



FINAL PROJECT

HOME CREDIT DEFAULT RISK

**GROUP 20 DATA ANALYTICS
STUDI INDEPENDEN ZENIUS**

GROUP 20



Saniya Adelia

Business
Understanding



Ila Syamitha

Data
Understanding



Jihan Pradita

Fitriani
Data Preparation



Tahta Yodya
Setiya Kusuma C
Data Analysis



M. Abdul Aziz
Data Visualizaion



CONTENT



01

BUSINESS UNDERSTANDING

02

DATA UNDERSTANDING

03

DATA PREPARATION

04

DATA ANALYSIS

05

DATA VISUALIZATION

BUSINESS UNDERSTANDING



TUJUAN / KONTEKS BISNIS

- Didirikan pada 1997, Home Credit adalah penyedia pinjaman konsumen yang beroperasi di 8 negara.
- Visi “....Secara bertanggung jawab memberikan layanan keuangan tepercaya....”
- Misi “....Terus meningkatkan manajemen risiko dengan memanfaatkan teknologi canggih....”
- Nilai perusahaan yaitu Kecerdasan Digital (Digital Savviness), dan Waspada Terhadap Risiko (Risk In Mind).



PROBLEM STATEMENT

- Home Credit memiliki rasio kredit bermasalah (non-performing loans/NPL) sebesar 8,1%.
- Sesuai dengan visi, misi, dan nilai perusahaan, Home Credit melakukan analisis risiko kredit dengan mempertimbangkan informasi eksternal berupa informasi mengenai riwayat kelancaran kredit debitur.



OBJECTIVE

- Metode yang digunakan yaitu Exploratory Data Analysis (EDA).
- EDA merupakan teknik menganalisis dan memahami data sehingga ditemukan tren tersembunyi, pola, hubungan antarvariabel, outlier atau anomali, menguji hipotesis, dan memeriksa asumsi dari data.



STRATEGI LANJUTAN

- Exploratory Data Analysis (EDA) menghasilkan output berupa insight data.
- Untuk mengelola risiko kredit, Home Credit menerapkan insight data pada machine learning.
- Machine learning melakukan credit scoring dengan berfokus membuat profil pelanggan secara komprehensif dan akurat.



DATA UNDERSTANDING



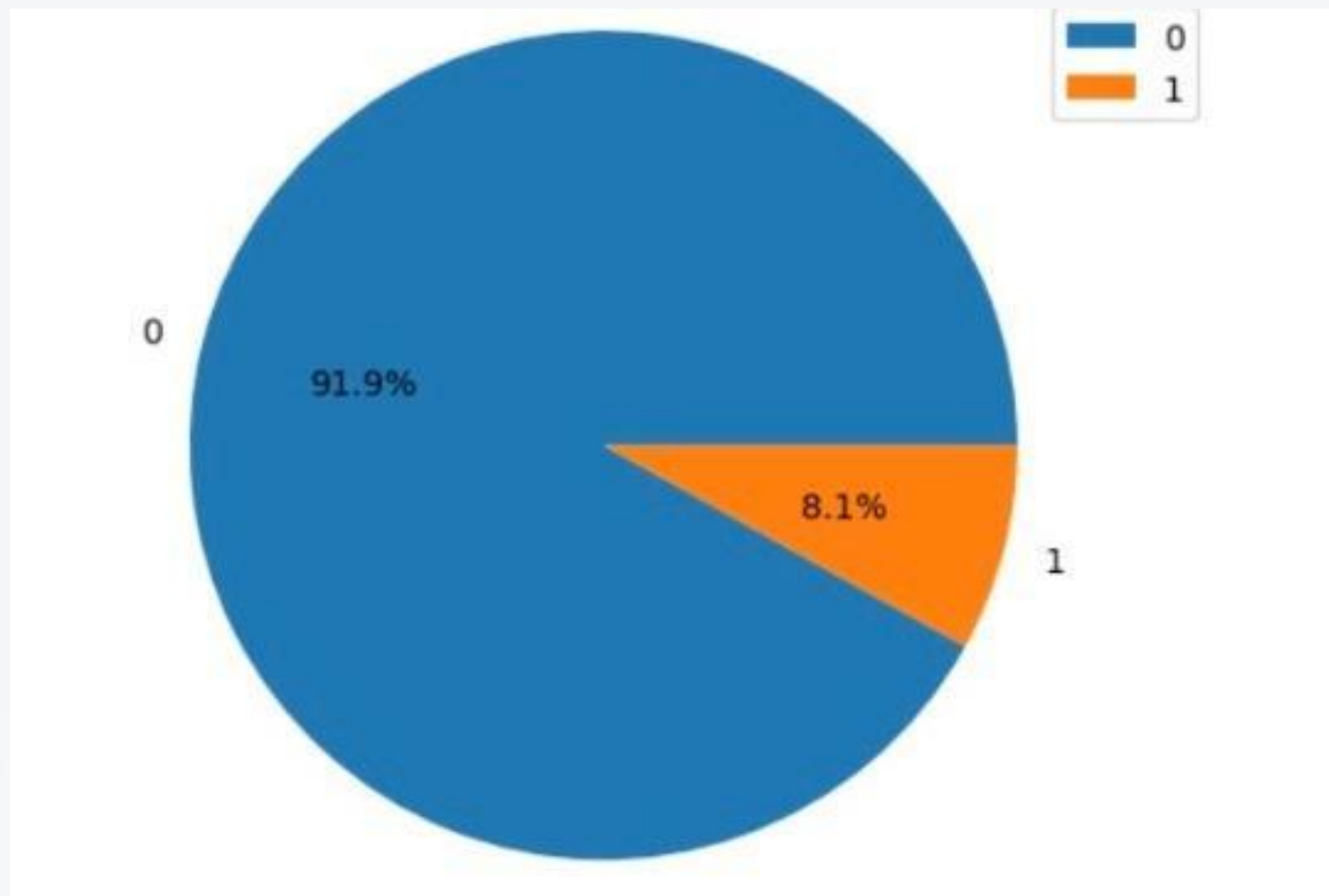
Data understanding adalah sebuah tahapan di dalam metodologi sains data dan pengembangan AI yang bertujuan untuk mendapatkan pemahaman awal mengenai data



