



# FINAL PROJECT



# FRAUD DETECTION

Fraud merupakan tindakan yang dilakukan secara sadar dengan tujuan untuk menguntungkan pihak tertentu namun bersifat merugikan pihak yang lainnya karena adanya motif penghilang harta yang dikuasai oleh pihak tertentu dengan cara yang ilegal (Syahputra & Urumsah, 2019).

Tindakan fraud mengakibatkan kerugian yang dapat mempengaruhi perusahaan. Pada pelaksanaannya fraud detection sangat sulit dideteksi dan membutuhkan waktu yang cukup lama maka dari itu dibutuhkan pendekripsi tidakan fraud secara cepat dan akurat dengan menerapkan algoritma machine learning.



# Langkah-langkah Analisis



- 01 Memasukkan library
- 02 Membaca Dataset
- 03 Mengexplore Dataset
- 04 Data Pre-processing
- 05 Analisis Regresi Logistik
- 06 Oversampling Using Smote to Resampling Data
- 07 Recursive Feature Elimination

01

## Memasukkan library

Library Python adalah kumpulan kode pada Python yang dapat digunakan kembali dalam beberapa program atau proyek.

- Pandas

Digunakan untuk melakukan pre-processing dan analisis data.

---

- Numpy

Digunakan dalam membantu proses komputasi numerik

---

- Matplotlib

Digunakan untuk menyajikan data ke dalam bentuk visual.

---

- Seaborn

Digunakan untuk membuat grafik visualisasi data.

---

- Scikit-learn

Digunakan untuk membuat model machine learning.

---

## 02

# Membaca Dataset

- Melakukan import source data kedalam google colabs melalui google drive.

```
from google.colab import drive  
drive.mount('/content/drive')
```

- Membaca file kedalam bentuk dataframe.

```
"pd.read_json"
```



03

## Mengexplore Dataset

Menampilkan Baris Paling Atas

"df.head()"

Menampilkan Baris Paling Bawah

"df.tail()"

Menampilkan Jumlah Baris & Kolom

"df.shape"

Menampilkan Kolom pada Data

"df.columns"

Menampilkan Jumlah Masing-masing Tipe Data

"df.dtypes.value\_counts()"

Menampilkan Informasi Mengenai Data

"df.info()"

Menampilkan Jumlah Nilai Unik pada Data

"df.unique()"

## 04

# Data Pre-processing

## a. Data Manipulation

- Mencari value dari kolom enterCVV dan CVV dengan membuat kolom baru berisi kesesuaian antar kedua kolom tersebut.
- Menghapus data yang duplikat.
- Menghapus kolom yang kosong dan yang tidak relevan.

"echoBuffer", "posOnPremises", "recurringAuthInd"

"merchantCity", "merchantState", "merchantZip", "customerId", "accountNumber", "cardCVV"

"enteredCVV", dan "cardLast4Digits"

- Melihat tipe data dari setiap kolom dan merubah tipe data yang salah kedalam tipe data yang seharusnya

(Object => Datetime) : "transactionDateTime", "currentExpDate", "accountOpenDate",  
"dateOfLastAddressChange".



## b. Data Quality Assessment

- Mengecek missing values

acqQountry	3913
merchantCountryCode	624
transactionType	589
posEntryMode	3345
posConditionCode	287

- Mengecek outlier pada kolom data numerik
- Melihat density dari kolom data numerik

## c. Data Cleaning

Pada tahap ini dilakukan handling missing values pada kolom yang terdapat missing value dengan mengisi nilai yang kosong tersebut dengan modus.

## d. Data Transformation

- Mencari importance feature dengan menemukan korelasi dari variabel dependen terhadap variabel target.

	column_name	isFraud
0	transactionAmount	0.088708
1	cardPresent	0.014946
2	CVV_Valid	0.011195
3	posConditionCode	0.009883
4	currentBalance	0.008611
5	posEntryMode	0.005542
6	creditLimit	0.003357
7	availableMoney	0.001379
8	expirationDateKeyInMatch	0.001239

- Menghapus kolom yang tidak memiliki korelasi atau yang bukan merupakan importance feature terhadap kolom variabel target.

# Dataset Untuk Membuat Model

	accountNumber	creditLimit	availableMoney	transactionAmount	posEntryMode	posConditionCode	isFraud	currentBalance	cardPresent	expirationDateKeyInMatch	CVV_Valid
0	733493772	5000	5000.00	111.33	5	1	1	0.00	0	0	1
1	733493772	5000	4888.67	24.75	9	1	0	111.33	0	0	1
2	733493772	5000	4863.92	187.40	5	1	0	136.08	0	0	1
3	733493772	5000	4676.52	227.34	2	1	1	323.48	0	0	1
4	733493772	5000	4449.18	0.00	2	1	0	550.82	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
641909	186770399	7500	2574.02	5.37	5	8	0	4925.98	0	0	1
641910	186770399	7500	2568.65	223.70	9	1	0	4931.35	0	0	1
641911	186770399	7500	2344.95	138.42	2	1	0	5155.05	0	0	1
641912	186770399	7500	2206.53	16.31	9	8	0	5293.47	0	0	1
641913	186770399	7500	2190.22	32.53	9	1	0	5309.78	0	0	1

641914 rows × 11 columns

# **MODELLING**

## **REGRESI LOGISTIK**

Regresi logistik adalah teknik klasifikasi dalam supervised learning yang digunakan untuk memprediksi berdasarkan hubungan antara variabel dependen dengan variabel target . Di mana, variabel target pada kasus ini bersifat binary(1/0).

