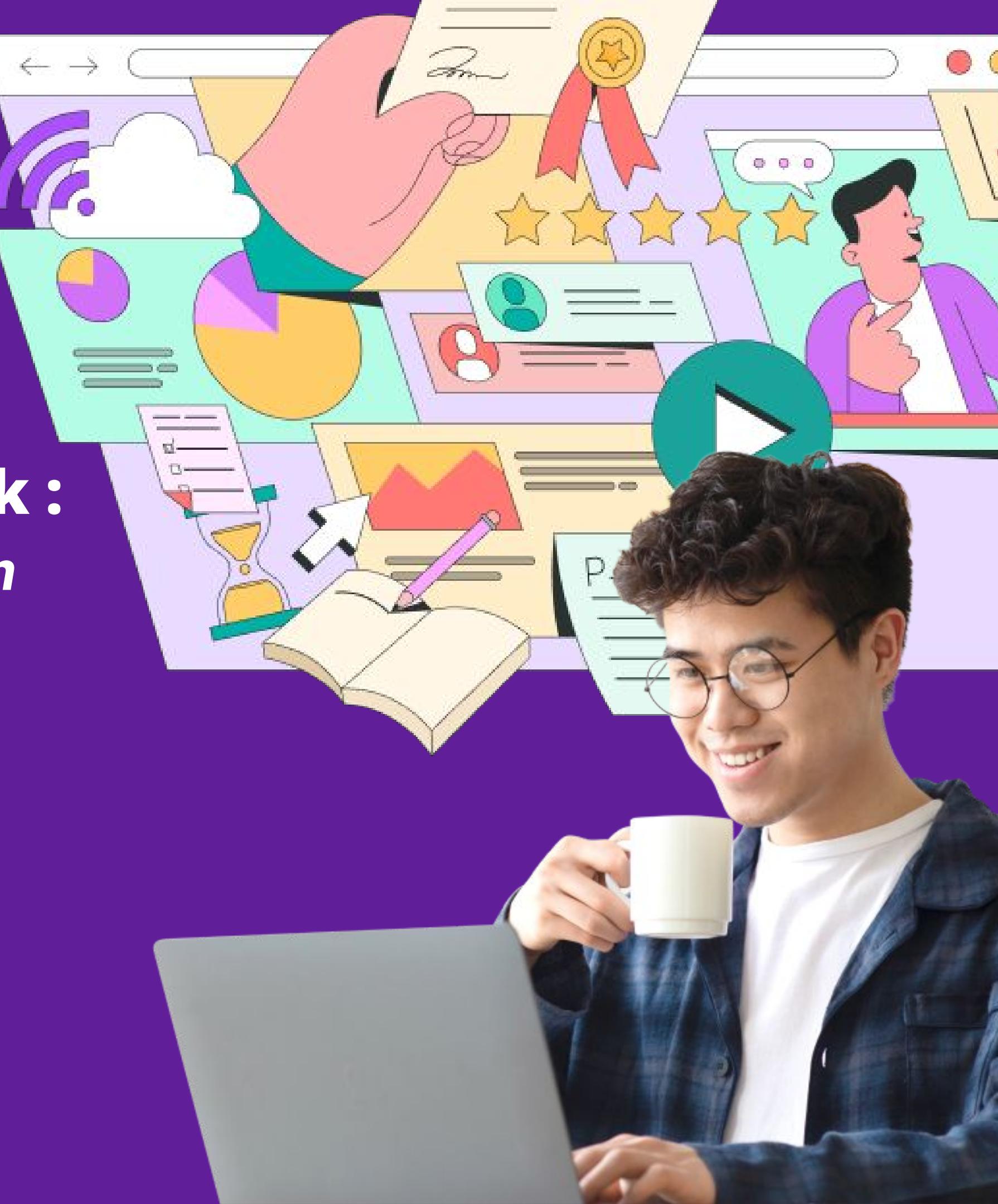
Prediksi Home Credit Default Risk : *Algoritma Logistic Regression, Random Forest, dan Decision Tree*

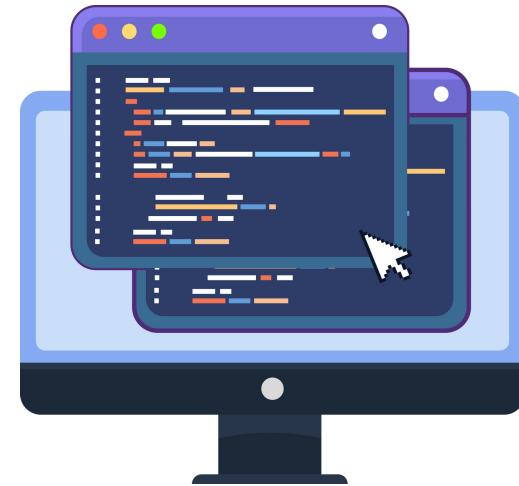
Kelompok 8

Mentor: Dede Brahma Arianto

Final Project Data Analytics

**Program Zenius Studi Independen Bersertifikat
Bersama Kampus Merdeka**





Google Colab

<https://colab.research.google.com/drive/1O6PMNtV8GL8bYolOv2cvtbIXhV8pS748?usp=sharing>



Dashboard

<https://lookerstudio.google.com/reporting/11f85d45-754f-4b35-8ad6-a25c22561299>

OUR TEAM



Melani Safwa Aprila
Universitas Diponegoro



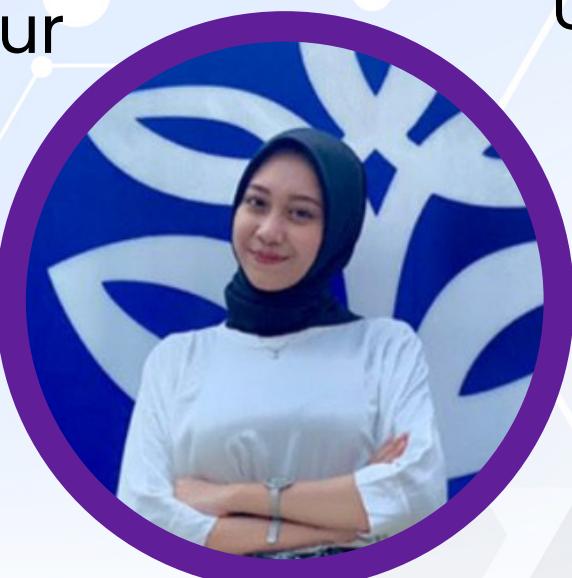
Syifa Saskia Elfaretta
UPN "Veteran" Jawa Timur



Widiawati
Universitas Galuh



Jeremia Maheswara A. S
Universitas Semarang



Nadila Rahmawati
Institut Pertanian Bogor

Outline



1. Business Understanding

2. Data Understanding

3. Data Preparation

4. Data Analysis

5. Dashboard

6. Deployment

Business Understanding

HOME CREDIT

Home Credit sebagai perusahaan pembiayaan dengan layanan **cicilan** berpotensi memiliki nasabah yang gagal bayar kredit.

Berdasarkan data Home Credit Group terdapat **8,1%** nasabah yang berisiko mengalami gagal bayar kredit. Hal ini salah satunya disebabkan oleh data riwayat kredit yang tidak tersedia.

Oleh karena itu, diperlukan model yang dapat memprediksi kemampuan pembayaran atau risiko kredit tiap nasabah.

Business Objective

- **Tujuan:** mengurangi kerugian perusahaan akibat adanya nasabah yang gagal bayar kredit.
- **Business Success Criteria:** berkurangnya pelanggan yang gagal bayar sehingga akan meningkatkan margin keuntungan dan mengurangi kerugian perusahaan.

Situation Assessment

Requirements: data yang berisi informasi tentang pelanggan. Variabel target mengindikasikan status pembayaran kredit. Variabel lainnya yang dibutuhkan dalam analisis ini, di antaranya jenis kelamin, pendidikan, usia, dan dokumen-dokumen identitas nasabah lainnya yang digunakan untuk pinjaman.