MENAMPILKAN DATA FRAME

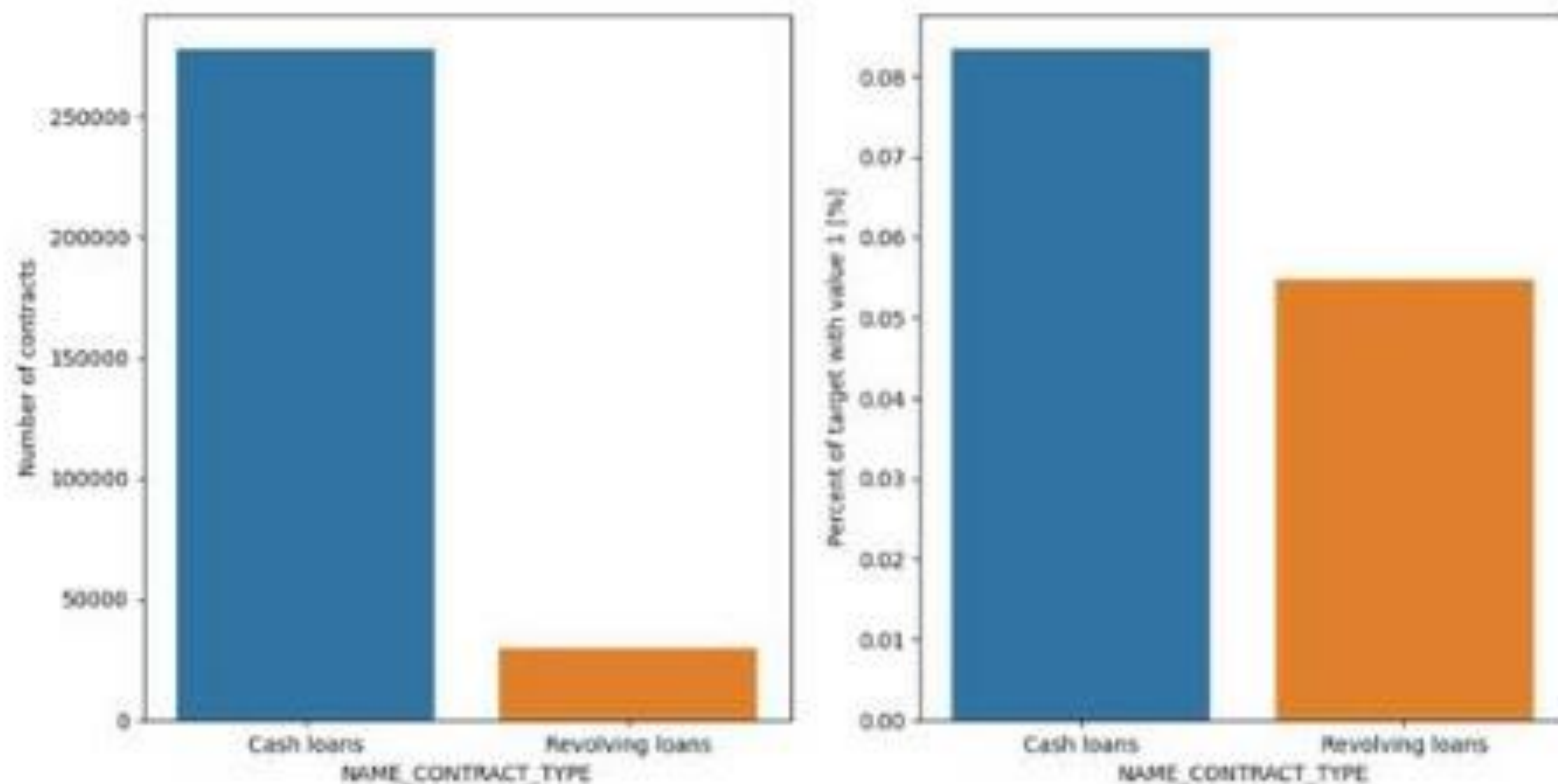
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	351000.0
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	1129000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	135000.0
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	297000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	513000.0

MELIHAT HUTANG KLIEN PADA DATA



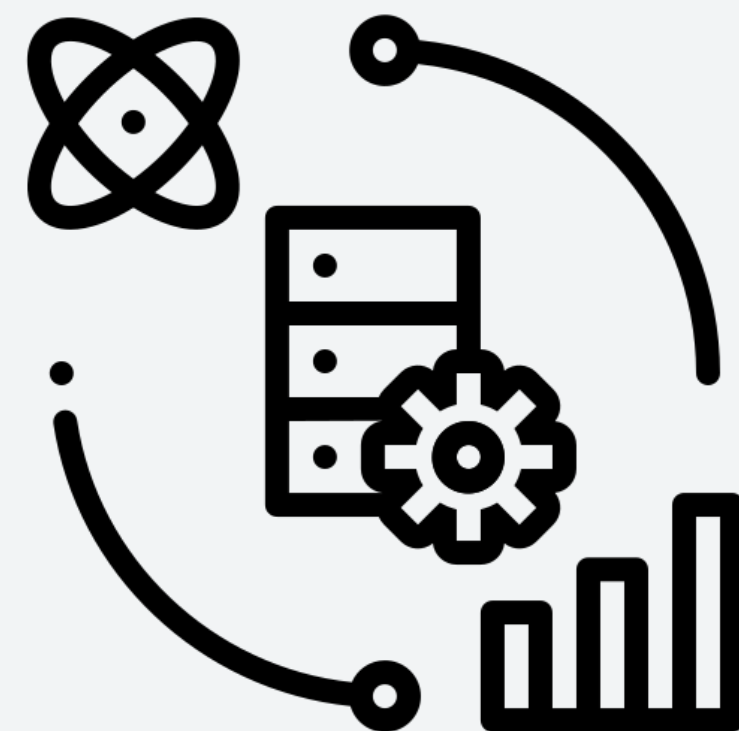
Dimana terdapat 8,1% klien yang tidak bisa melunasi hutang dan ada 91,9% klien yang dapat melunasi hutang





Melihat jumlah kontrak berdasarkan jenis pinjaman yang diambil dan persentase pinjaman bisa di lihat pada gambar di atas

