



# FINAL PROJECT DATA SCIENCE: STUDY CASE DARI DATASET CREDIT CARD CHURN

MAVALDI RIZQY HAZDI

DATA SCIENCE

# LATAR BELAKANG MASALAH

Tim bisnis di sebuah bank semakin resah dengan semakin banyaknya nasabah yang menutup layanan kartu kreditnya. Untuk mencegah hal ini terjadi, tim berencana untuk memberikan *treatment*/layanan yang khusus kepada customers yang berencana akan *churn* (menutup layanan kartu kreditnya), sehingga customers tersebut berubah pikiran dan mengurungkan niat untuk *churn*.

Agar rencana *treatment* yang dilakukan tepat sasaran, tim harus membuat sebuah *predictive* model untuk memprediksi siapa saja *customers* yang akan *churn*.

Sebagai Data Scientist pada sebuah perusahaan yang bergerak di bidang keuangan, Anda diberikan dataset customer yang menutup (*churn*) layanan kartu kreditnya dan yang tidak beserta variabel observasinya (**Credit Card Churn - Dataset.csv**).


# TUJUAN PROYEK

Dari dataset yang tersedia, tujuan dari proyek ini adalah untuk mencari pola customer yang *churn* dan yang tidak supaya Perusahaan dapat mencegah bertambahnya jumlah customer yang churn.

Dengan menganalisa semua faktor-faktor dan beberapa hal yang berpengaruh berat pada pelanggan yang melakukan *churn*, kita dapat menentukan dan memprediksi apa saja permasalahan dan kendala yang dialami oleh pelanggan yang melakukan *churn* dan kita dapat mengatasi hal tersebut agar Perusahaan selalu memiliki pelanggan yang akan setia pada layanan kartu kredit yang kita ciptakan.