### **Model**

membuat model regresi logistik menggunakan dataset yang sudah melalui tahap data preprocesing

### **Model**

membuat model Regresi Logistik dengan melakukan resampling

### **Model**

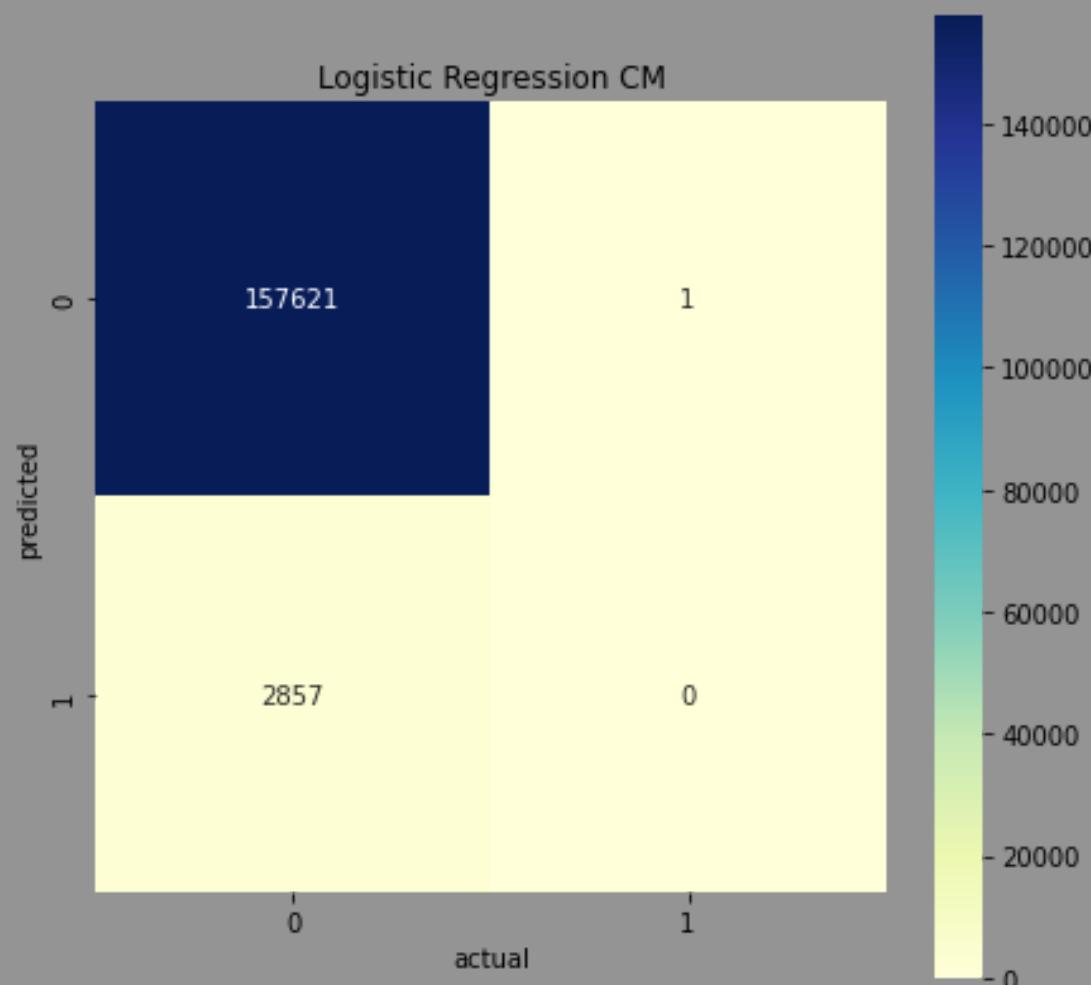
membuat model Regresi Logistik menggunakan Recursive Feature elimination

## Classification Report

	precision	recall	f1-score	support
0	0.98	1.00	0.99	157622
1	0.00	0.00	0.00	2857
accuracy			0.98	160479
macro avg	0.49	0.50	0.50	160479
weighted avg	0.96	0.98	0.97	160479

```
Sensitivity score : 0 %
specificity score : 100 %
```