Data Mining Goals

Model prediksi yang dikembangkan, yaitu ***logistic regression, random forest regression, dan decision tree***

Tujuan Model: mempelajari pola-pola dari atribut-atribut dalam dataset dan membangun hubungan antara atribut dengan status pembayaran kredit.

Ketiga model akan dibandingkan tingkat akurasinya.

Model Success Criteria

Akurasi prediksi model minimal 65%

Produce Project Plan

Dataset

<https://www.kaggle.com/competitions/home-credit-default-risk/dataLinks to an external site.>

Dari data tersebut, akan dibuat solusi machine learning untuk permasalahan *credit scoring* dengan langkah-langkah mengikuti framework CRISP-DM.

Business Understanding : 2 Juni – 4 Juni 2023

Data Understanding : 4 Juni – 6 Juni 2023

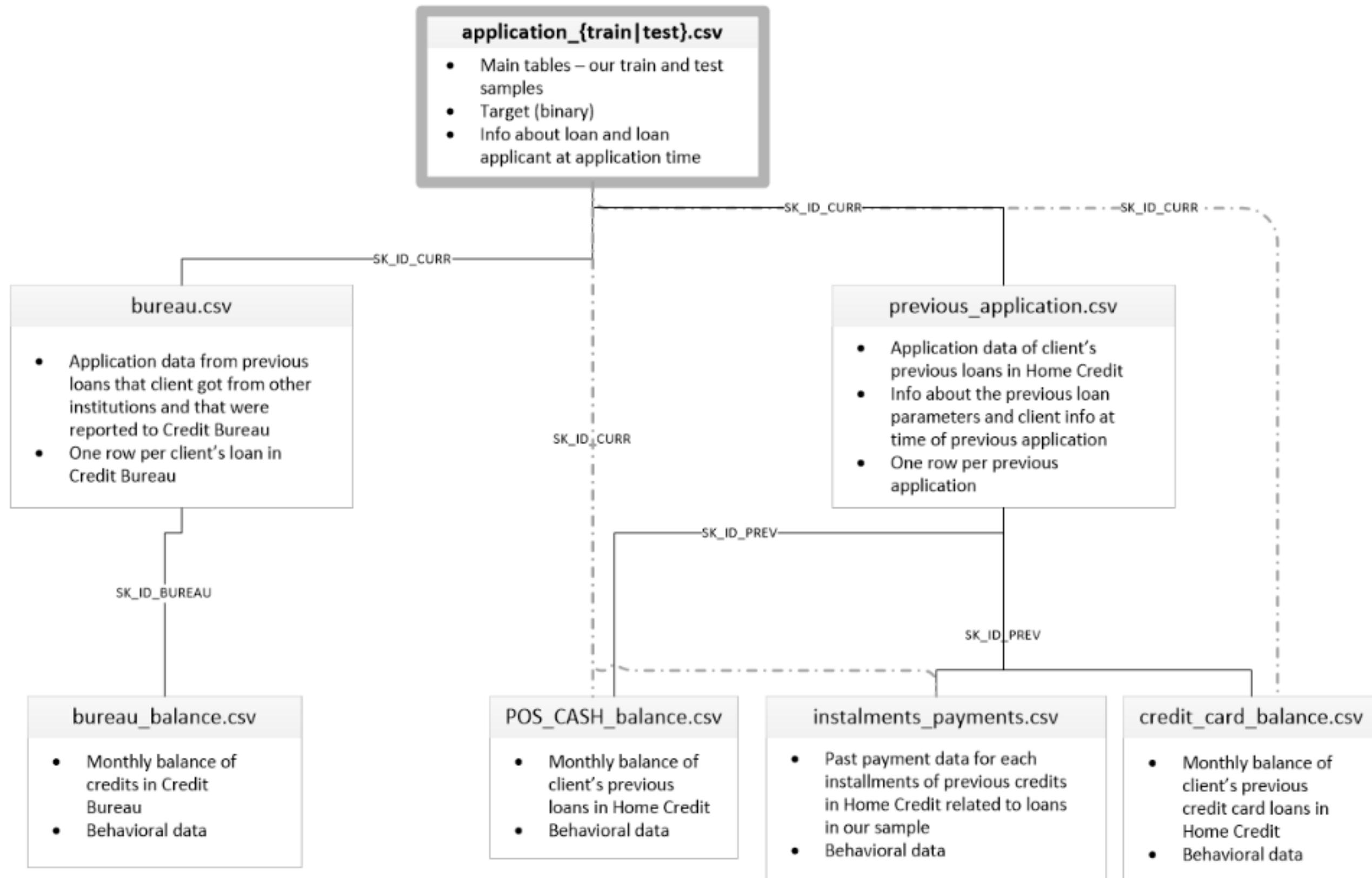
Data Preparation : 6 Juni – 11 Juni 2023

Data Analysis : 11 Juni – 13 Juni 2023

Dashboard : 13 Juni – 15 Juni 2023

Data Understanding

Dataset Home Credit Default Risk



Ada 7 Sumber Data yang Dapat diambil:

1. application_{train|test}.csv:

- Dataset ini terdiri dari data pelatihan dan pengujian utama dengan informasi tentang setiap aplikasi pinjaman di Home Credit. Dataset train memiliki column target sedangkan dataset test tidak memiliki column target.

2. bureau.csv:

- Dataset ini berisi semua kredit klien sebelumnya yang diberikan oleh lembaga keuangan lain yang dilaporkan ke Biro Kredit.

3. bureau_balance.csv:

- Dataset yang berisi tentang saldo bulanan kredit sebelumnya di Biro Kredit.

4. POS_CASH_bureau.csv:

- Dataset yang berisi saldo bulanan dari POS (point of sales) sebelumnya dan pinjaman tunai yang dimiliki pemohon dengan Home Credit.

5. credit_card_balance.csv:

- Dataset yang berisi saldo bulanan kartu kredit sebelumnya yang dimiliki pemohon dengan Home Credit.

6. previous_application.csv:

- Dataset yang terdiri dari data mengenai aplikasi sebelumnya untuk pinjaman di Home Credit klien yang memiliki pinjaman dalam data aplikasi.

7. installments_payments.csv:

- Dataset yang terdiri dari riwayat pembayaran untuk kredit yang dicairkan sebelumnya di Home Credit terkait dengan pinjaman.

Menggunakan *data application_train* dan *application_test*:

1. *application_train.csv*:

- Terdiri dari 307,511 Records, 122 Columns
- Dataset memiliki missing values
- Dataset memiliki outliers
- Target (0 - klien dengan kasus lain, 1 - klien dengan kesulitan pembayaran)

2. *application_test.csv*:

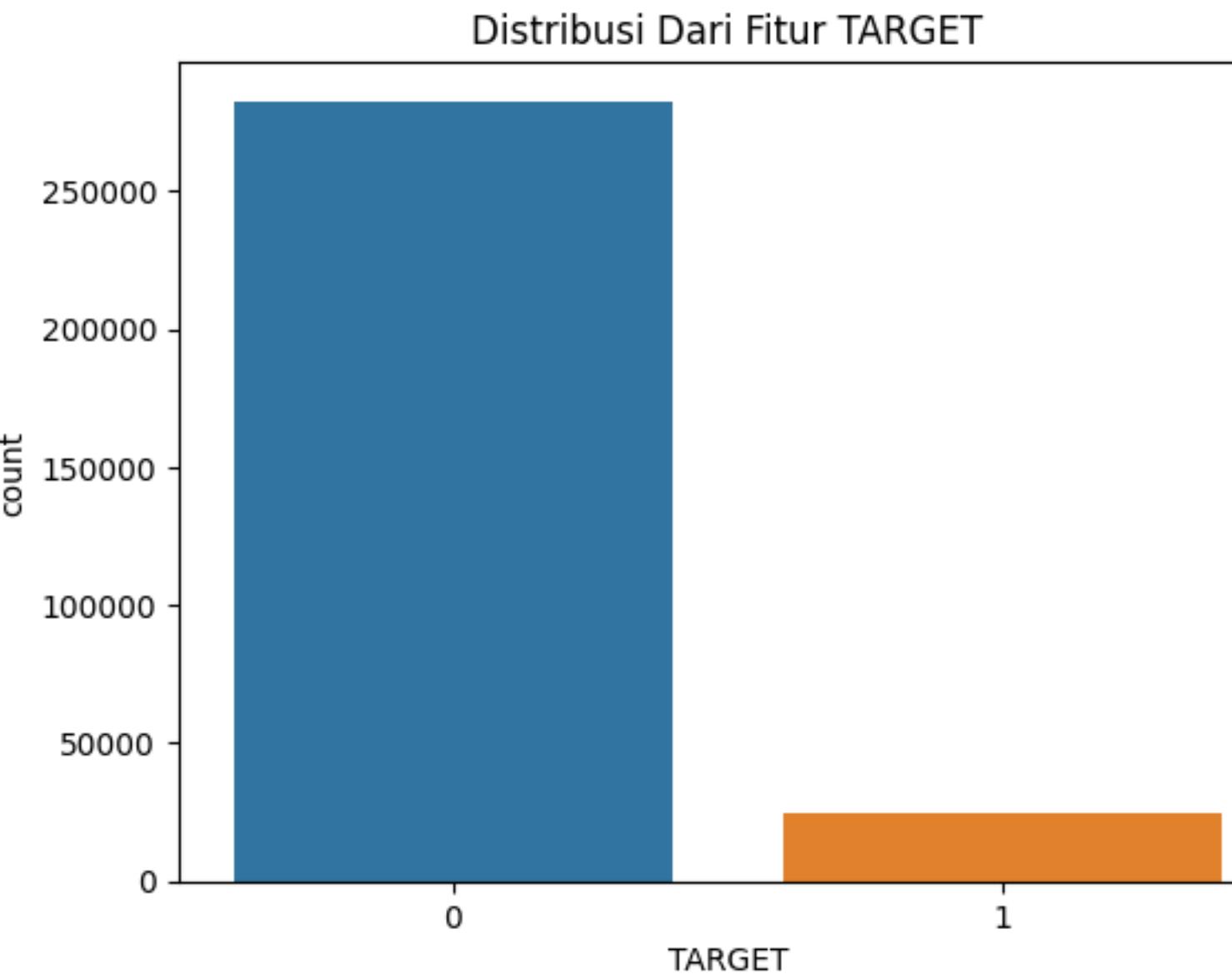
- Terdiri dari 48,744 Records, 121 Columns
- Dataset memiliki missing values
- Dataset memiliki outliers

Fitur-fitur yang digunakan:

- **TARGET** = Data target (1 - klien dengan kesulitan membayar, 0 - kasus lainnya)
- **CODE_GENDER** = Data gender berisi F (Female) dan M (Male)
- **NAME_EDUCATION_TYPE** = Berisi tingkat pendidikan tertinggi yang dicapai klien
- **DAYS_BIRTH** = Usia klien dalam hari pada saat melakukan peminjaman
- **DAYS_ID_PUBLISH** = Berapa hari sebelum klien mengubah dokumen identitas yang dia gunakan untuk pinjaman
- **REGION_RATING_CLIENT** = Peringkat untuk wilayah tempat tinggal klien (1,2,3)
- **REGION_RATING_CLIENT_W_CITY** = Peringkat untuk wilayah tempat klien tinggal dengan mempertimbangkan kota (1,2,3)
- **REG_CITY_NOT_WORK_CITY** = Berisi tanda jika alamat tetap klien tidak cocok dengan alamat kantor (1=berbeda, 0=sama, di tingkat kota)
- **DAYS_LAST_PHONE_CHANGE** = Banyak hari sebelum klien mengganti ponsel

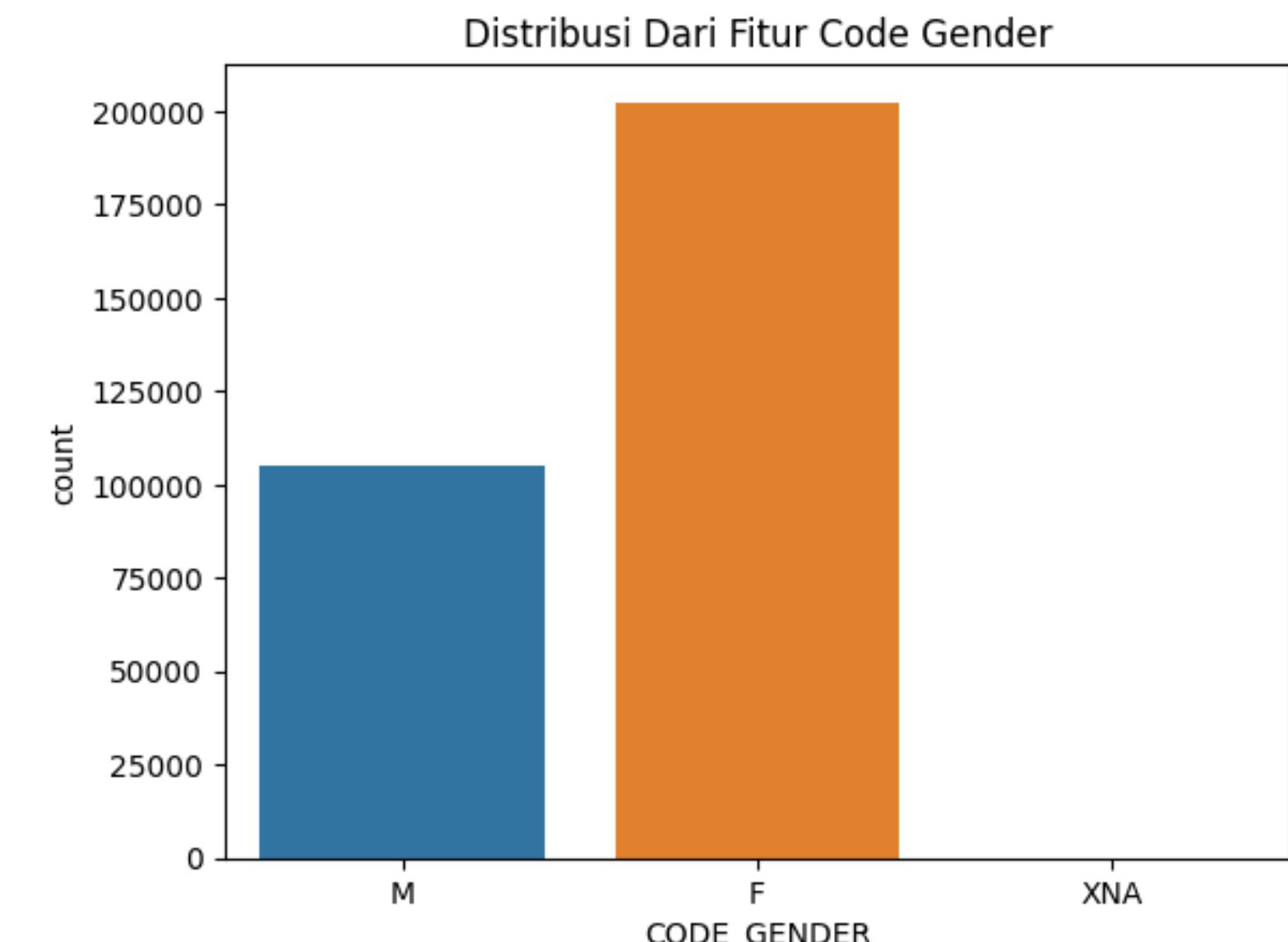
Fitur **TARGET**

- 0 = 282.686
- 1 = 24.825
- Tidak Ada Missing Values



Fitur **CODE_GENDER**

- F = 202.448
- M = 105.059
- XNA = 4
- Tidak Ada Missing Values



Data Preparation

Handle Missing Values



Menangani nilai-nilai yang hilang atau kosong

Data Train memiliki 122 kolom dengan 67 kolom missing values

Data Test memiliki 121 kolom dengan 64 kolom missing values

1. Drop Column

- Hapus kolom yang memiliki nilai null di atas 50%

2. Fillna(median)

- Isi kolom numerik dengan nilai median

3. Fillna(mode)

- Isi kolom kategorik dengan nilai modus

```
[ ] missing_values = missing_val(df_train)  
missing_values.head(50)
```

Your selected dataframe has 81 columns.
There are 0 columns that have missing values.
Missing Values % of Total Values

```
[ ] missing_values_test = missing_val_test(df_test)  
missing_values_test.head(50)
```

Your selected dataframe has 92 columns.
There are 0 columns that have missing values.