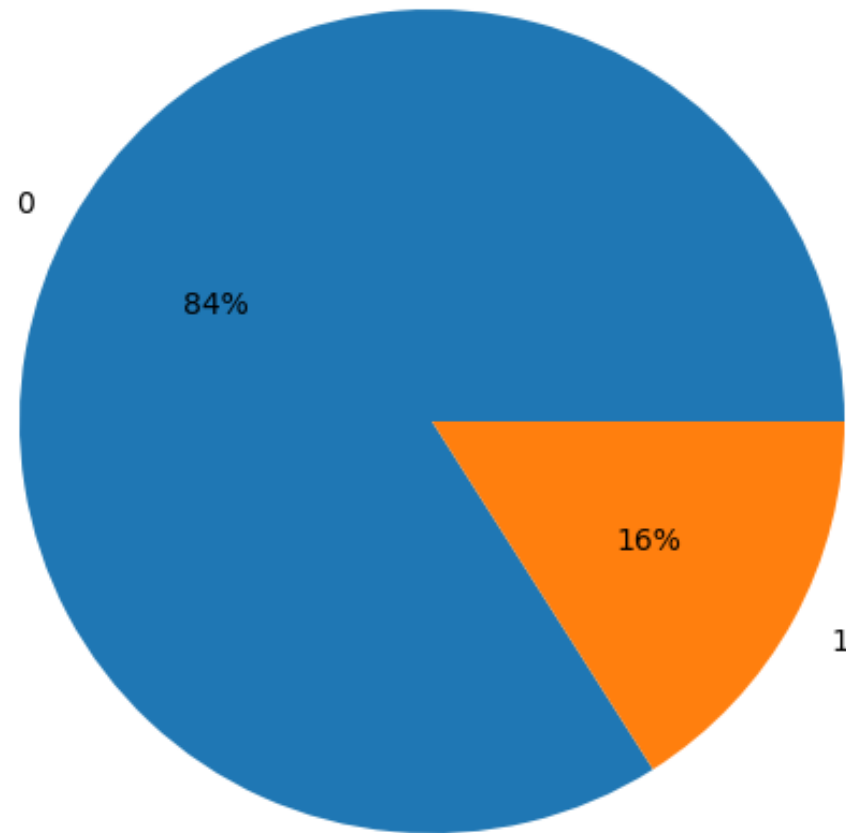
# PENJELASAN FITUR-FITUR PADA DATASET

- `client_id` : ID dari masing-masing customer
- `label` : status churn dari customer (1 : churn ; 0 : no churn)
- `usia` : usia dari customer (tahun)
- `gender` : jenis kelamin (M = Male, F = Female)
- `jumlah_tanggungan` : jumlah tanggungan customer
- `pendidikan` : tingkat pendidikan customer (high school, college graduate, etc.)
- `status_nikah` : status nikah (married, single, divorced, unknown)
- `penghasilan_tahunan` : kategori penghasilan tahunan customer (dalam \$)
- `tipe_kartu_kredit` : tipe kartu kredit yang dipegang customer (Blue, Silver, Gold, Platinum)
- `lama_nasabah` : lama customer sudah menjadi nasabah (bulan)
- `jumlah_produk` : jumlah produk yang dimiliki customer
- `bulan_nonactive` : lama customers tidak aktif dalam 12 bulan terakhir (dalam bulan)
- `jumlah_kontak` : jumlah kontak dalam 12 bulan terakhir
- `total_limit_kredit` : total limit kredit yang diperoleh oleh customer
- `total_limit_kredit_dipakai` : total limit kredit yang sudah dipakai oleh customer
- `sisa_limit_kredit` : sisa limit kredit yang dimiliki customer
- `rasio_transaksi_Q4_Q1` : rasio total nominal transaksi di Q4 terhadap Q1
- `total_transaksi` : total nominal transaksi dalam 12 bulan terakhir
- `jumlah_transaksi` : jumlah transaksi dalam 12 bulan terakhir
- `rasio_jumlah_transaksi_Q4_Q1` : rasio jumlah transaksi di Q4 terhadap Q1
- `rasio_pemakaian` : rasio pemakaian limit kredit