DATA PREPARATION



HANDLING MISSING VALUE

```
float64, (69.9 persen) COMMONAREA_AVG : 214865
float64, (53.3 persen) ELEVATORS_AVG : 163891
float64, (50.3 persen) ENTRANCES_AVG : 154828
float64, (49.8 persen) FLOORSMAX_AVG : 153020
float64, (67.8 persen) FLOORSMIN_AVG : 208642
float64, (59.4 persen) LANDAREA_AVG : 182590
float64, (68.4 persen) LIVINGAPARTMENTS_AVG : 210199
float64, (50.2 persen) LIVINGAREA_AVG : 154350
float64, (69.4 persen) NONLIVINGAPARTMENTS_AVG : 213514
float64, (55.2 persen) NONLIVINGAREA_AVG : 169682
float64, (50.7 persen) APARTMENTS_MODE : 156061
float64, (58.5 persen) BASEMENTAREA_MODE : 179943
float64, (48.8 persen) YEARS_BEGINEXPLUATATION_MODE : 150007
float64, (66.5 persen) YEARS_BUILD_MODE : 204488
float64, (69.9 persen) COMMONAREA_MODE : 214865
float64, (53.3 persen) ELEVATORS_MODE : 163891
float64, (50.3 persen) ENTRANCES_MODE : 154828
float64, (49.8 persen) FLOORSMAX_MODE : 153020
float64, (67.8 persen) FLOORSMIN_MODE : 208642
float64, (59.4 persen) LANDAREA_MODE : 182590
float64, (68.4 persen) LIVINGAPARTMENTS_MODE : 210199
float64, (50.2 persen) LIVINGAREA_MODE : 154350
float64, (69.4 persen) NONLIVINGAPARTMENTS_MODE : 213514
float64, (55.2 persen) NONLIVINGAREA_MODE : 169682
float64, (50.7 persen) APARTMENTS_MEDI : 156061
float64, (58.5 persen) BASEMENTAREA_MEDI : 179943
float64, (48.8 persen) YEARS_BEGINEXPLUATATION_MEDI : 150007
float64, (66.5 persen) YEARS_BUILD_MEDI : 204488
float64, (69.9 persen) COMMONAREA_MEDI : 214865
```

- Melakukan drop kolom dengan missing value $\geq 50\%$
- Menghapus kolom yang tidak diperlukan
- Melakukan filling missing value data numerik dengan menggunakan nilai median
- Melakukan filling missing value data kategorik dengan menggunakan nilai yang paling sering muncul

ENCODING

CODE_GENDER_F	CODE_GENDER_M	NAME_INCOME_TYPE_Pensioner	NAME_INCOME_TYPE_Working	NAME_EDUCATION_TYPE_Higher education
0	1	0	1	0
1	0	0	0	1
0	1	0	1	0
1	0	0	1	0
0	1	0	1	0

- **Label Encoding**
- Penggunaan Label Encoding digunakan untuk mengubah categorical variabel dengan unique ≤ 2 agar lebih mudah untuk mengetahui kategori kelompok dari suatu variable.
- **One Hot Encoding**
- Penggunaan one-hot disini digunakan untuk mengubah categorical variabel dengan unique > 2 agar lebih mudah untuk mengetahui kategori kelompok dari suatu variable.

CORRELATION

```
Feature dengan korelasi kuat:  
Index(['TARGET', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION',  
      'DAYS_ID_PUBLISH', 'FLAG_EMP_PHONE', 'REGION_RATING_CLIENT',  
      'REGION_RATING_CLIENT_W_CITY', 'REG_CITY_NOT_LIVE_CITY',  
      'REG_CITY_NOT_WORK_CITY', 'DAYS_LAST_PHONE_CHANGE', 'CODE_GENDER_F',  
      'CODE_GENDER_M', 'NAME_INCOME_TYPE_Pensioner',  
      'NAME_INCOME_TYPE_Working', 'NAME_EDUCATION_TYPE_Higher education',  
      'NAME_EDUCATION_TYPE_Secondary / secondary special',  
      'ORGANIZATION_TYPE_XNA'],  
      dtype='object')  
Jumlah feature yang berkorelasi kuat: 18
```

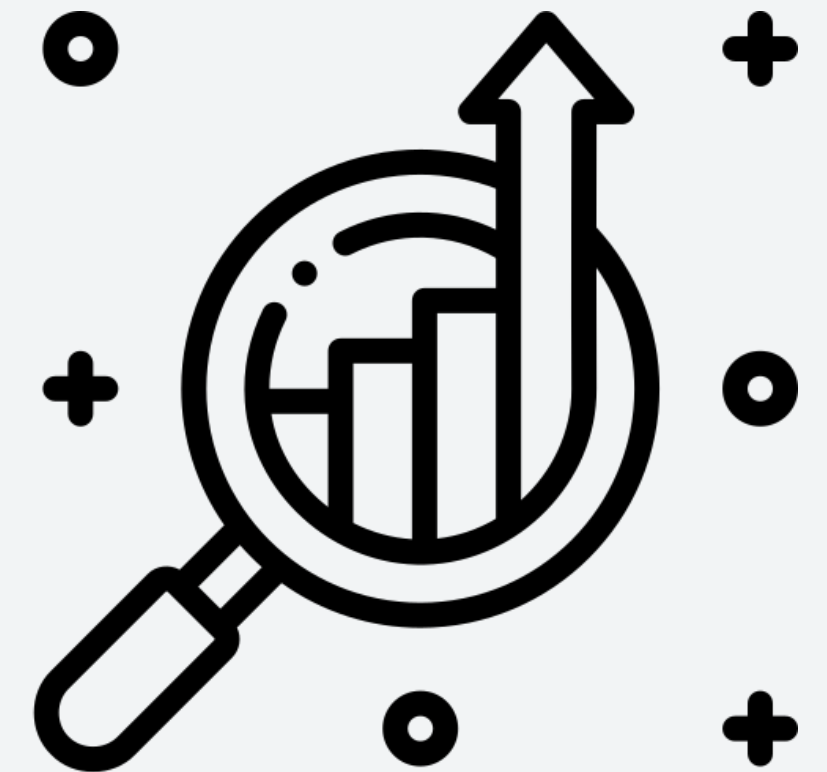
- **Correlation**
 - Mencari nilai korelasi antar feature
 - Mencari nilai korelasi setiap feature terhadap variabel TARGET
 - Memilih feature yang memiliki korelasi kuat terhadap variabel TARGET dengan nilai korelasi $> 0,04$
- **Heatmap**
 - Melihat feature yang memiliki korelasi kuat terhadap variabel TARGET dengan menggunakan Heatmap