Missing Values % of Total Values

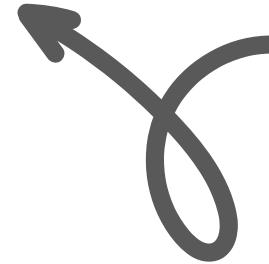
Categorical Data Encoding



Mengubah variabel kategorikal menjadi numerikal

- Menggunakan library **LabelEncoder** untuk mengubah kolom tipe kategorik menjadi numerik dengan proses pelabelan
- Terdapat **13 kolom** pada data **train** yang diubah pada proses ini
- Sedangkan pada data **test** terdapat **15 kolom**

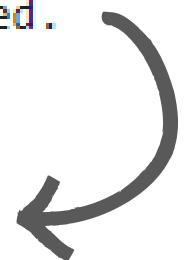
Data Test



```
NAME_CONTRACT_TYPE  
CODE_GENDER  
FLAG_OWN_CAR  
FLAG_OWN_REALTY  
NAME_TYPE_SUITE  
NAME_INCOME_TYPE  
NAME_EDUCATION_TYPE  
NAME_FAMILY_STATUS  
NAME_HOUSING_TYPE  
OCCUPATION_TYPE  
WEEKDAY_APPR_PROCESS_START  
ORGANIZATION_TYPE  
EMERGENCystate_MODE  
15 columns were label encoded.
```

```
NAME_CONTRACT_TYPE  
CODE_GENDER  
FLAG_OWN_CAR  
FLAG_OWN_REALTY  
NAME_TYPE_SUITE  
NAME_INCOME_TYPE  
NAME_EDUCATION_TYPE  
NAME_FAMILY_STATUS  
NAME_HOUSING_TYPE  
OCCUPATION_TYPE  
WEEKDAY_APPR_PROCESS_START  
ORGANIZATION_TYPE  
EMERGENCystate_MODE  
13 columns were label encoded.
```

Data Train



Memilih Feature



Seleksi fitur yang memiliki korelasi kuat dengan target

- Variabel dependen yang digunakan ialah '**TARGET**', sehingga dilakukan..
- Seleksi fitur yang memiliki **korelasi > 0.05** dengan '**TARGET**'
- Didapat **9 fitur**

9 fitur korelasi kuat (> 0.05) dengan 'TARGET'



highly correlated feature:

```
Index(['TARGET', 'CODE_GENDER', 'NAME_EDUCATION_TYPE', 'DAYS_BIRTH',  
       'DAYS_ID_PUBLISH', 'REGION_RATING_CLIENT',  
       'REGION_RATING_CLIENT_W_CITY', 'REG_CITY_NOT_WORK_CITY',  
       'DAYS_LAST_PHONE_CHANGE'],  
      dtype='object')
```

No. of highly correlated features: 9

Ganti Value Negatif



Value yang negatif diubah menjadi positif

Kolom dengan value negatif:

- **DAY_S_BIRTH**
- **DAY_S_ID_PUBLISH**
- **DAY_S_LAST_PHONE_CHANGE**

```
df_train[['DAY_S_BIRTH', 'DAY_S_ID_PUBLISH', 'DAY_S_LAST_PHONE_CHANGE']].apply(lambda x: abs(x))
```

Diubah menjadi positif menggunakan **apply(lambda x: abs(x))**

- **apply**: menerapkan fungsi yang terdapat di dalam kurung, dalam hal ini lambda x
- **x**: value dari ketiga kolom yang dipanggil
- **lambda x**: mengambil setiap argumen x dan mengembalikan nilai,
- **abs(x)**: absolut (nilai numerik positif) dari x

Output setelah diubah menjadi positif

	DAY_S_BIRTH	DAY_S_ID_PUBLISH	DAY_S_LAST_PHONE_CHANGE
0	9461	2120	1134.0
1	16765	291	828.0
2	19046	2531	815.0
3	19005	2437	617.0
4	19932	3458	1106.0