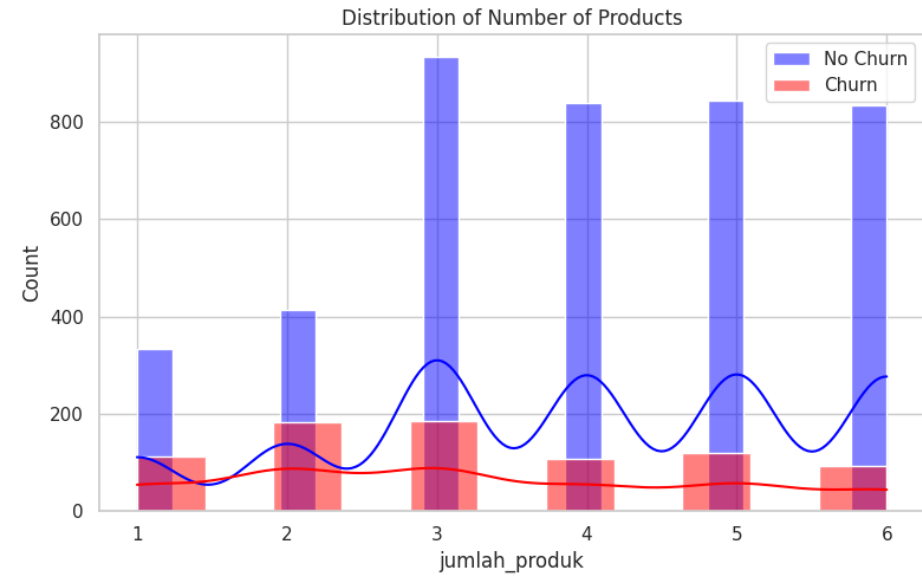
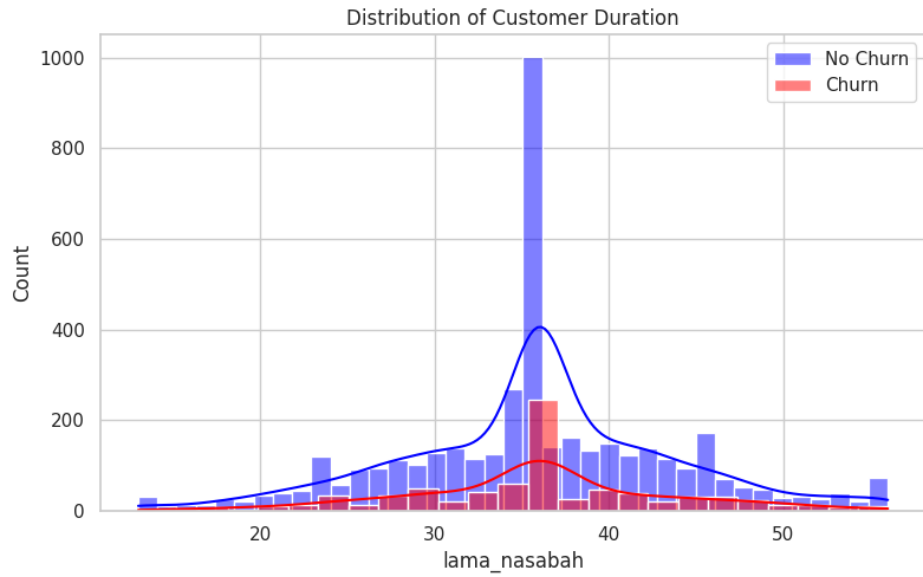
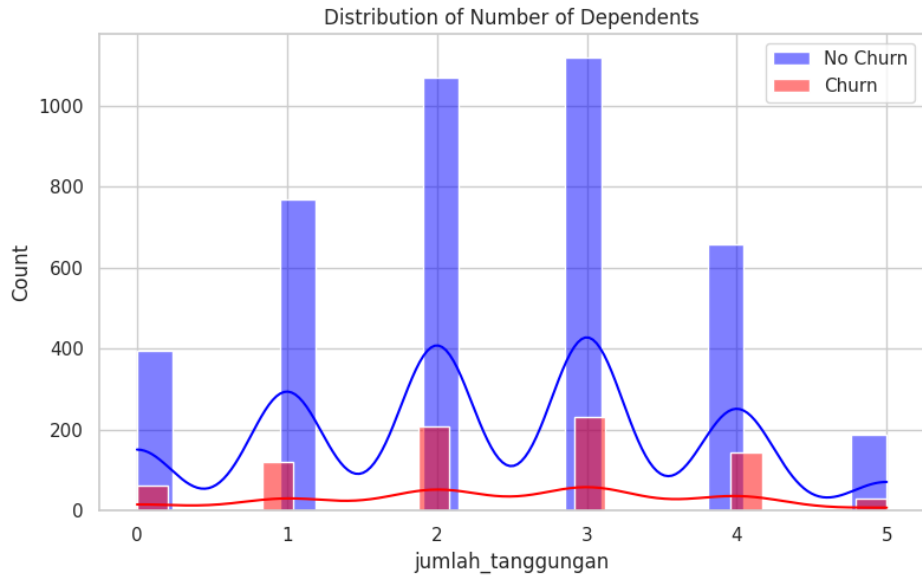
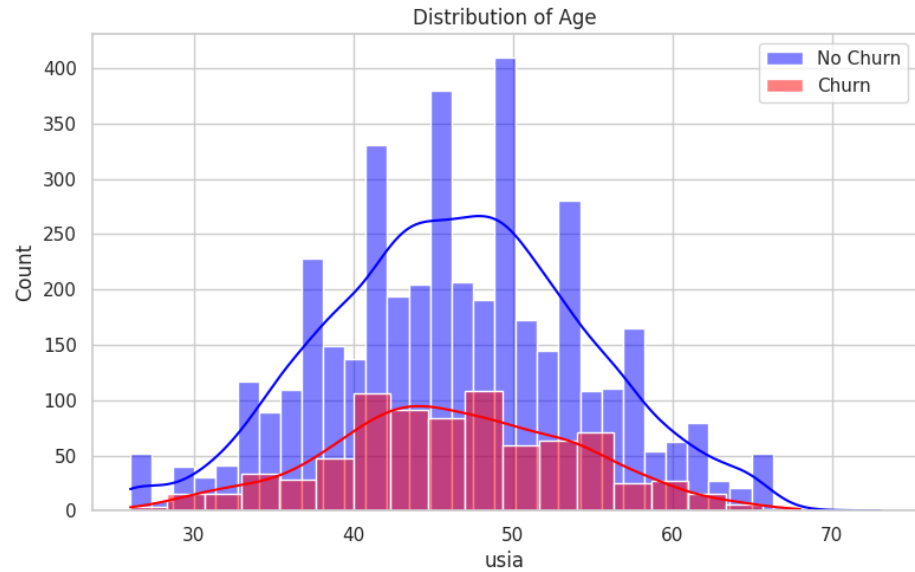
A series of white, thin, overlapping geometric lines and polygons on a black background, located on the left side of the slide. The lines form various shapes, including triangles and quadrilaterals, some of which are nested or intersecting.