## Confusion Matrix



# MODEL

## Data:

- Variabel Dependen: IsFraud(0 dan 1)
- Variabel Independen: creditLimit, availableMoney, transactionAmount, posEntryMode, posConditionCode, currentBalance, cardPresent, expirationDateKeyInMatch, CVV\_Valid

## Hasil:

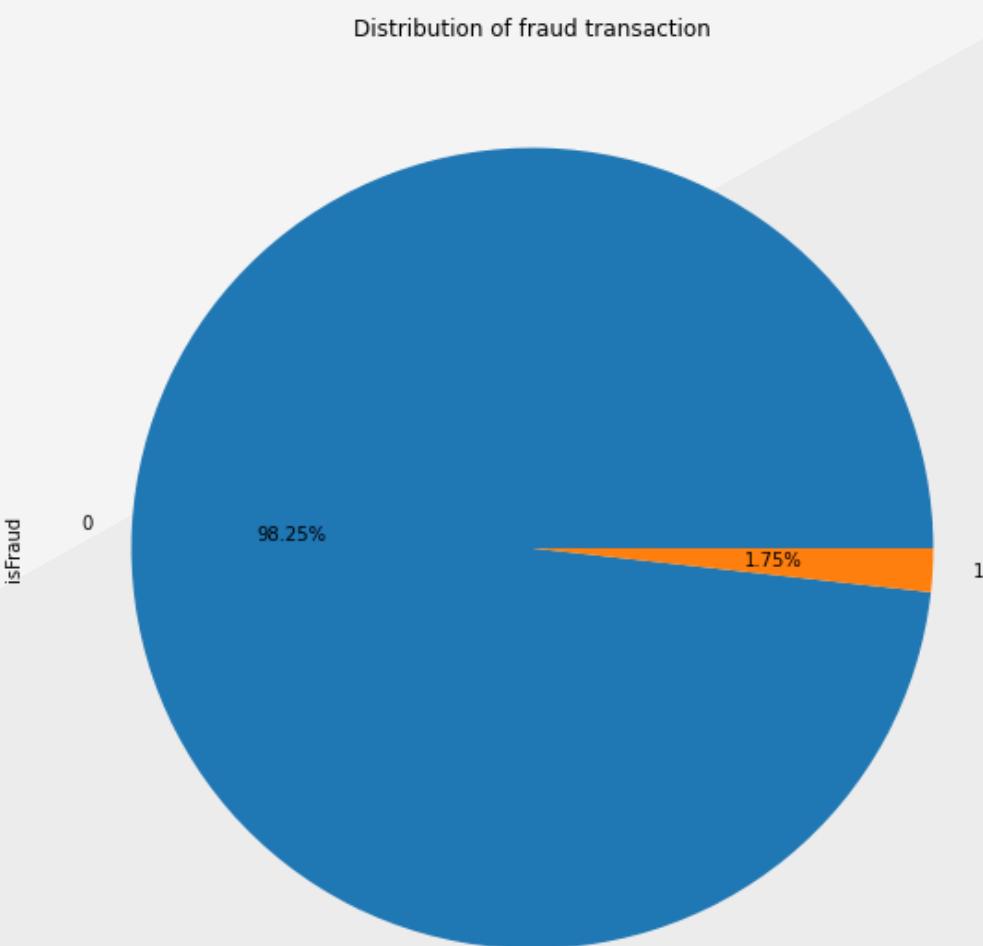
1. Akurasi prediksi yang didapatkan ialah sebesar 98%
2. Sensitivity ( kebenaran memprediksi true positif dibandingkan dengan keseluruhan data yang true) sebesar 0%
3. Specificity (Merupakan kebenaran memprediksi false dibandingkan dengan keseluruhan data yang false) sebesar 100%
4. Precision merupakan rasio prediksi benar dibandingkan dengan keseluruhan hasil yang diprediksi , sebesar 98% untuk prediksi 0 dan 0% untuk prediksi 1