FEATURE SELECTION

```
There are 145 columns to remove.  
Application Train shape after removing each variables = (307511, 18)
```

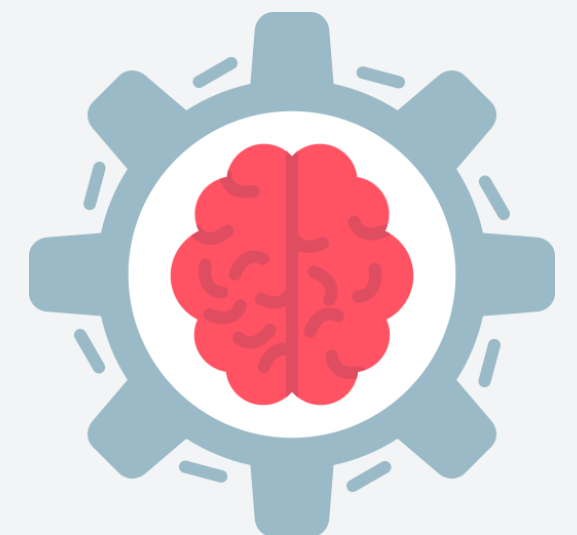
Feature yang dipilih untuk dilakukan modeling adalah data-data yang nilainya absolute atau tidak melihat negatif maupun positifnya dan data yang diambil memiliki korelasi yang cenderung cukup kuat sampai dengan kuat terhadap 'TARGET' berdasarkan tabel korelasi yaitu dengan nilai >0.04

DATA ANALYSIS

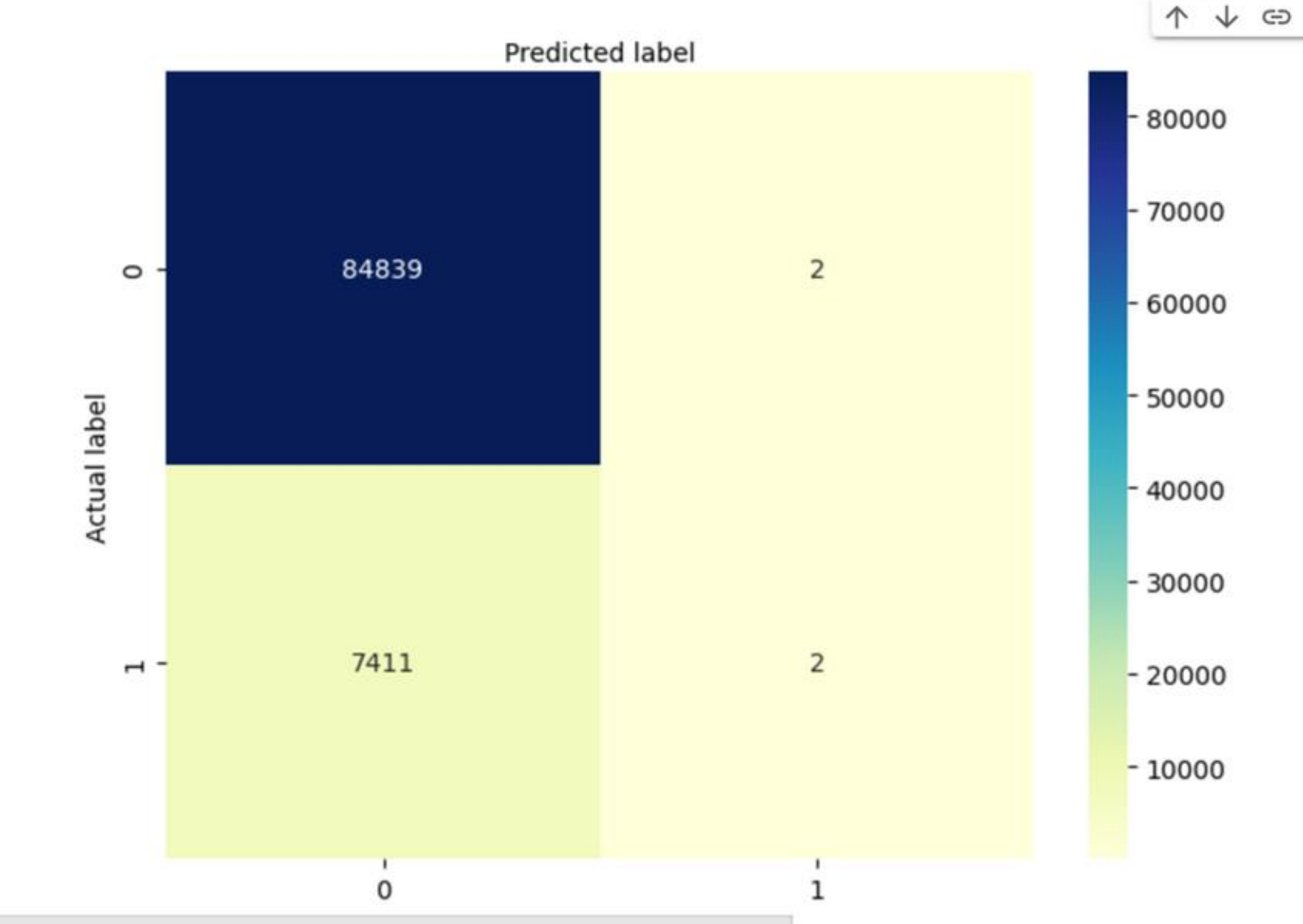



MODELLING

- Memilah data dependen dan independen
- Membagi data training dan testing dengan ukuran data testing sebesar 30%
- Melakukan normalisasi data untuk memudahkan tahapan analisis
- Melakukan analisis data menggunakan lima model diantaranya Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbor, dan Random Forest.