# EXPLORATORY DATA ANALYSIS (EDA)

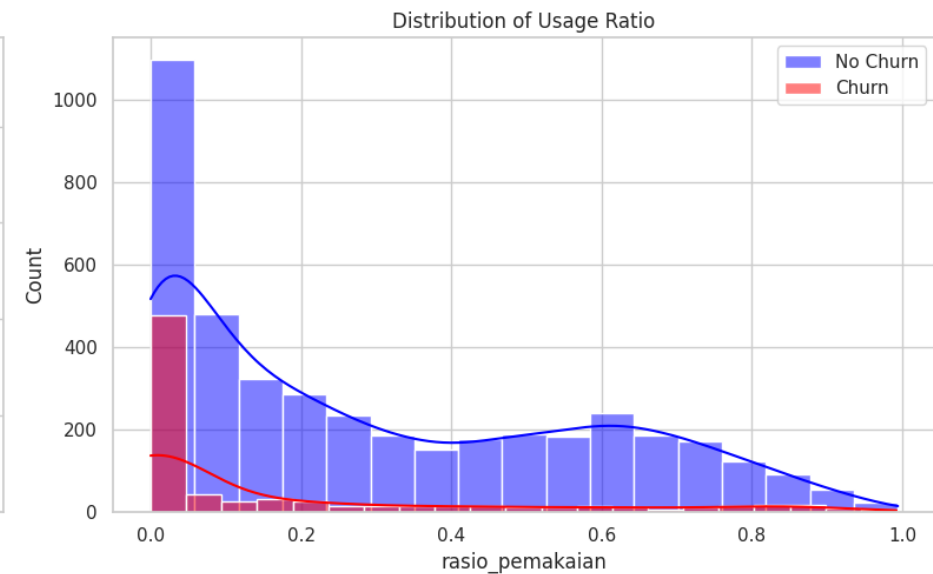
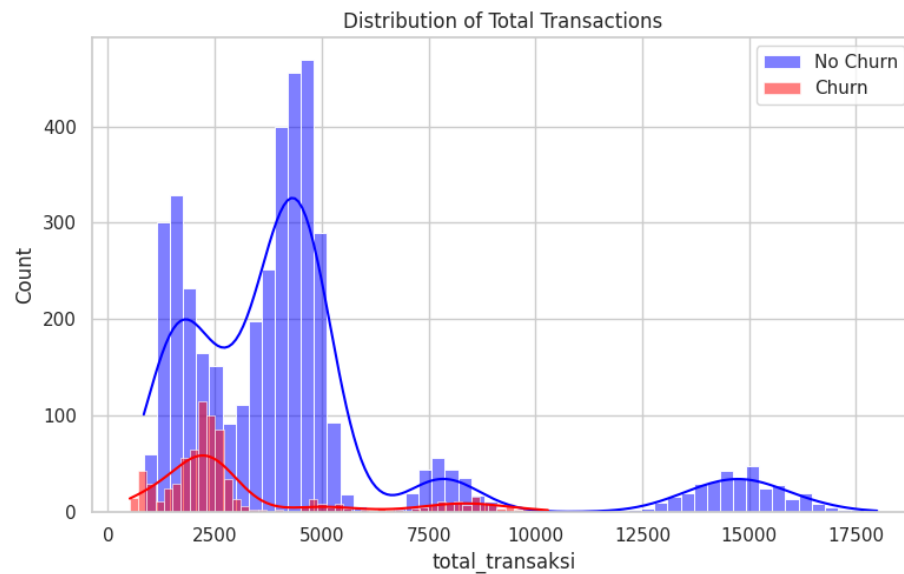
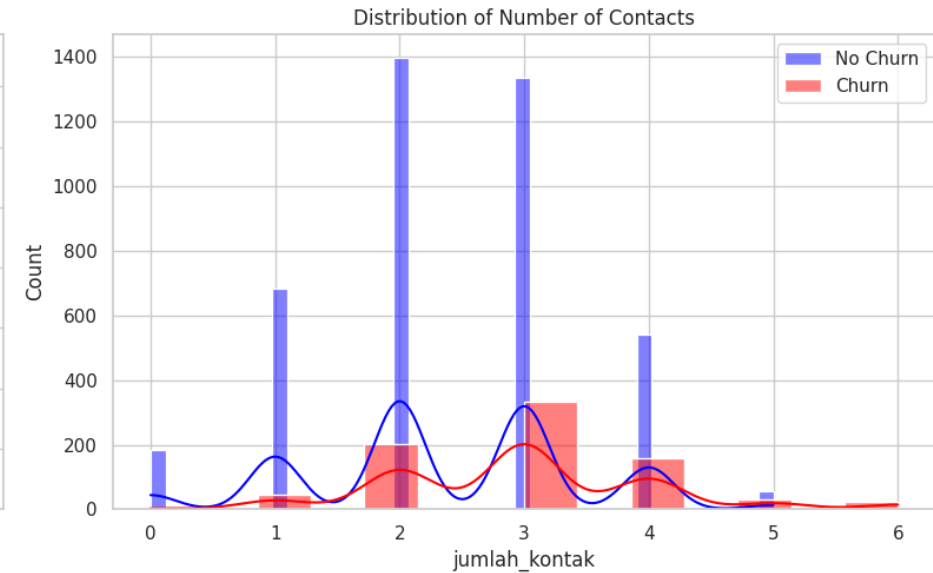
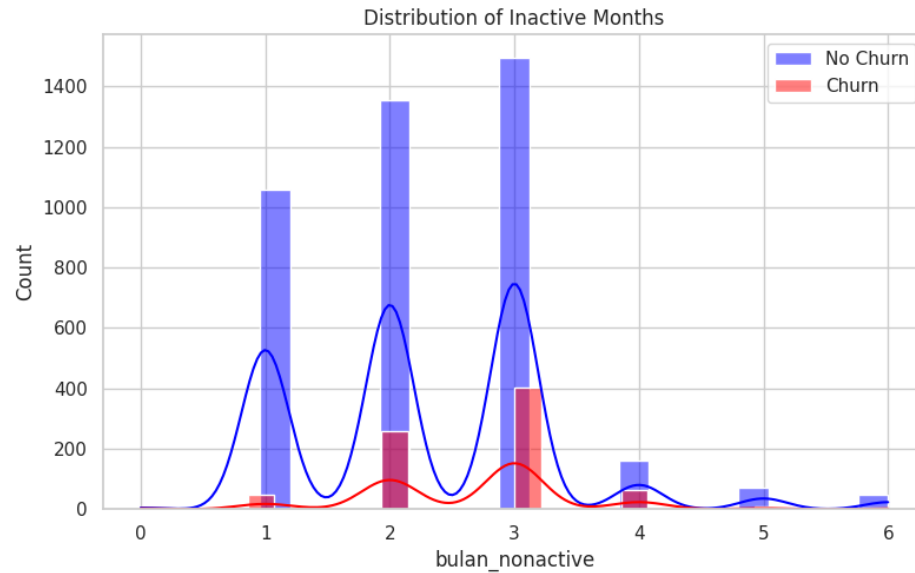
## PIE CHART DATA CHURN DAN TIDAK CHURN