○ ○ ○

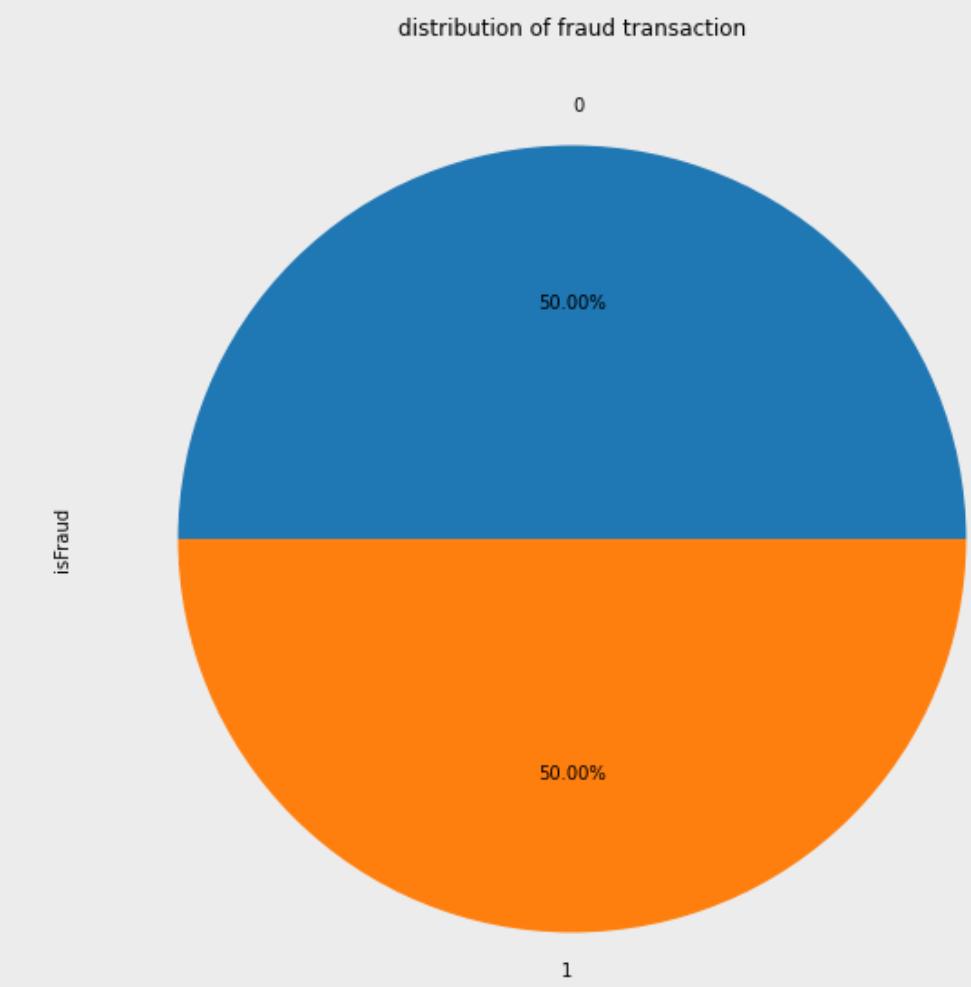
# OVERSAMPLING

Dengan melakukan resampling ini akan mengetahui kinerja model jika pembagian data seimbang teknik yang dipakai adalah oversampling , Secara umum teknik ini adalah mengambil kelas minoritas sedemikian rupa sehingga proporsinya dalam sample lebih besar dibandingkan proporsi asalnya. Yang dilakukan umumnya pada kasus pemodelan klasifikasi adalah dengan cara menduplikasi kelas minoritas

- Data yang fraud dalam kasus ini merupakan kejadian langka dimana hanya terdapat 1.75% data dinyatakan fraud
- Sebagian besar algoritma pembelajaran mesin tidak bekerja dengan baik dengan dataset tidak seimbang.