LOGISTIC REGRESSION





Recall: 0.0002697963037906381
Accuracy: 0.9196457606174258

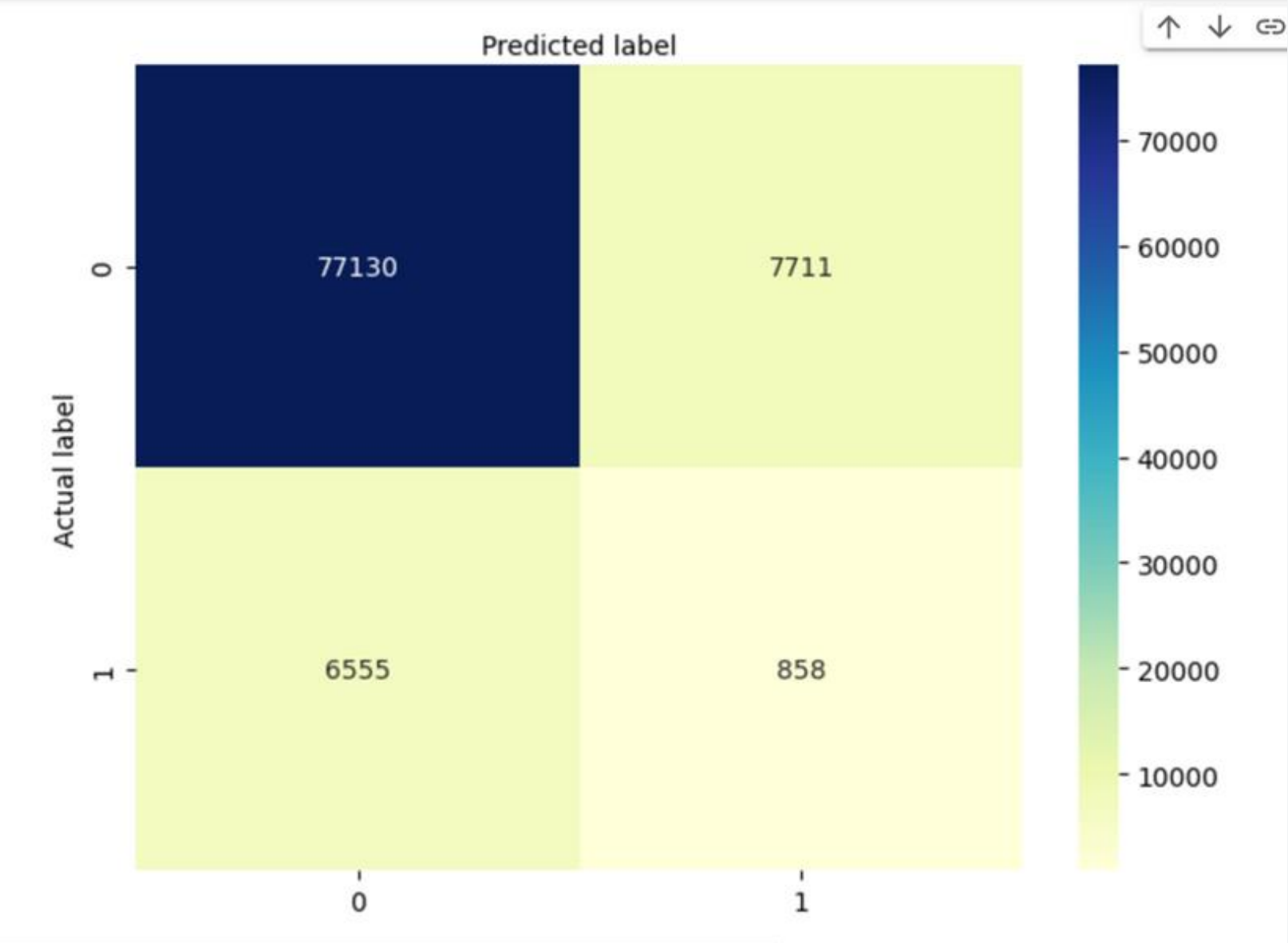
=====
Classification Report:

	precision	recall	f1-score	support
0	0.92	1.00	0.96	84841
1	0.50	0.00	0.00	7413
accuracy			0.92	92254
macro avg	0.71	0.50	0.48	92254
weighted avg	0.89	0.92	0.88	92254

=====
ROC AUC Score: 0.5001231113978495



DECISSION TREE



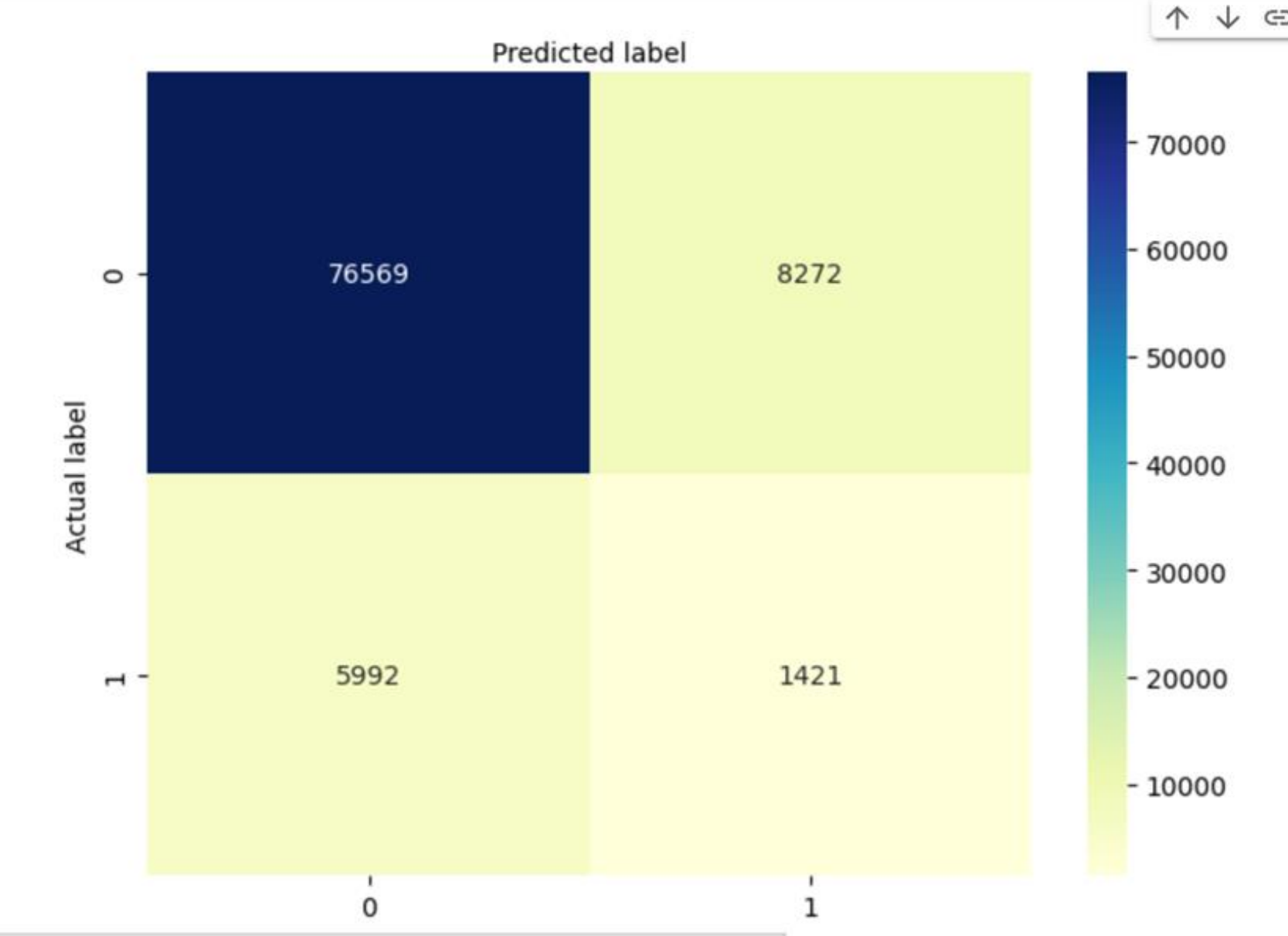
Recall: 0.11574261432618373
Accuracy: 0.8453617187330631