# BAR GRAPH DARI BEBERAPA FITUR PADA DATASET BERDASARKAN LABEL CHURN DAN NO CHURN



# BAR GRAPH DARI BEBERAPA FITUR PADA DATASET BERDASARKAN LABEL CHURN DAN NO CHURN





## KELOMPOK TIDAK CHURN (LABEL 0)

- **Jumlah Sampel:** 4200 klien
- **Usia:** Usia rata-rata sekitar 46,41 tahun.
- **Jumlah Tanggungan:** Rata-rata memiliki 2,34 tanggungan.
- **Durasi Menjadi Nasabah:** Rata-rata durasi sekitar 35,90 bulan.
- **Jumlah Produk:** Rata-rata memiliki 3,94 produk.
- **Bulan Tidak Aktif:** Rata-rata 2,27 bulan tidak aktif.
- **Jumlah Kontak:** Rata-rata memiliki 2,37 kontak.
- **Total Limit Kredit:** Rata-rata limit sebesar 8718,07.
- **Limit Kredit yang Digunakan:** Rata-rata kredit yang digunakan sebesar 1265,80.
- **Sisa Limit Kredit:** Rata-rata sisa kredit sebesar 7452,28.
- **Rasio Transaksi Q4/Q1:** Rasio rata-rata sebesar 0,77.
- **Total Transaksi:** Total transaksi rata-rata sebesar 4644,64.
- **Jumlah Transaksi:** Rata-rata memiliki 68,13 transaksi.
- **Rasio Jumlah Transaksi Q4/Q1:** Rasio rata-rata sebesar 0,73.
- **Rasio Penggunaan:** Rasio penggunaan rata-rata sebesar 0,30.