**Sebelum resampling dataset**



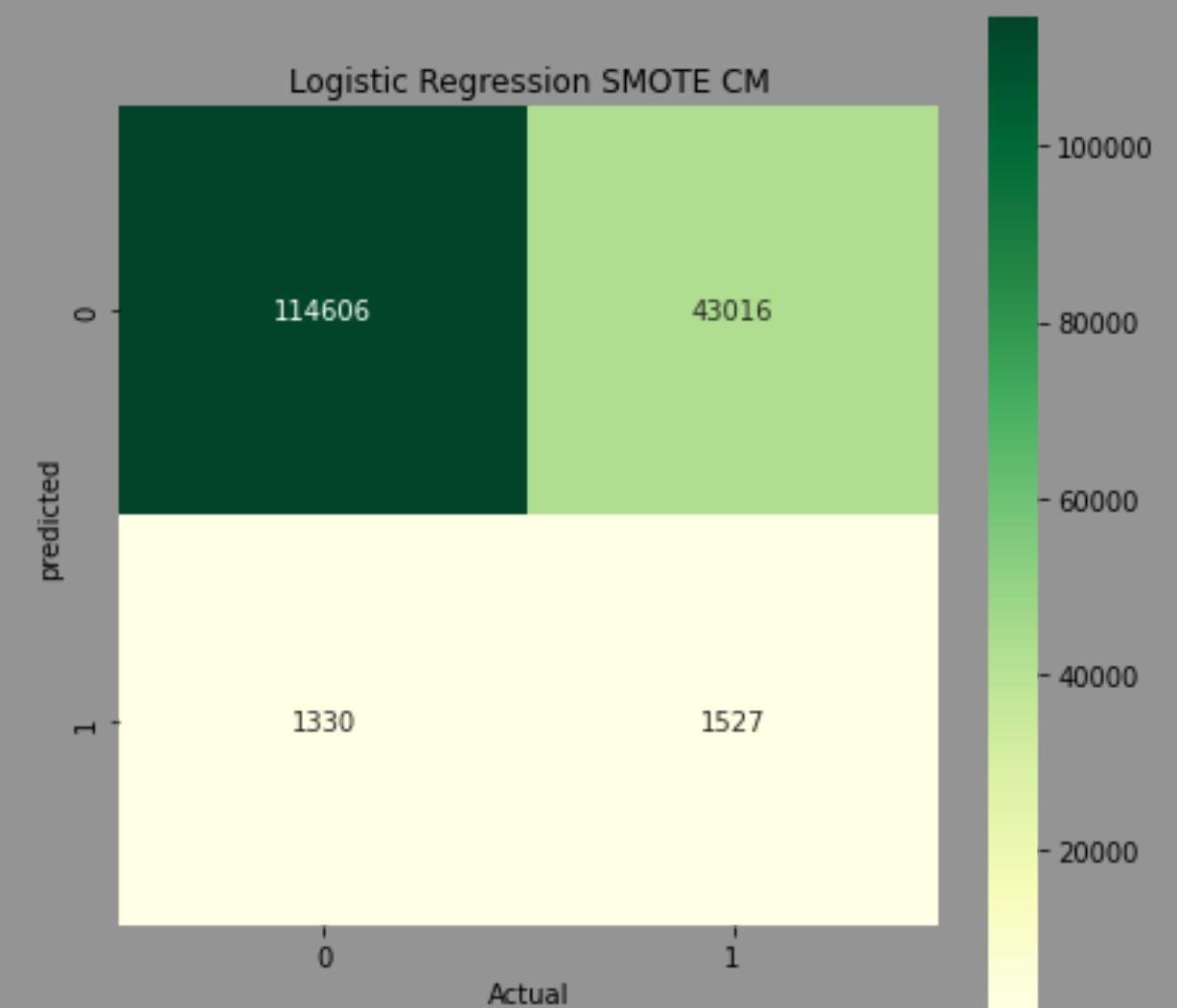
**Setelah resampling dataset**

○ ○ ○

## Classification Report

	precision	recall	f1-score	support
0	0.99	0.73	0.84	157622
1	0.03	0.54	0.06	2857
accuracy			0.72	160479
macro avg	0.51	0.63	0.45	160479
weighted avg	0.97	0.72	0.82	160479
Sensitivity score :	53 %			
specificity score :	73 %			

## Confussion Matrix



O O O O

# MODEL

## Data:

- Menggunakan data setelah resampling
- Variabel Dependen: IsFraud(0 dan 1)
- Variabel Independen: creditLimit, availableMoney, transactionAmount, posEntryMode, posConditionCode, currentBalance, cardPresent, expirationDateKeyInMatch, CVV\_Valid

## Hasil:

1. Akurasi yang didapatkan ialah sebesar 73%
2. Sensitivity ( kebenaran memprediksi true positif dibandingkan dengan keseluruhan data yang true) sebesar 53%
3. Specificity (Merupakan kebenaran memprediksi false dibandingkan dengan keseluruhan data yang false) sebesar 73%
4. Precision merupakan rasio prediksi benar dibandingkan dengan keseluruhan hasil yang diprediksi , sebesar 99% untuk hasil prediksi 0 dan 0,3% untuk prediksi 1

O O O O

# MODEL

## Data:

- Variabel Dependen: IsFraud(0 dan 1)
- Variabel Independen ditentukan dengan menggunakan metode RFE (Recursive Feature Elimination) yaitu transactionAmount, posConditionCode, cardPresent, CVV\_Valid