=====
Classification Report:

	precision	recall	f1-score	support
0	0.92	0.91	0.92	84841
1	0.10	0.12	0.11	7413
accuracy			0.85	92254
macro avg	0.51	0.51	0.51	92254
weighted avg	0.86	0.85	0.85	92254

=====
ROC AUC Score: 0.5124274769394971

NAIVE BAYES CLASSIFIER



Recall: 0.19169027384324835
Accuracy: 0.8453833980098424

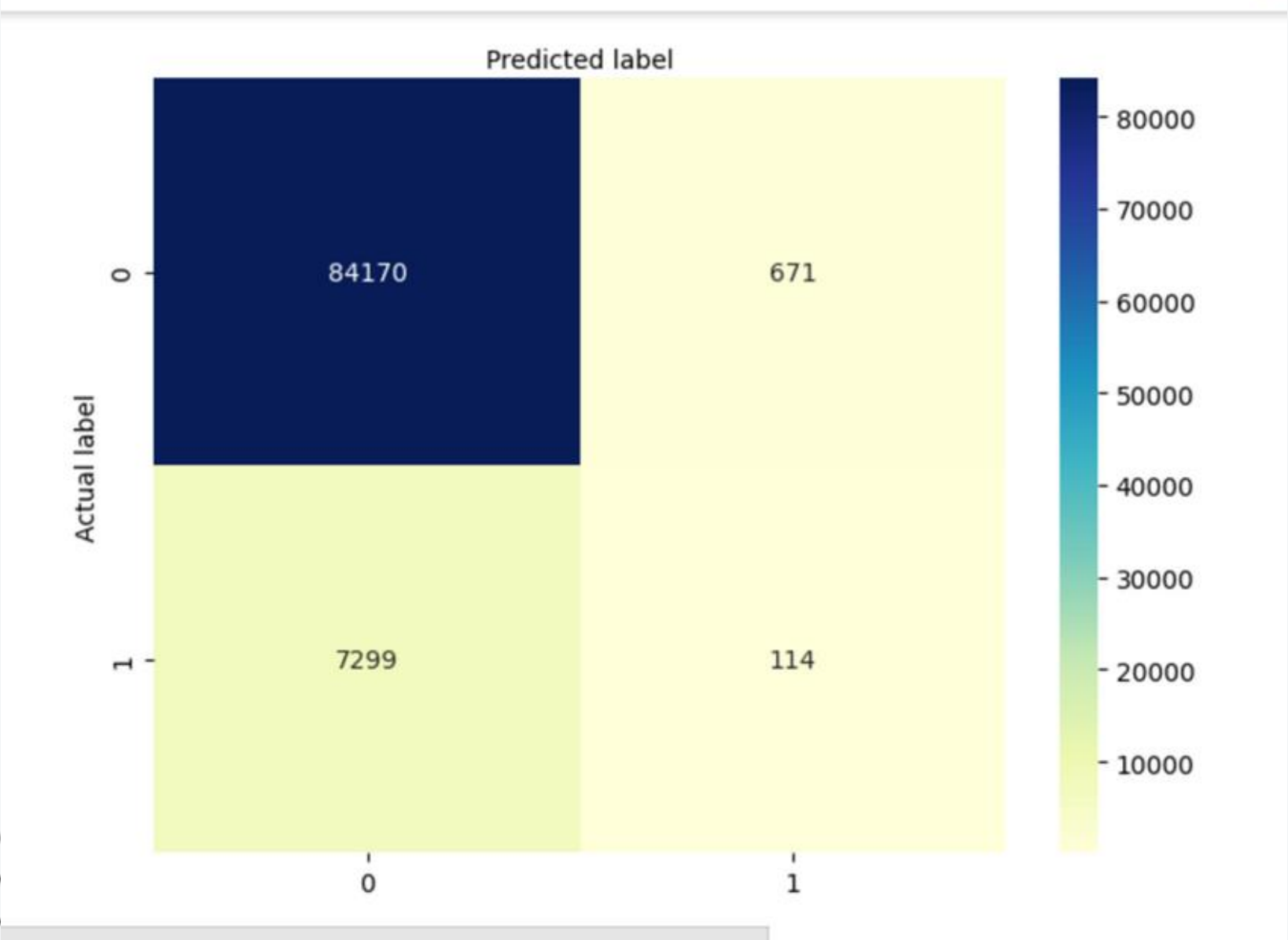
=====
Classification Report:
=====

	precision	recall	f1-score	support
0	0.93	0.90	0.91	84841
1	0.15	0.19	0.17	7413
accuracy			0.85	92254
macro avg	0.54	0.55	0.54	92254
weighted avg	0.86	0.85	0.85	92254