# KELOMPOK CHURN (LABEL 1)

- **Jumlah Sampel:** 800 klien
- **Usia:** Usia rata-rata sekitar 47,28 tahun.
- **Jumlah Tanggungan:** Rata-rata memiliki 2,36 tanggungan.
- **Durasi Menjadi Nasabah:** Rata-rata durasi sekitar 37,55 bulan.
- **Jumlah Produk:** Rata-rata memiliki 3,44 produk.
- **Bulan Tidak Aktif:** Rata-rata 3,10 bulan tidak aktif.
- **Jumlah Kontak:** Rata-rata memiliki 2,72 kontak.
- **Total Limit Kredit:** Rata-rata limit sebesar 8725,21.
- **Limit Kredit yang Digunakan:** Rata-rata kredit yang digunakan sebesar 678,47.
- **Sisa Limit Kredit:** Rata-rata sisa kredit sebesar 7535,45.
- **Rasio Transaksi Q4/Q1:** Rasio rata-rata sebesar 0,69.
- **Total Transaksi:** Total transaksi rata-rata sebesar 3166,55.
- **Jumlah Transaksi:** Rata-rata memiliki 45,55 transaksi.
- **Rasio Jumlah Transaksi Q4/Q1:** Rasio rata-rata sebesar 0,56.
- **Rasio Penggunaan:** Rasio penggunaan rata-rata sebesar 0,17.

# KESIMPULAN SEMENTARA

- **Durasi:** Klien yang churn memiliki durasi menjadi nasabah yang sedikit lebih lama (37,55 bulan) dibandingkan dengan yang tidak churn (35,90 bulan).
- **Jumlah Produk:** Klien yang churn memiliki lebih sedikit produk rata-rata (3,44) dibandingkan dengan yang tidak churn (3,94).
- **Bulan Tidak Aktif:** Klien yang churn memiliki lebih banyak bulan tidak aktif (3,10) dibandingkan dengan yang tidak churn (2,27).
- **Jumlah Kontak:** Klien yang churn memiliki lebih banyak kontak rata-rata (2,72) dibandingkan dengan yang tidak churn (2,37).
- **Limit Kredit yang Digunakan:** Klien yang churn menggunakan lebih sedikit dari limit kredit mereka (678,47) dibandingkan dengan yang tidak churn (1265,80).
- **Transaksi:** Klien yang churn memiliki total transaksi yang lebih sedikit (3166,55) dan jumlah transaksi yang lebih sedikit (45,55) dibandingkan dengan yang tidak churn (4644,64 dan 68,13).
- **Rasio Penggunaan:** Klien yang churn memiliki rasio penggunaan yang lebih rendah (0,17) dibandingkan dengan yang tidak churn (0,30).

Perbedaan ini menunjukkan bahwa klien yang churn cenderung menggunakan lebih sedikit produk, memiliki lebih banyak bulan tidak aktif, dan memiliki keterlibatan yang lebih rendah dengan penggunaan kredit dan transaksi dibandingkan dengan mereka yang tidak churn. Informasi ini dapat berguna untuk menciptakan strategi untuk mengurangi churn dengan fokus pada peningkatan keterlibatan pelanggan dan penggunaan produk.

A series of white, thin, overlapping geometric lines and polygons on a black background, located on the left side of the slide. The lines form various shapes, including triangles and quadrilaterals, some of which are nested or intersecting.

