

---

# **Classification: Non Performing Loan (NPL) Credit Scoring**

Capstone - Deschie Tri Aksara

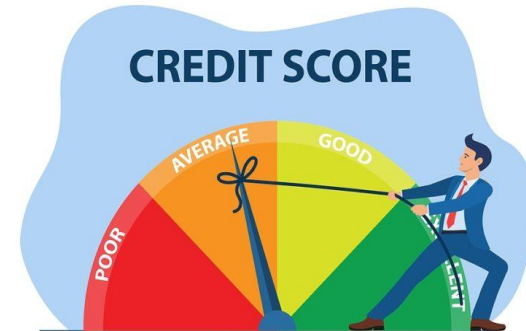
---

# Problem

- Banyak faktor yang dipertimbangkan analis kredit sebelum memberikan rekomendasi menerima atau menolak sebuah pengajuan kredit
- Bank A ingin membuat sistem penilaian kelayakan kredit secara otomatis berdasarkan data yang mereka miliki.

# Goal

- Melakukan analisis dan prediksi kredit macet berdasarkan variabel yang ada.
- Metrics yang dipertimbangkan untuk credit scoring yaitu recall



# Data Description

1. **X:** Nomor
2. **jumlah\_kartu:** Jumlah kartu aktif yang dimiliki pelanggan
3. **outstanding:** Total saldo pemakaian kartu kredit
4. **limit\_kredit:** Jumlah maksimum limit kredit yang dapat digunakan
5. **tagihan:** Jumlah tagihan pada bulan terakhir
6. **total\_pemakaian\_tunai:** Total pemakaian transaksi tunai pada bulan terakhir (cash advance)
7. **total\_pemakaian\_retail:** Total pemakaian transaksi retail pada bulan terakhir
8. **sisa\_tagihan\_tidak\_terbayar:** Jumlah tagihan tidak terbayar pada bulan terakhir
9. **rasio\_pembayaran:** Rasio perbandingan jumlah yang dibayar dengan tagihan pada bulan terakhir
10. **persentasi\_overlimit:** Presentasi overlimit
11. **rasio\_pembayaran\_3bulan:** Rata-rata rasio pembayaran 3 bulan terakhir
12. **rasio\_pembayaran\_6bulan:** Rata-rata rasio pembayaran 3 bulan sebelum 3 bulan terakhir
13. **skor\_delikueni:** Kecenderungan pembayaran menunggak selama 12 bulan ke depan (skor rendah berarti buruk)
14. **flag\_kredit\_macet:** Flag yang menandakan apakah kartu kredit macet atau tidak
15. **jumlah\_tahun\_sejak\_pembukaan\_kredit:** Jumlah tahun sejak pembuatan kartu kredit pertama kali

# Data Description

- 16. **total\_pemakaian**: Total pemakaian kartu kredit (tunai dan retail) pada bulan terakhir
- 17. **sisa\_tagihan\_per\_jumlah\_kartu**: Jumlah tagihan yang tidak terbayar pada bulan terakhir dibagi jumlah kartu aktif
- 18. **sisa\_tagihan\_per\_limit**: Jumlah tagihan yang tidak terbayar pada bulan terakhir dibagi total limit
- 19. **total\_pemakaian\_per\_limit**: Total pemakaian kartu kredit (tunai dan retail) pada bulan terakhir dibagi total limit
- 20. **pemakaian\_3bln\_per\_limit**: Rata-rata pemakaian selama 3 bulan terakhir dibagi total limit
- 21. **pemakaian\_6bln\_per\_limit**: Rata-rata pemakaian selama 3 bulan sebelum 3 bulan terakhir dibagi total limit
- 22. **utilisasi\_3bulan**: Utilisasi kartu kredit selama 3 bulan terakhir
- 23. **utilisasi\_6bulan**: Utilisasi kartu kredit selama 3 bulan sebelum 3 bulan terakhir