=====
ROC AUC Score: 0.5470951221881817



K-NEAREST NEIGHBOR



Recall: 0.01537838931606637
Accuracy: 0.9136080820343834

=====

Classification Report:

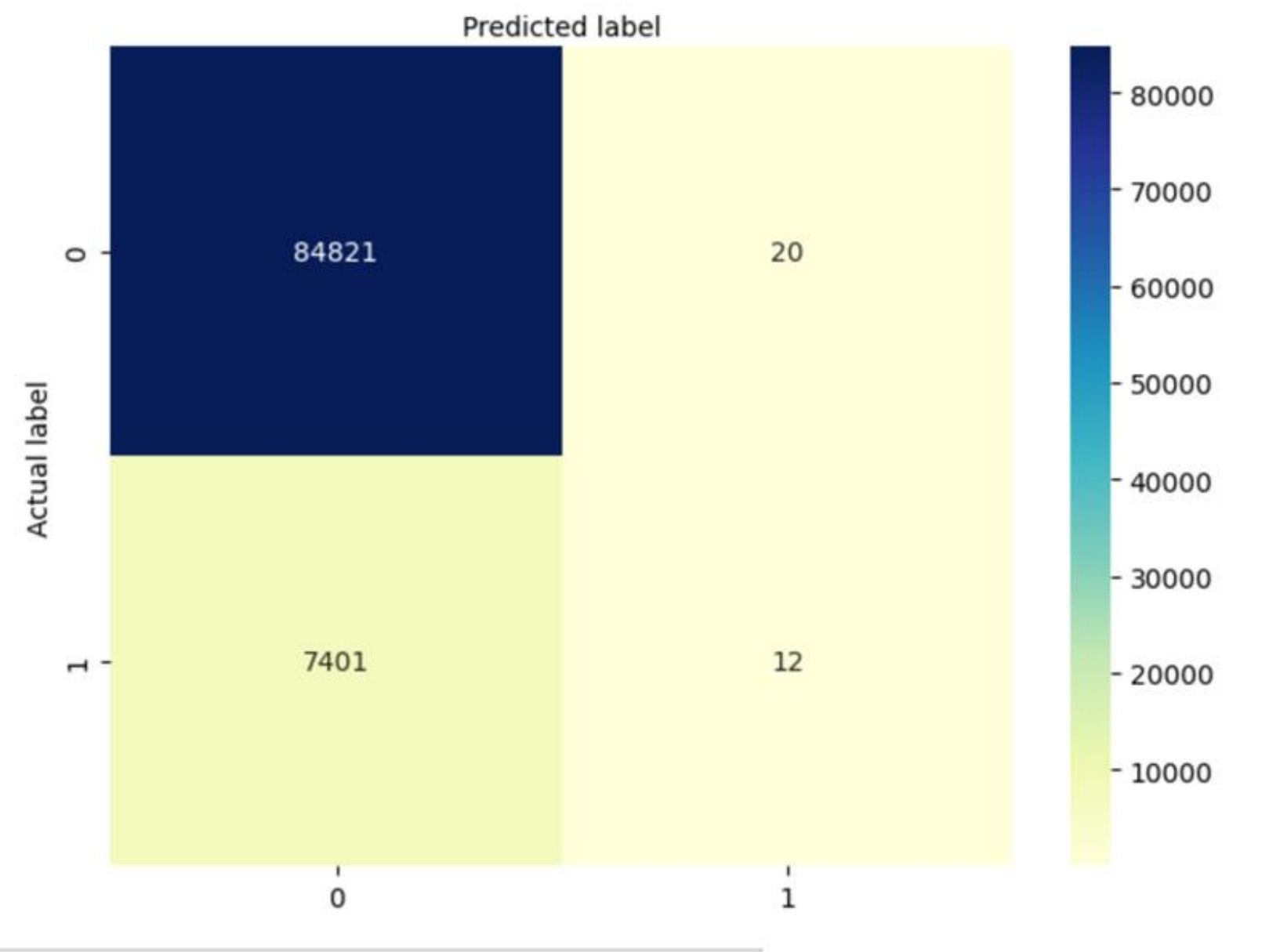
	precision	recall	f1-score	support
0	0.92	0.99	0.95	84841
1	0.15	0.02	0.03	7413
accuracy			0.91	92254
macro avg	0.53	0.50	0.49	92254
weighted avg	0.86	0.91	0.88	92254

=====

ROC AUC Score: 0.5037347386756662



RANDOM FOREST

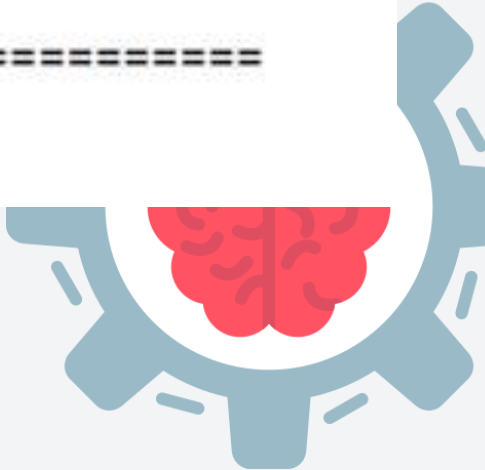


Recall: 0.0016187778227438284
Accuracy: 0.9195590435103085

=====
Classification Report:

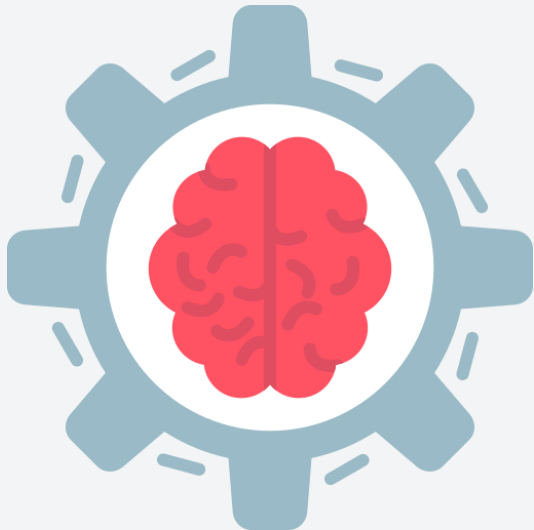
	precision	recall	f1-score	support
0	0.92	1.00	0.96	84841
1	0.38	0.00	0.00	7413
accuracy			0.92	92254
macro avg	0.65	0.50	0.48	92254
weighted avg	0.88	0.92	0.88	92254