# DATA PREPROCESSING

# IMPORT DATA MENGGUNAKAN LIBRARY PANDAS

```
# Load Dataset  
data = pd.read_csv('churn.csv')
```

# MELAKUKAN CEK PADA DATA APAKAH DATA YANG DI IMPORT SUDAH BENAR

data.head()

	client_id	label	usia	gender	jumlah_tanggungan	pendidikan	status_nikah	penghasilan_tahunan	tipe_kartu_kredit	lama_nasabah	...	bulan_nonactive	jumlah_kontak	total_limit_kr
0	719455083	0	48	F	3	Uneducated	Single	Less than \$40K	Blue	39	...	3	4	29
1	773503308	0	59	M	1	Uneducated	Single	Less than \$40K	Blue	53	...	5	4	21
2	715452408	0	37	F	2	Graduate	Divorced	Less than \$40K	Blue	36	...	3	3	17
3	711264033	0	47	M	3	Doctorate	Divorced	\$40K - \$60K	Blue	36	...	2	3	47
4	718943508	0	42	M	3	Unknown	Single	\$80K - \$120K	Blue	33	...	3	2	37

5 rows × 21 columns

# MENGHAPUS FITUR CLIENT\_ID DARI DATASET KARENA TIDAK AKAN BERGUNA DALAM PEMBUATAN MODEL

```
[ ] clean_data = data.drop(['client_id'], axis=1)
clean_data.head()
```



	label	usia	gender	jumlah_tanggungan	pendidikan	status_nikah	penghasilan_tahunan	tipe_kartu_kredit	lama_nasabah	jumlah_produk	bulan_nonactive	jumlah_kontak	total_limit_kre
0	0	48	0	3	5	2	4	0	39	4	3	4	299
1	0	59	1	1	5	2	4	0	53	5	5	4	219
2	0	37	0	2	2	0	4	0	36	4	3	3	173
3	0	47	1	3	1	0	1	0	36	4	2	3	478
4	0	42	1	3	6	2	3	0	33	3	3	2	371




# MENGONVERSI KOLOM-KOLOM KATEGORIKAL MENJADI FORMAT NUMERIK MENGGUNAKAN `LABELENCODER` DARI PUSTAKA `SKLEARN`.

```
▶ label_encoders = {}  
categorical_columns = ['gender', 'pendidikan', 'status_nikah', 'penghasilan_tahunan', 'tipe_kartu_kredit']  
  
for col in categorical_columns:  
    le = LabelEncoder()  
    data[col] = le.fit_transform(data[col])  
    label_encoders[col] = le  
  
# Memisahkan features dan target  
X = data.drop(['client_id', 'label'], axis=1)  
y = data['label']
```

# MEMISAHKAN ATAU SPLIT DATA MENJADI TRAINING SET DAN TESTING SET

```
[ ] # Split data menjadi training dan testing set
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[ ] print(y_train.shape)
    print(y_test.shape)
```



```
(4000,)
(1000,)
```

\*

Dataset untuk training sebanyak 4000

Dataset untuk testing sebanyak 1000

A series of white, thin, overlapping geometric lines on a black background, forming a complex, abstract pattern on the left side of the slide.

# MACHINE LEARNING MODELING

## MODEL RANDOM FOREST CLASSIFIER (TRAINING MODEL)

Accuracy: 0.905

Classification Report Training Model (Random Forest) :

	precision	recall	f1-score	support
0	0.92	0.97	0.94	3349
1	0.77	0.59	0.67	651
accuracy			0.91	4000
macro avg	0.85	0.78	0.81	4000
weighted avg	0.90	0.91	0.90	4000

Confusion Matrix:

```
[[3236  113]
 [ 267  384]]
```

## MODEL RANDOM FOREST CLASSIFIER (TESTING MODEL)

Accuracy: 0.968

Classification Report Testing Model (Random Forest) :

	precision	recall	f1-score	support
0	0.97	1.00	0.98	851
1	0.97	0.81	0.88	149
accuracy			0.97	1000
macro avg	0.97	0.90	0.93	1000
weighted avg	0.97	0.97	0.97	1000