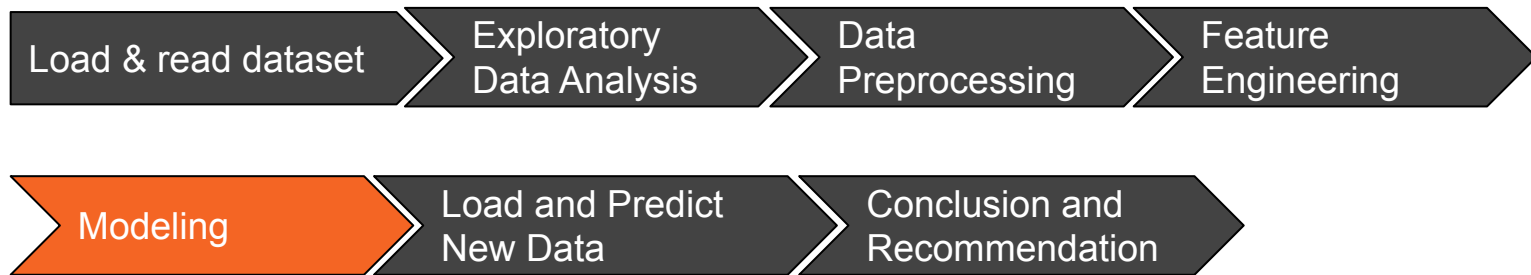
## Response Variable/ Target

flag\_kredit\_macet

0: Tidak Macet

1: Macet

# Data Processing Schema

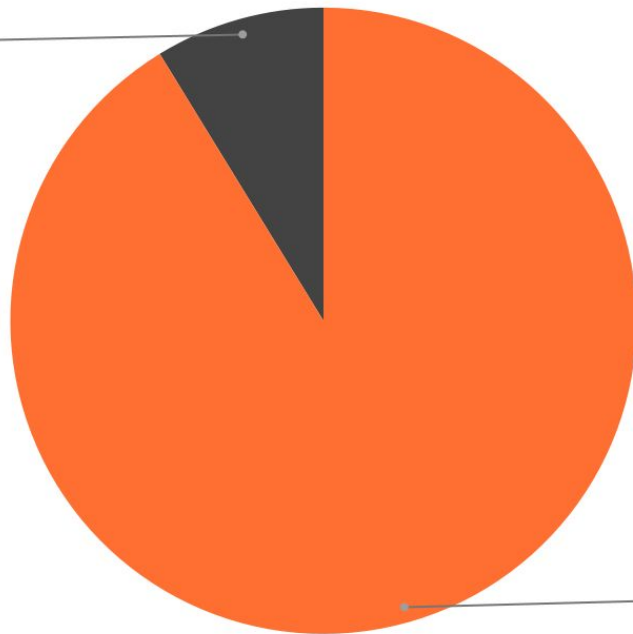


1. Logistic Regression
2. Random Forest Classifier
3. Gradient Boosting Classifier

### Perbandingan Label/ Class

Macet

8.8%



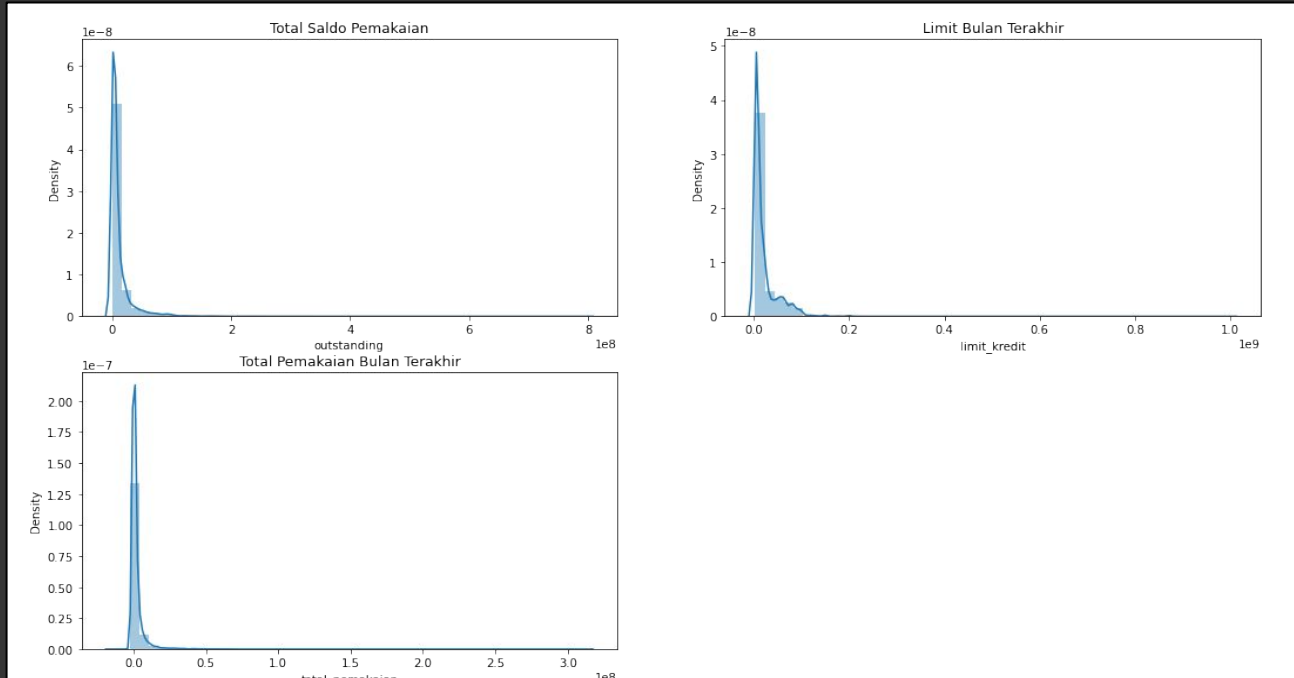
Tidak Macet

91.2%

### Imbalanced Data

Proporsi data tidak seimbang sehingga akan dilakukan resampling dengan menggunakan **SMOTE**.

- Jumlah kartu terbanyak adalah 16 kartu.
- Dari beberapa kolom seperti total\_pemakaian, rasio\_pembayaran, dan total\_pemakaian\_per\_limit, nilai minimum bernilai negatif.