## Hasil:

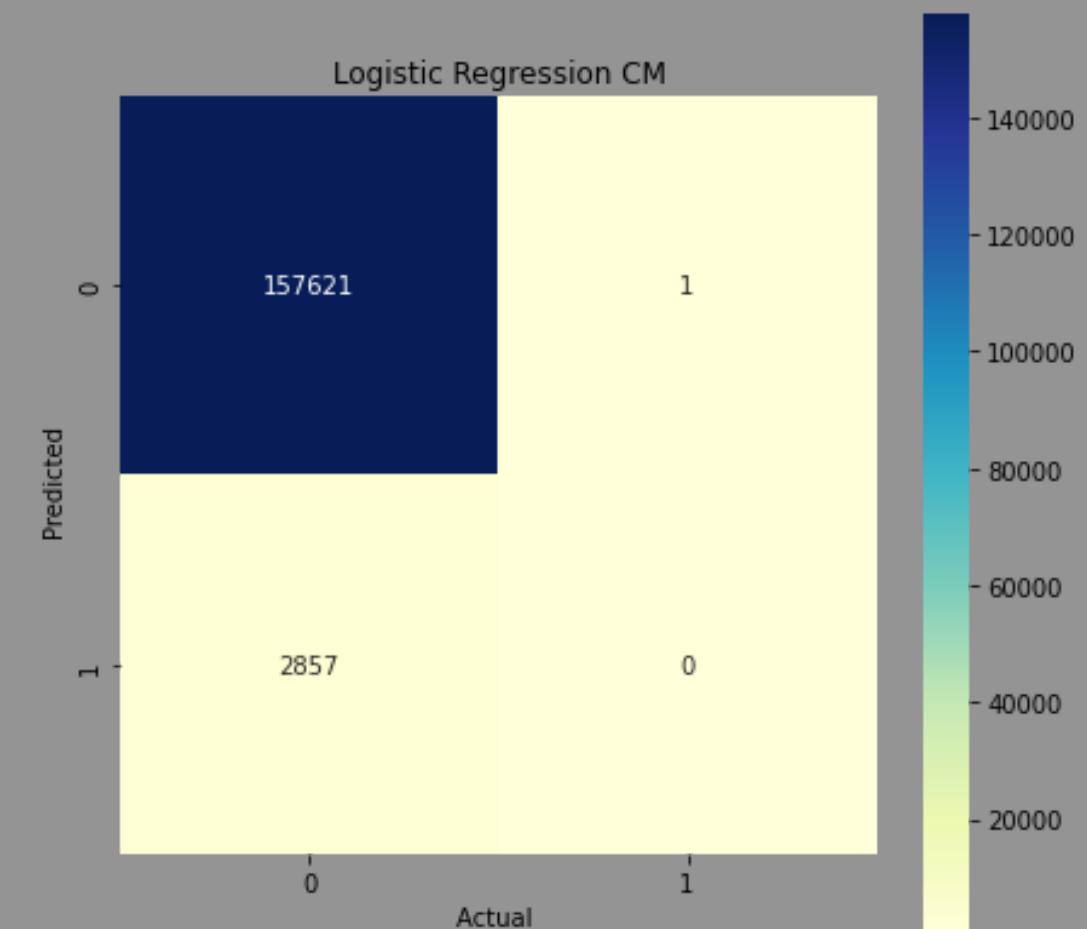
1. Akurasi yang didapatkan ialah sebesar 98%
2. Sensitivity (kebenaran memprediksi true positif dibandingkan dengan keseluruhan data yang true) sebesar 0%
3. Specificity (Merupakan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif) sebesar 100%
4. Precision merupakan rasio prediksi benar dibandingkan dengan keseluruhan hasil yang diprediksi, sebesar 99% untuk hasil prediksi 0 dan 0,3% untuk hasil prediksi 1

## Classification Report

	precision	recall	f1-score	support
0	0.98	1.00	0.99	157622
1	0.00	0.00	0.00	2857
accuracy			0.98	160479
macro avg	0.49	0.50	0.50	160479
weighted avg	0.96	0.98	0.97	160479

Sensitivity score : 0 %  
specificity score : 100 %

## Confusion Matrix



# **MODEL RFE SUMMARY RESULT**

```

Results: Logit
=====
Model:           Logit          Pseudo R-squared: -6.846
Dependent Variable: y            AIC:                 667142.4519
Date:            2022-07-08 03:39 BIC:                 667186.7900
No. Observations: 481435        Log-Likelihood:   -3.3357e+05
Df Model:         3             LL-Null:            -42515.
Df Residuals:    481431        LLR p-value:       1.0000
Converged:        1.0000        Scale:               1.0000
No. Iterations:   3.0000

-----
```

	Coef.	Std.Err.	z	P> z	[0.025	0.975]
x1	0.0466	0.0029	16.0948	0.0000	0.0409	0.0522
x2	0.0058	0.0029	2.0147	0.0439	0.0002	0.0115
x3	-0.0096	0.0029	-3.3239	0.0009	-0.0152	-0.0039
x4	-0.0070	0.0029	-2.4203	0.0155	-0.0126	-0.0013