Confusion Matrix:

```
[[847  4]
 [ 28 121]]
```

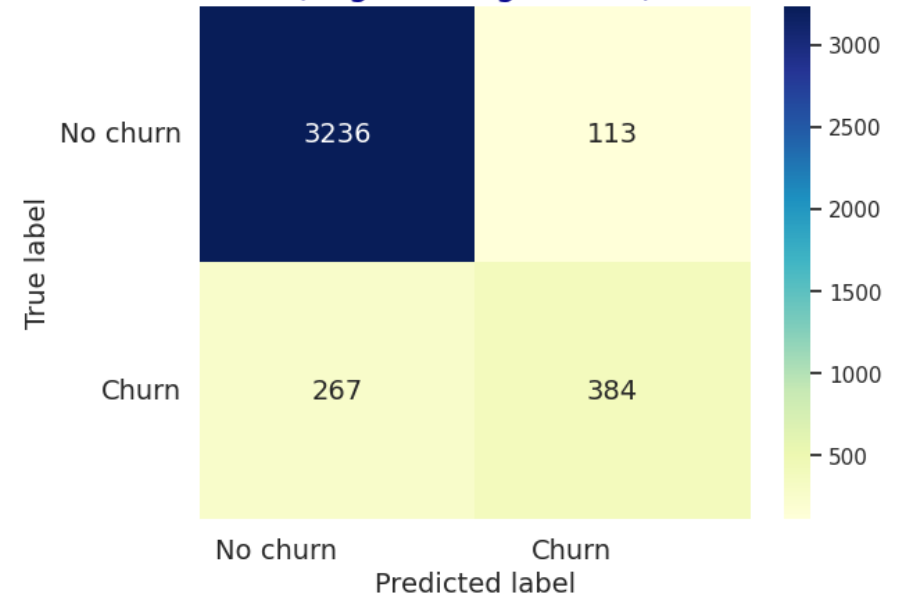
# MODEL LOGISTIC REGRESSION (TRAINING MODEL)

Accuracy: 0.905

Classification Report Training Model (Logistic Regression) :

	precision	recall	f1-score	support
0	0.92	0.97	0.94	3349
1	0.77	0.59	0.67	651
accuracy			0.91	4000
macro avg	0.85	0.78	0.81	4000
weighted avg	0.90	0.91	0.90	4000

Confusion Matrix for Training Model  
(Logistic Regression)



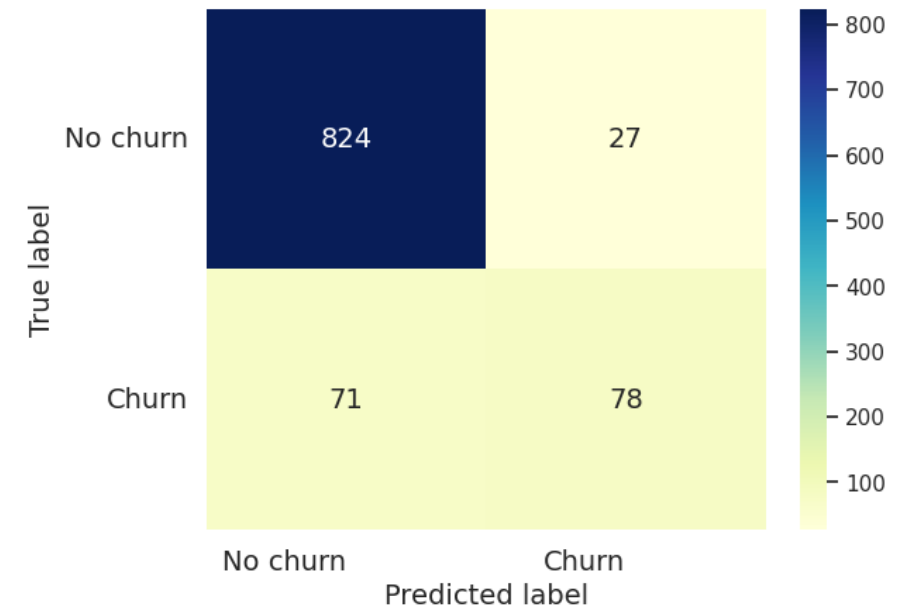
# MODEL LOGISTIC REGRESSION (TESTING MODEL)

Accuracy: 0.968

Classification Report Testing Model (Logistic Regression) :


	precision	recall	f1-score	support
0	0.92	0.97	0.94	851
1	0.74	0.52	0.61	149
accuracy			0.90	1000
macro avg	0.83	0.75	0.78	1000
weighted avg	0.89	0.90	0.89	1000

Confusion Matrix for Testing Model  
(Logistic Regression)



Dilihat dari kedua model yang menghasilkan performa hampir sama baiknya, maka kedua model tersebut dapat dipakai untuk dataset ini. Tetapi jika dilihat pada **Testing Model** dari kedua model dapat dilihat jika testing model pada **Random Forest Classifier** lebih baik daripada Logistic Regression.



A series of white, thin, overlapping geometric lines on a black background, forming a complex, abstract shape on the left side of the slide.

# TERIMA KASIH