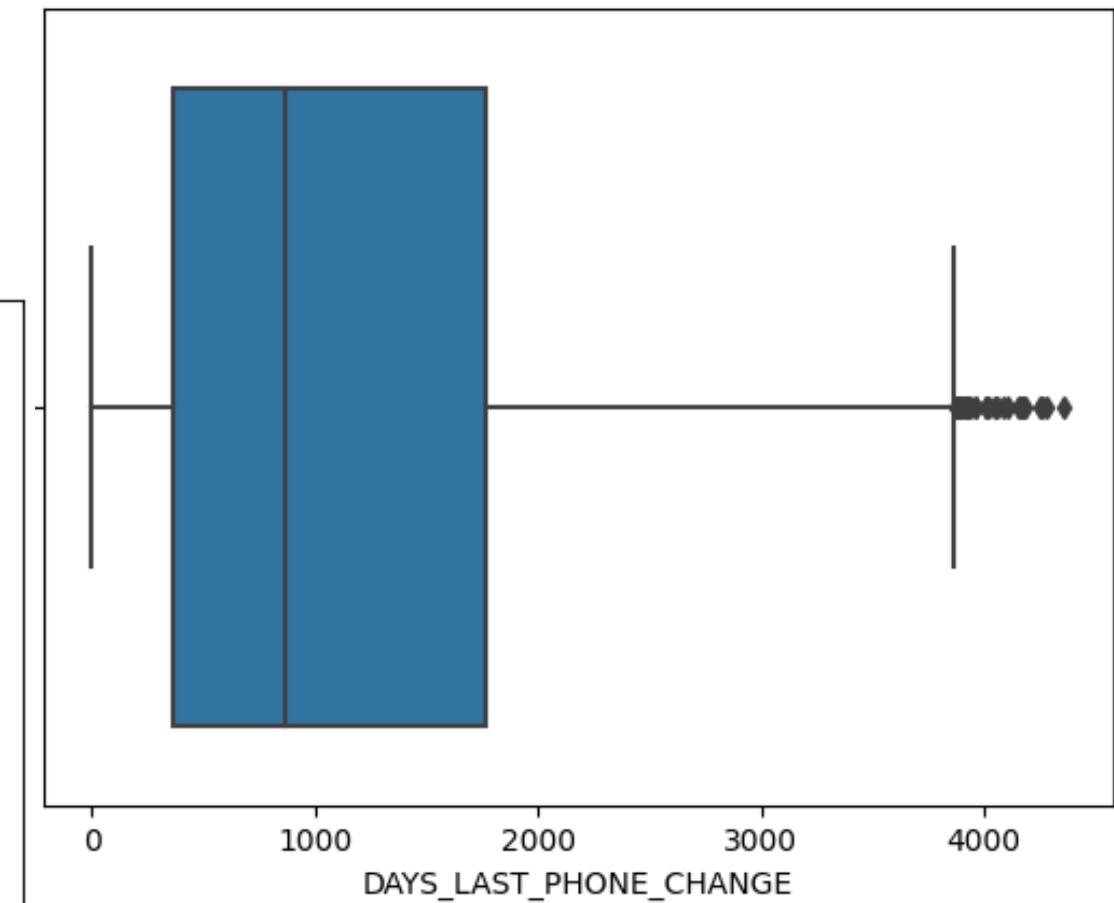
Handle Outliers



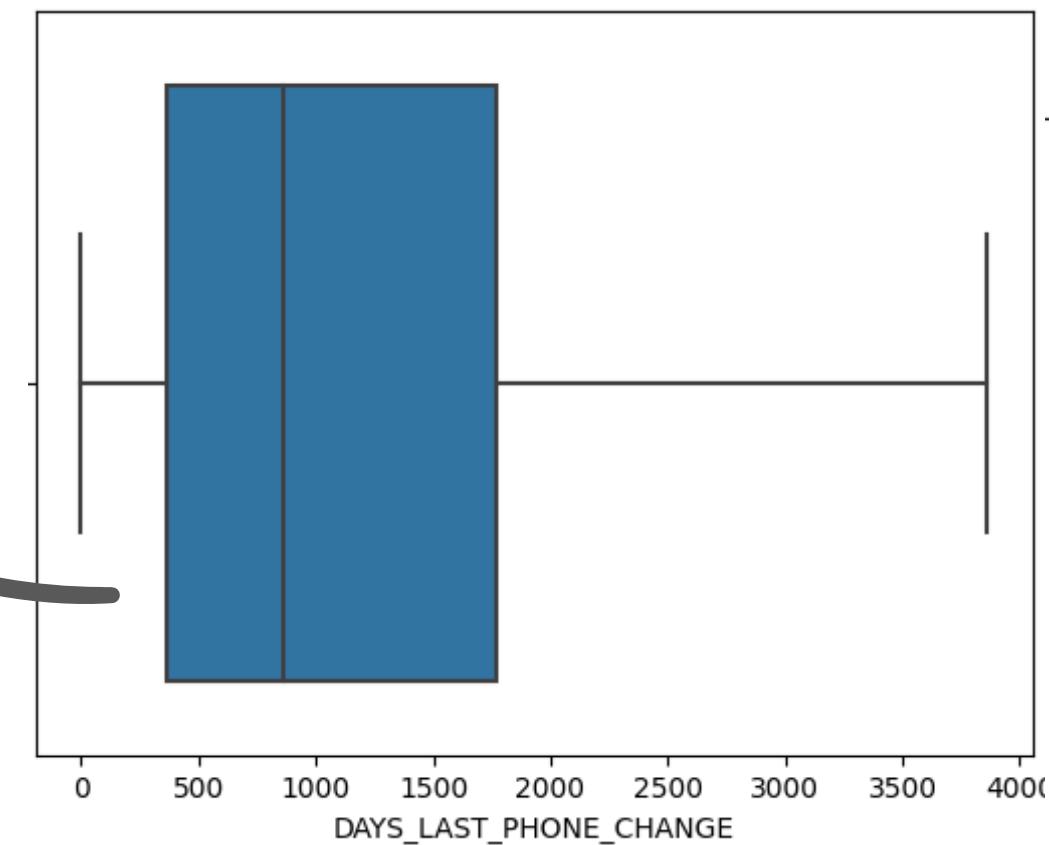
Menangani nilai dengan pola yang jauh dari umumnya

- Outliers ditemukan pada kolom '**DAYs_LAST_PHONE_CHANGE**'
- Handle dilakukan dengan metode **IQR (Interquartile Range)**
- Metode ini didasarkan pada **perhitungan jarak antara kuartil pertama (Q1) dan kuartil ketiga (Q3)** dalam distribusi data

DAYs_LAST_PHONE_CHANGE
sebelum handle outliers



DAYs_LAST_PHONE_CHANGE
setelah handle outliers



Data Analysis

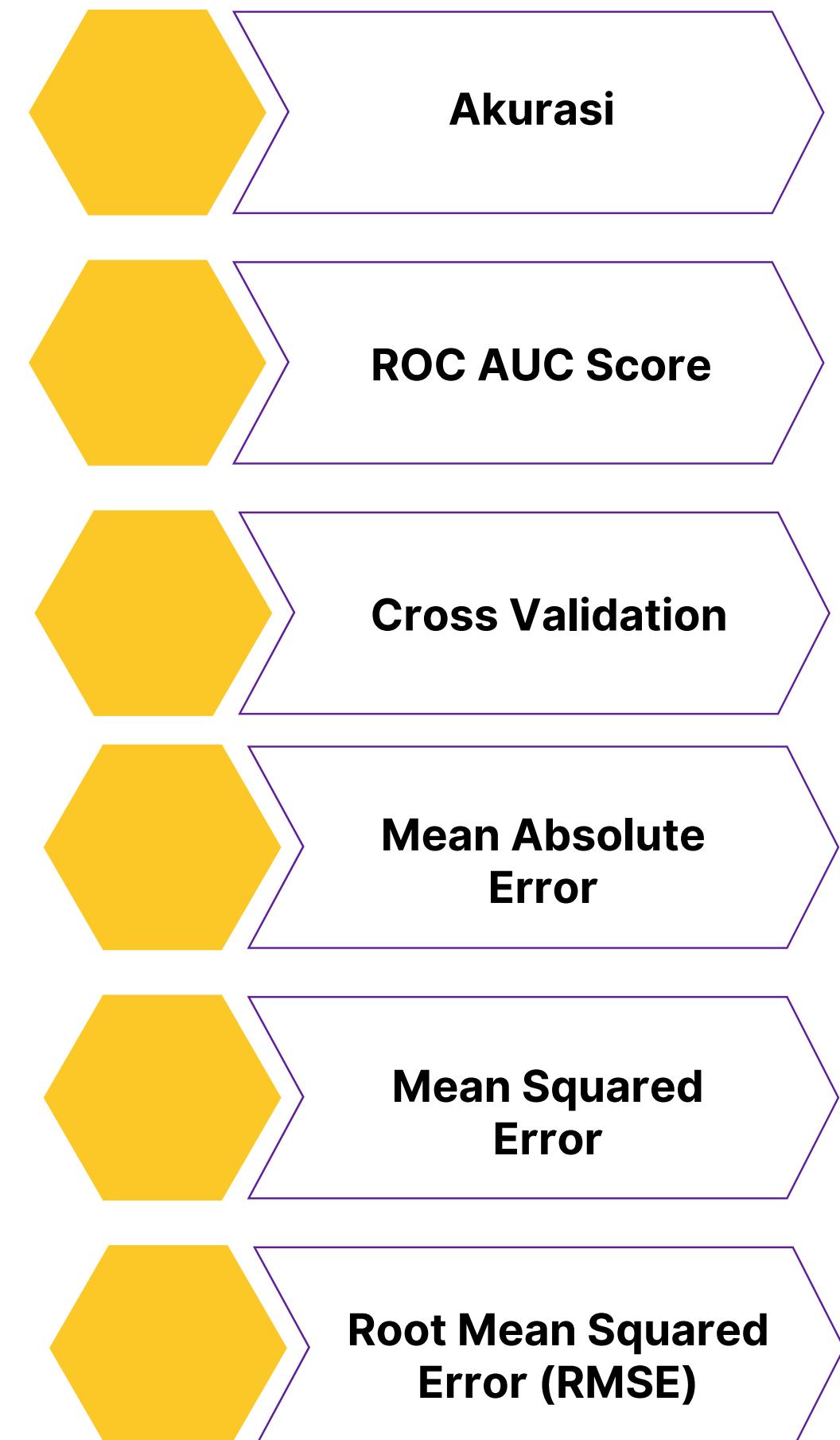
Model Technique

Beberapa model yang dipertimbangkan dalam prediksi Home Credit Default Risk:

- ***Logistic Regression***: teknik yang dipinjam dari bidang statistik dengan pembelajaran mesin. Model ini merupakan strategi masuk untuk masalah klasifikasi biner.
- ***Decision Tree***: bentuk pembelajaran mesin yang diawasi di mana data terus dibagi dengan parameter tertentu.
- ***Random Forest***: ansambel Pohon Keputusan, umumnya dilatih melalui metode pengantongan.