=====

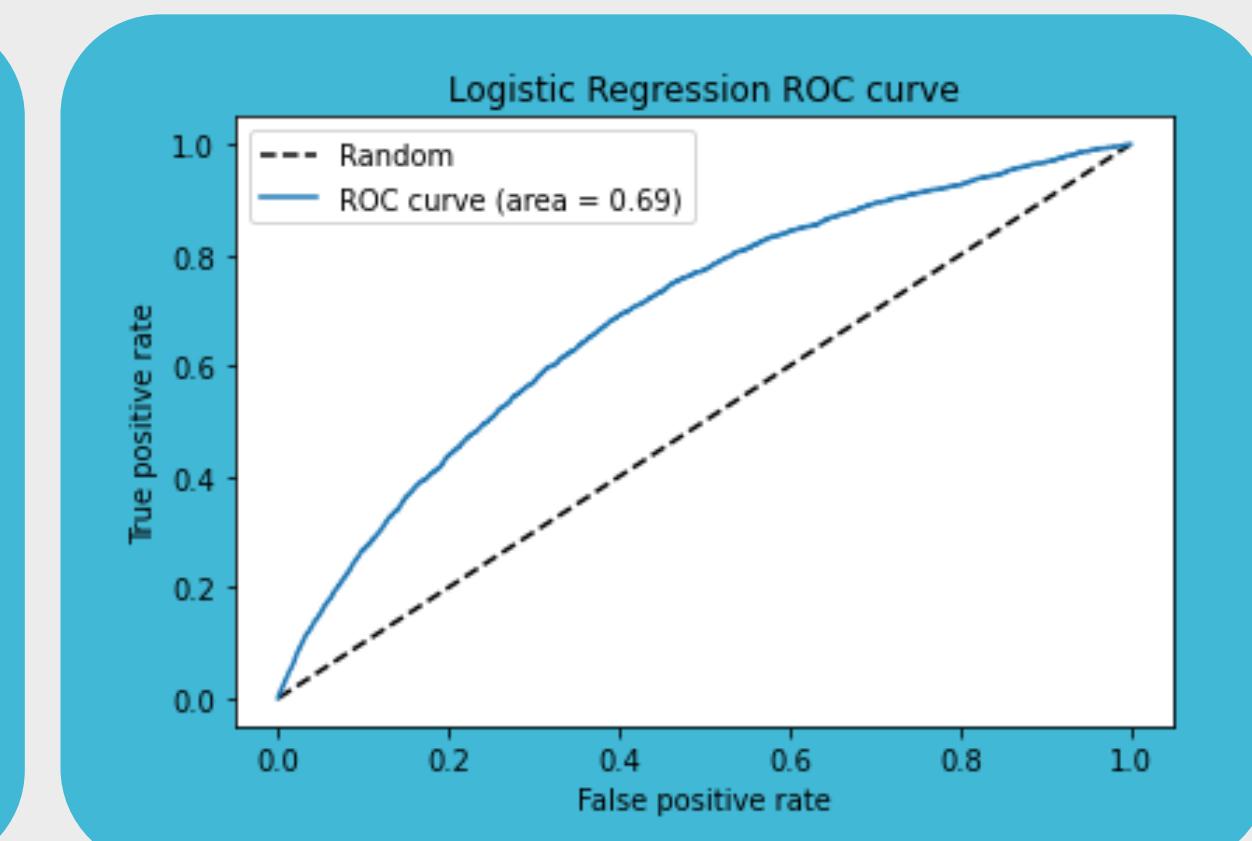
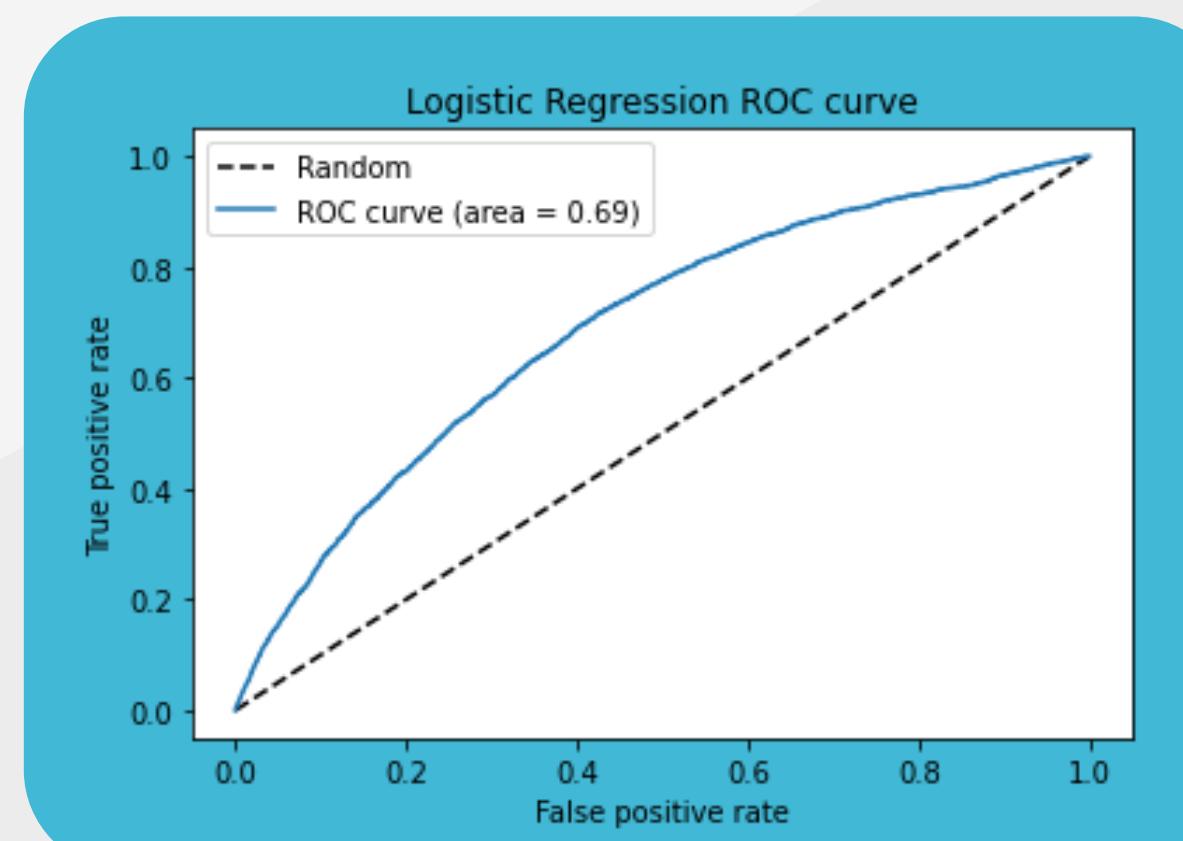
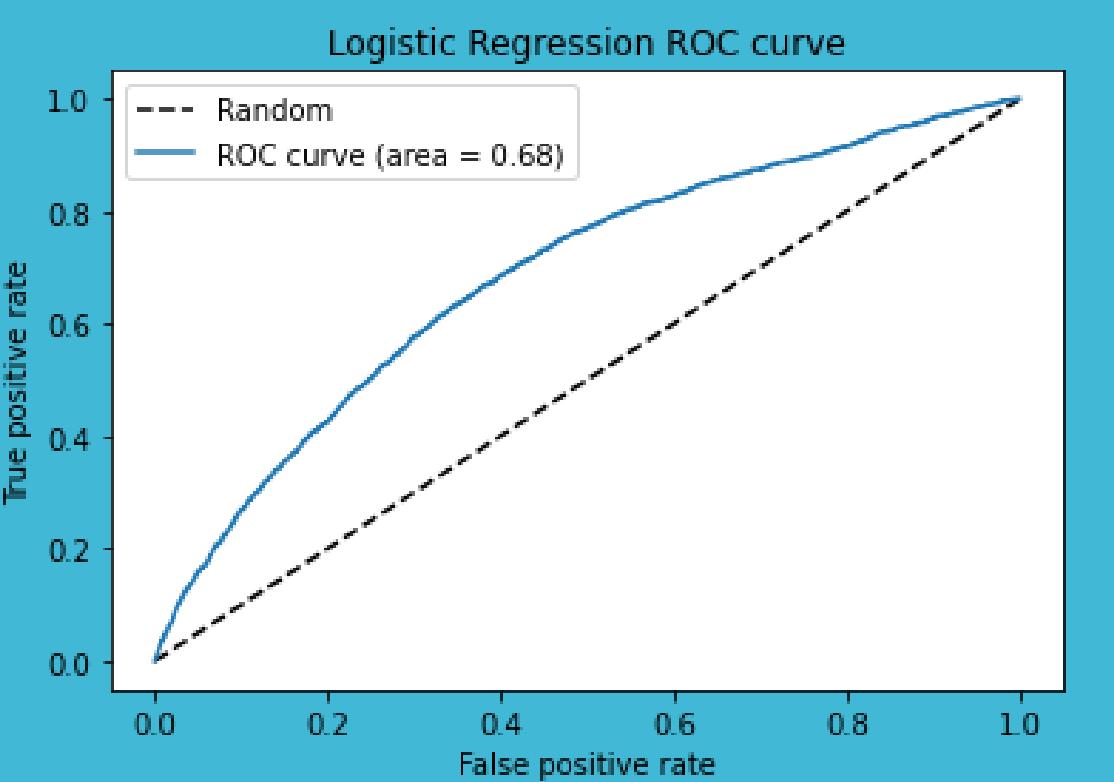
# PERBANDINGAN ROC CURVE