=====
ROC AUC Score: 0.5006915213709139



MODEL RECAP

Model	Accuracy	Recall	ROC AUC
Regression Logistic	0,9197	0,0003	0,5001
Decission Tree	0,8454	0,1157	0,5124
Naive Bayes Classification	0,8454	0,1917	0,5471
K-Nearest Neighbor	0,9136	0,0154	0,5037
Random Forest	0,9196	0,016	0,5007



PREDICTION

- Membuat prediksi data target menggunakan model terbaik yaitu Naive Bayes Classifier

```
▶ predictions = NBC.predict(df.drop('TARGET', axis=1))
  predictions = pd.DataFrame({'Prediction': predictions})
  predictions.head()

  result = pd.concat([predictions, df], axis=1)
  result.rename(columns = {'TARGET': 'Actual', 'Prediction': 'Predicted'}, inplace=True)
  result.head()
```

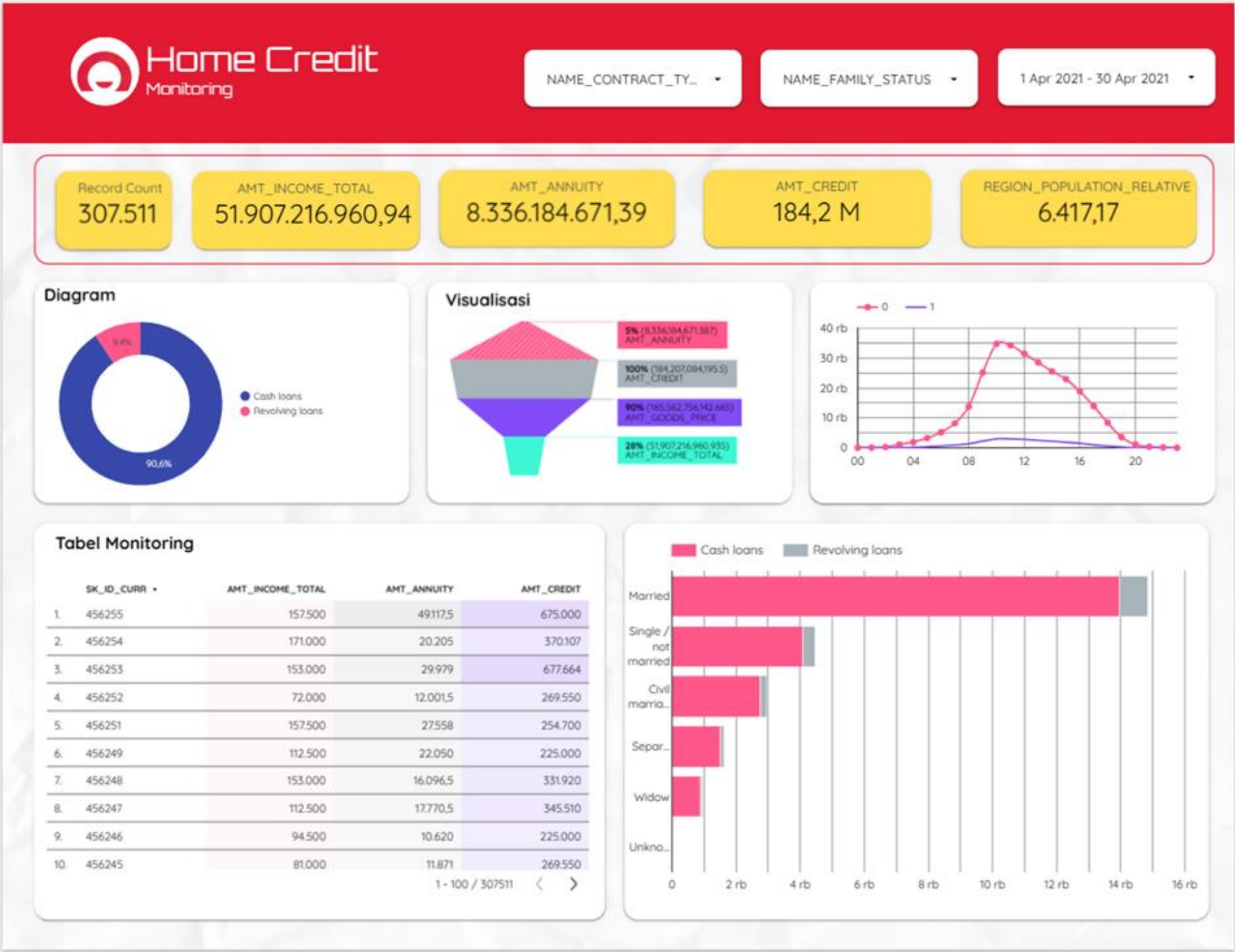
↗

	Predicted	SK_ID_CURR	Actual	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_P
0	0	100002	1	0	202500.0	406597.5	24700.5	351000.0	
1	0	100003	0	0	270000.0	1293502.5	35698.5	1129500.0	
2	0	100004	0	0	67500.0	135000.0	6750.0	135000.0	
3	0	100006	0	0	135000.0	312682.5	29686.5	297000.0	
4	0	100007	0	0	121500.0	513000.0	21865.5	513000.0	

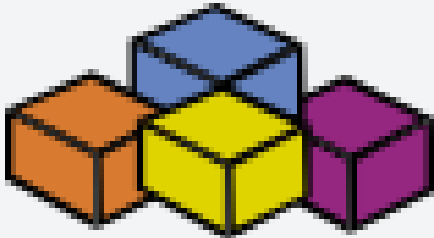


DATA VISUALIZATION





Data Terakhir Diperbarui: 16/6/2023 20:32:30 | Kebijakan Privasi



THANK YOU!