T
E
S
T

D
E
S
I
G
N



Build & Asses Model

Feature yang digunakan berdasarkan korelasi tertinggi (>0.05):

- TARGET
- CODE_GENDER
- NAME_EDUCATION_TYPE
- DAYS_BIRTH
- DAYS_ID_PUBLISH
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- REG_CITY_NOT_WORK_CITY
- DAYS_LAST_PHONE_CHANGE

TARGET tidak balance, sehingga lakukan undersampling :

```
from imblearn.under_sampling import NearMiss
nm = NearMiss(version=3, n_neighbors_ver3=3)
x, y = nm.fit_resample(x, y)
```

Pembagian train test split menggunakan perbandingan 75 : 25

```
x_train, x_test, y_train, y_test = train_test_split(x, y, stratify=y, test_size=0.25,
print(x_train.shape, x_test.shape)
print(y_test.shape)
```

Lakukan fitting model Logistic Regression, Decision Tree, dan Random Forest:

```
LR = LogisticRegression()
LR.fit(x_train, y_train)
```

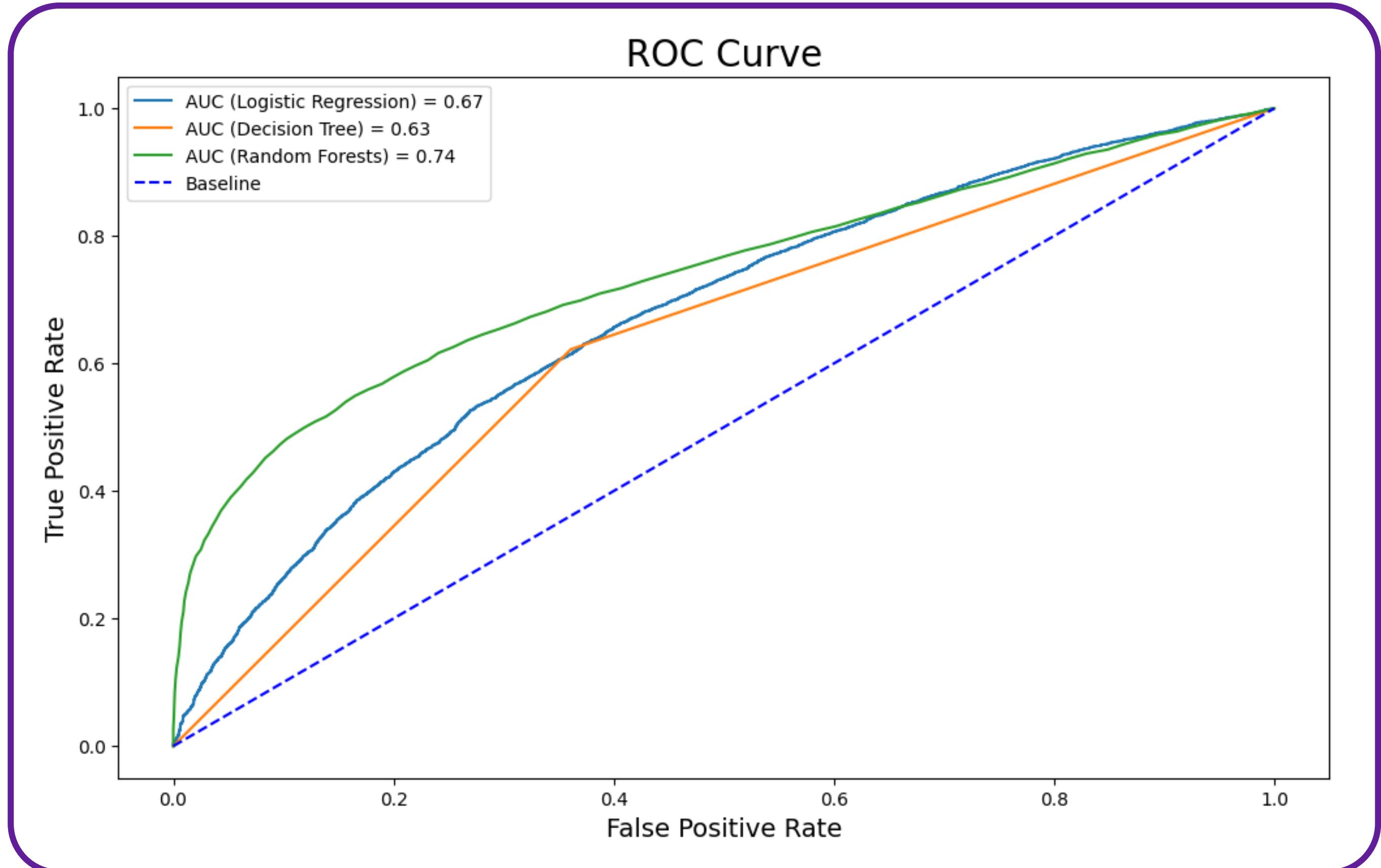
```
DT = DecisionTreeClassifier()
DT.fit(x_train, y_train)
```

```
RF = RandomForestClassifier()
RF.fit(x_train, y_train)
```

Setelah dilakukan permodelan, didapatkan evaluasi metrics dari ketiga model sebagai berikut:

	Model	Akurasi	MAE	MSE	RMSE	Cross Validation	AUC Score
0	Logistic Regression	0.627246	0.372754	0.372754	0.610536	61.42	0.672962
1	Decision Tree	0.630549	0.369451	0.369451	0.607825	62.98	0.630548
2	Random Forest	0.685733	0.314267	0.314267	0.560595	67.93	0.737920

Kinerja dari ketiga model yang digunakan berdasarkan ROC AUC Score dan evaluation metrics sebelumnya, terlihat bahwa model yang terbaik untuk melakukan prediksi adalah **Random Forest**