Titik-titik di ROC graph itu menggambarkan semua kemungkinan true positive dan false positive atau confusion matrix nya jika kita jalankan threshold(ambang batas keputusan nya) dari bawah sampe atas. Mulai dari threshold paling bawah yaitu 0, dengan kata lain kita klasifikasikan semua transaction adalah fraud hingga paling atas yaitu 1 atau dengan kata lain kita klasifikasikan semua transactionnya tidak fraud

## ROC CURVE WITHOUT RESAMPLING

## ROC CURVE WITH RESAMPLING (OVERSAMPLING USING SMOTE)

## ROC CURVE WITH RFE





# KESIMPULAN

- Model 1 dan Model 3 memiliki classification report yang serupa , akurasi yang sama yaitu 98%, memiliki nilai AUC yang berbeda yaitu 0.68 dan 0.69 Hal ini menunjukkan bahwa untuk variabel yang mempengaruhi bisa diperoleh dengan mencari korelasinya menggunakan RFE
- Masalah dalam model sebenarnya ialah pada imbalance data, dapat diliat perbedaan pada model 2 ketika data seimbang maka model akan memprediksi dengan lebih baik, dan dapat dilihat juga dengan naik nya nilai AUC menjadi 0.69
- Semakin banyak data yang seimbang serta menggunakan RFE untuk memilih variabel penting yang mempengaruhi target, itu akan membuat model semakin baik dalam memprediksi fraud,





THANK  
YOU