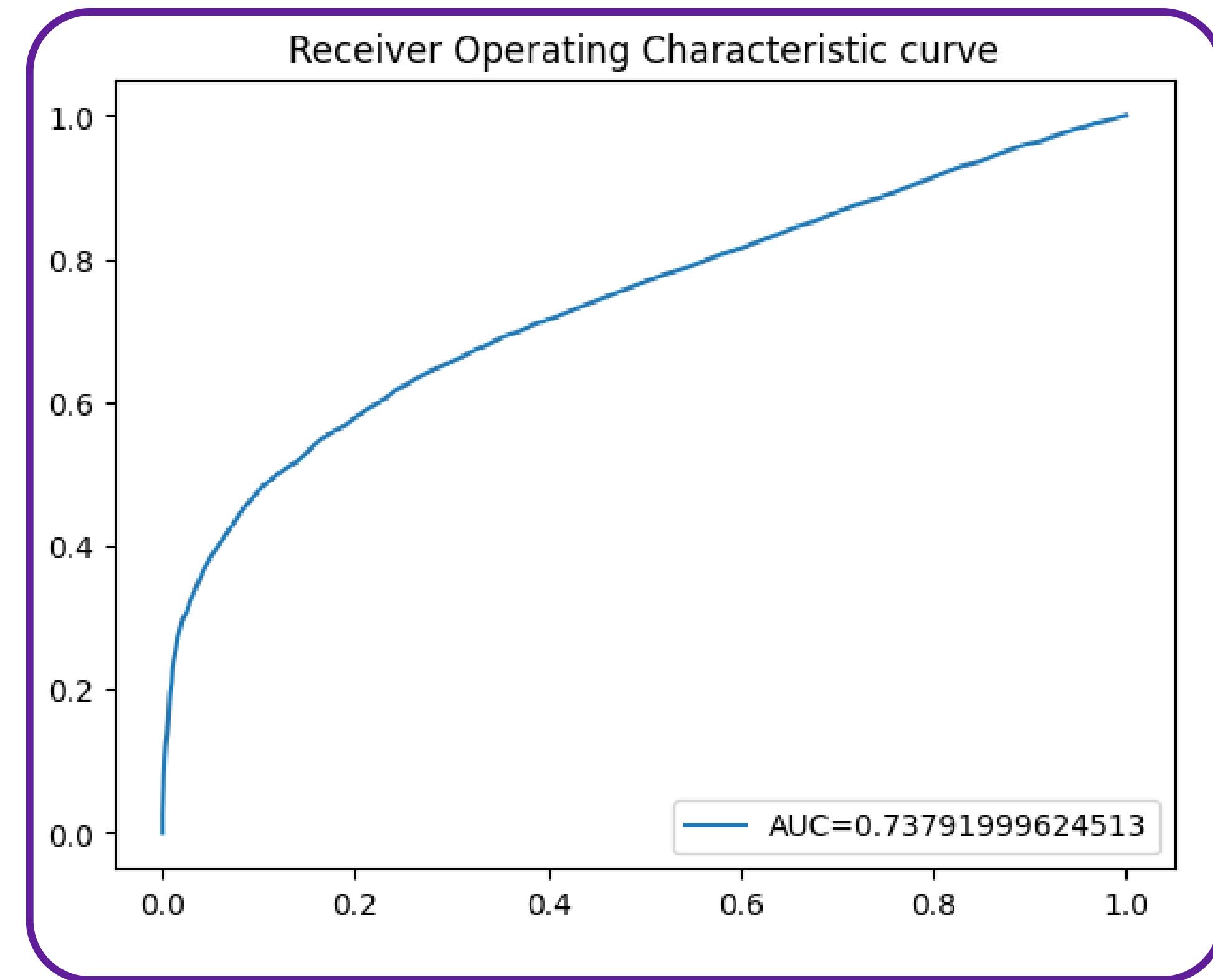


Best Model: Random Forest

Dari evaluasi metric beberapa model sebelumnya, didapatkan bahwa model terbaik yang dapat digunakan untuk prediksi gagal bayar kredit adalah Random Forest.

Untuk ukuran kinerja model ini lebih lanjut dapat dilihat pada report berikut:

	precision	recall	f1-score	support
0	0.67	0.75	0.70	6207
1	0.71	0.63	0.67	6206
accuracy			0.69	12413
macro avg	0.69	0.69	0.68	12413
weighted avg	0.69	0.69	0.68	12413



Best Model: Random Forest

Hasil prediksi menggunakan random forest didapatkan bahwa nilai target yang kesulitan membayar kredit telah banyak menurun hingga terdapat hanya 1 dari 48744 data

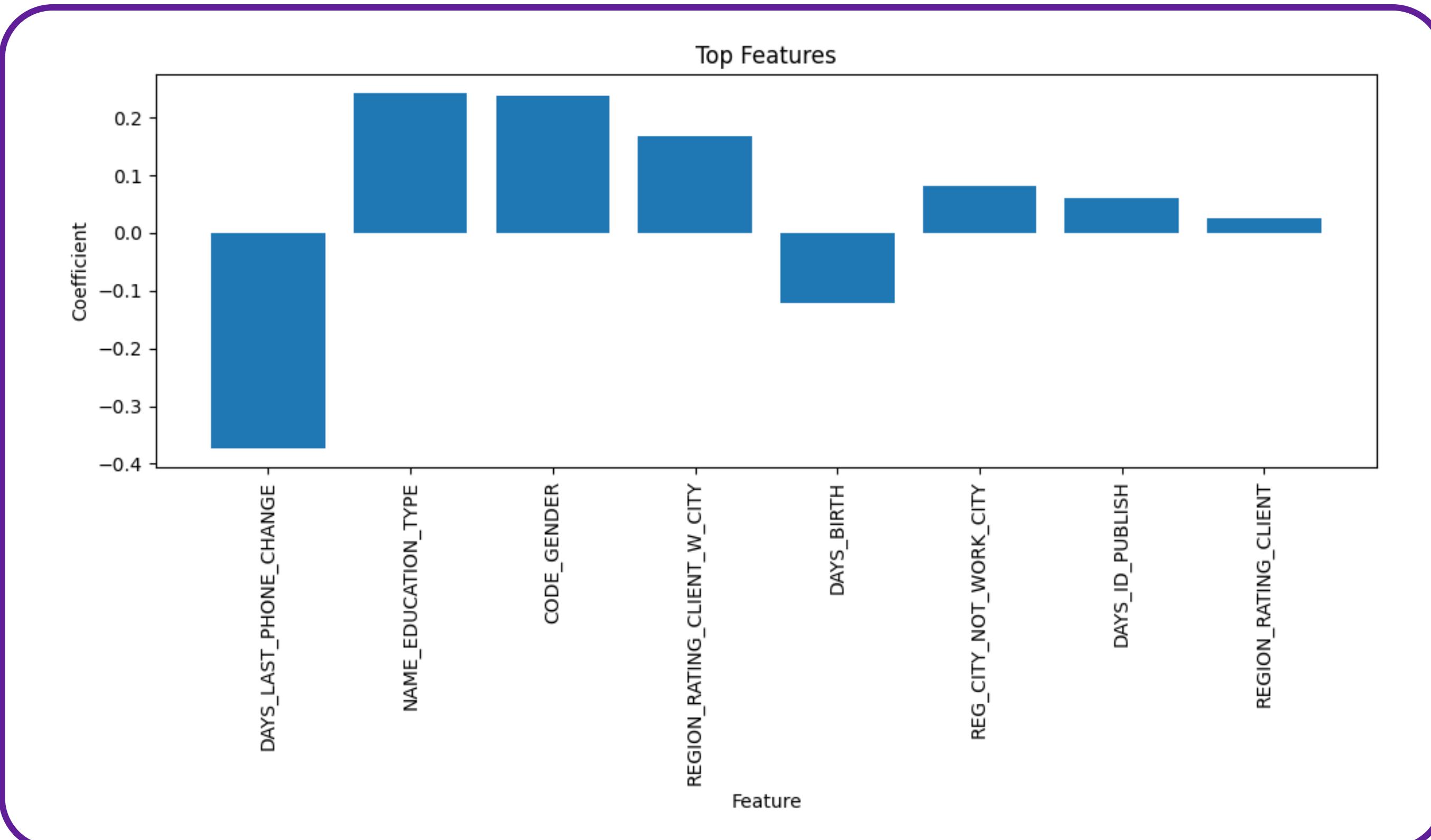
	CODE_GENDER	NAME_EDUCATION_TYPE	DAYS_BIRTH	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	REG_CITY_NOT_WORK_CITY	DAYSLAST_PHONE_CHANGE	TARGET
0	0		1	19241	812	2	2	0	1740.0 0
1	1		4	18064	1623	2	2	0	0.0 0
2	1		1	20038	3503	2	2	0	856.0 0
3	0		4	13976	4208	2	2	0	1805.0 0
4	1		4	13040	4262	2	2	1	821.0 0
...
48739	0		4	19970	3399	3	3	0	684.0 0
48740	0		4	11186	3003	2	2	1	0.0 0
48741	0		4	15922	1504	2	2	0	838.0 0
48742	1		1	13968	1364	2	2	1	2308.0 0
48743	0		4	13962	4220	2	2	0	327.0 0

48744 rows × 9 columns

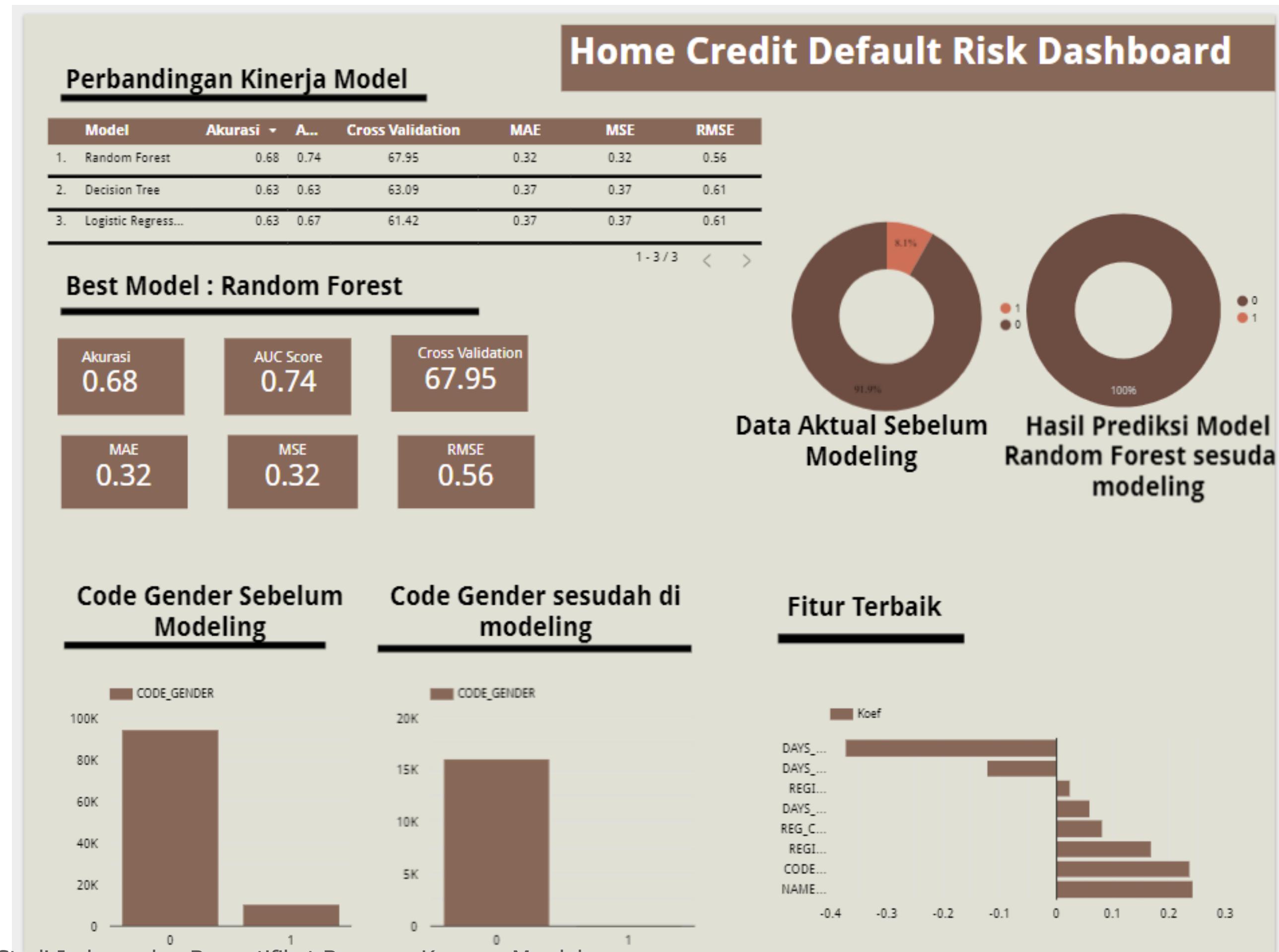
	CODE_GENDER	NAME_EDUCATION_TYPE	DAYS_BIRTH	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	REG_CITY_NOT_WORK_CITY	DAYSLAST_PHONE_CHANGE	TARGET
34791	1		4	16460	0	1	1	0	0.0 1

Best Model: Random Forest

Feature - feature yang paling berpengaruh dalam proses prediksi dalam model ini adalah:

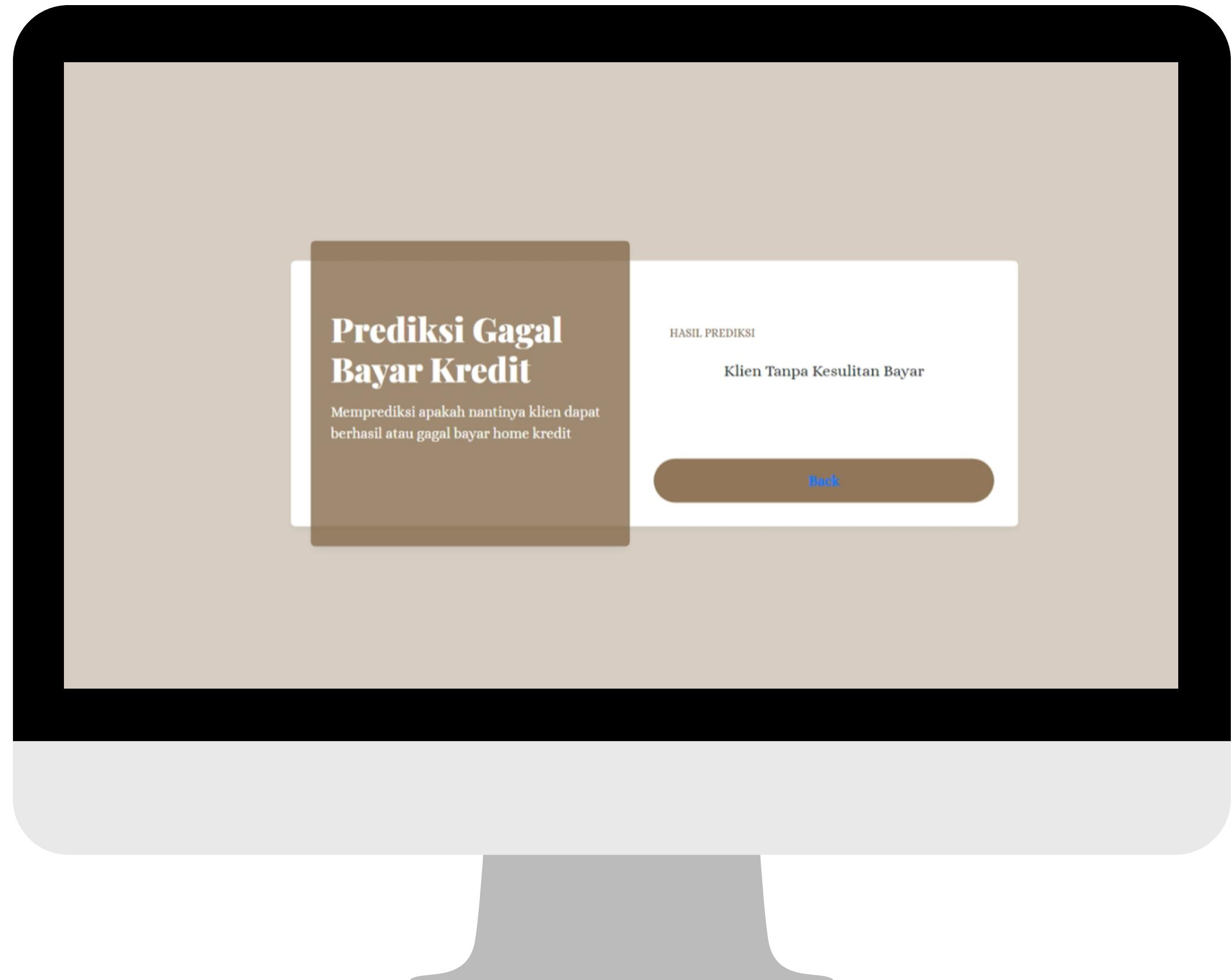


Dashboard



Deployment





Thank you!