## Outlier

Terdapat perbedaan yang signifikan antara mean dan median dari kolom outstanding, limit\_kredit, dan total\_pemakaian yang mengindikasikan **outlier** atau **skewed distribution**.

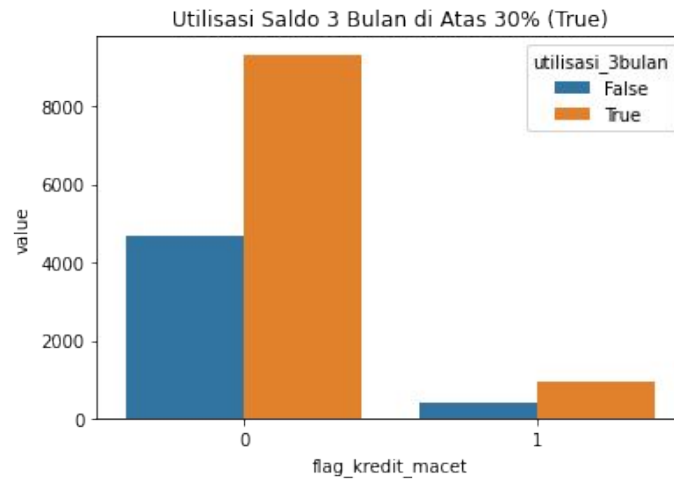
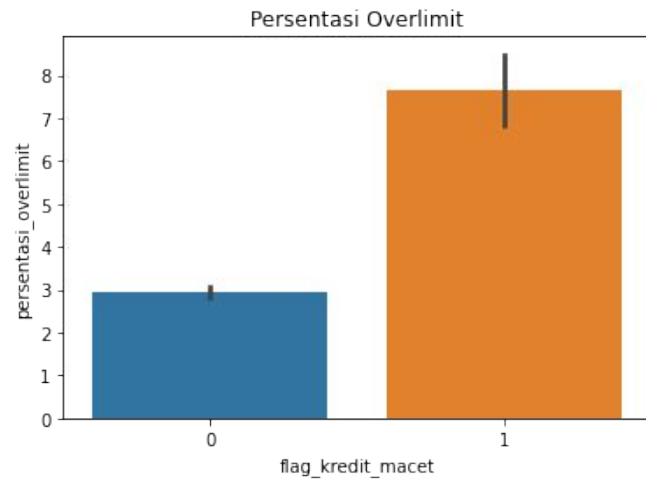
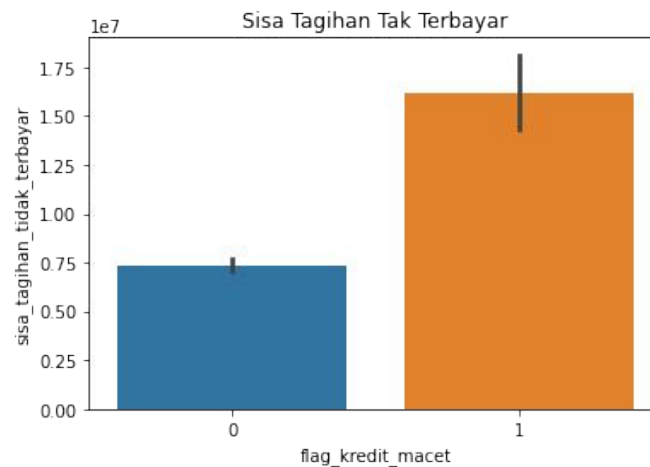
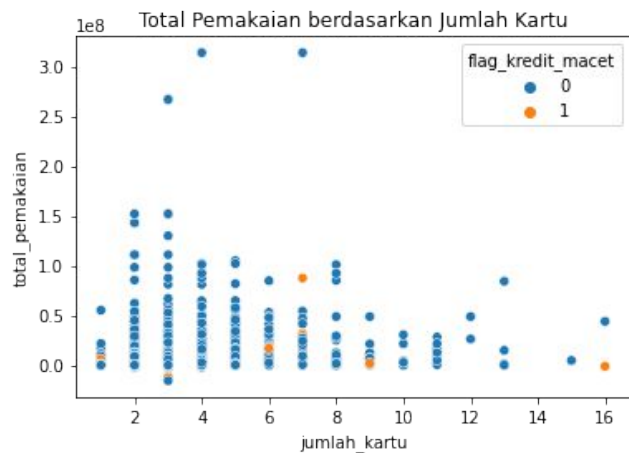


## Exploratory Data Analysis

1. Apakah semakin banyak jumlah kartu semakin besar total saldo yang dipakai?
2. Apakah sisa tagihan tidak terbayar yang tinggi mengindikasikan kredit macet?
3. Bagaimana dengan persentasi overlimit yang besar terhadap flag kredit macet?
4. Apakah nasabah yang menggunakan saldo kreditnya di atas 30% (utilitas kredit  $> 30\%$ ) cenderung terindikasi kredit macet?



# Exploratory Data Analysis



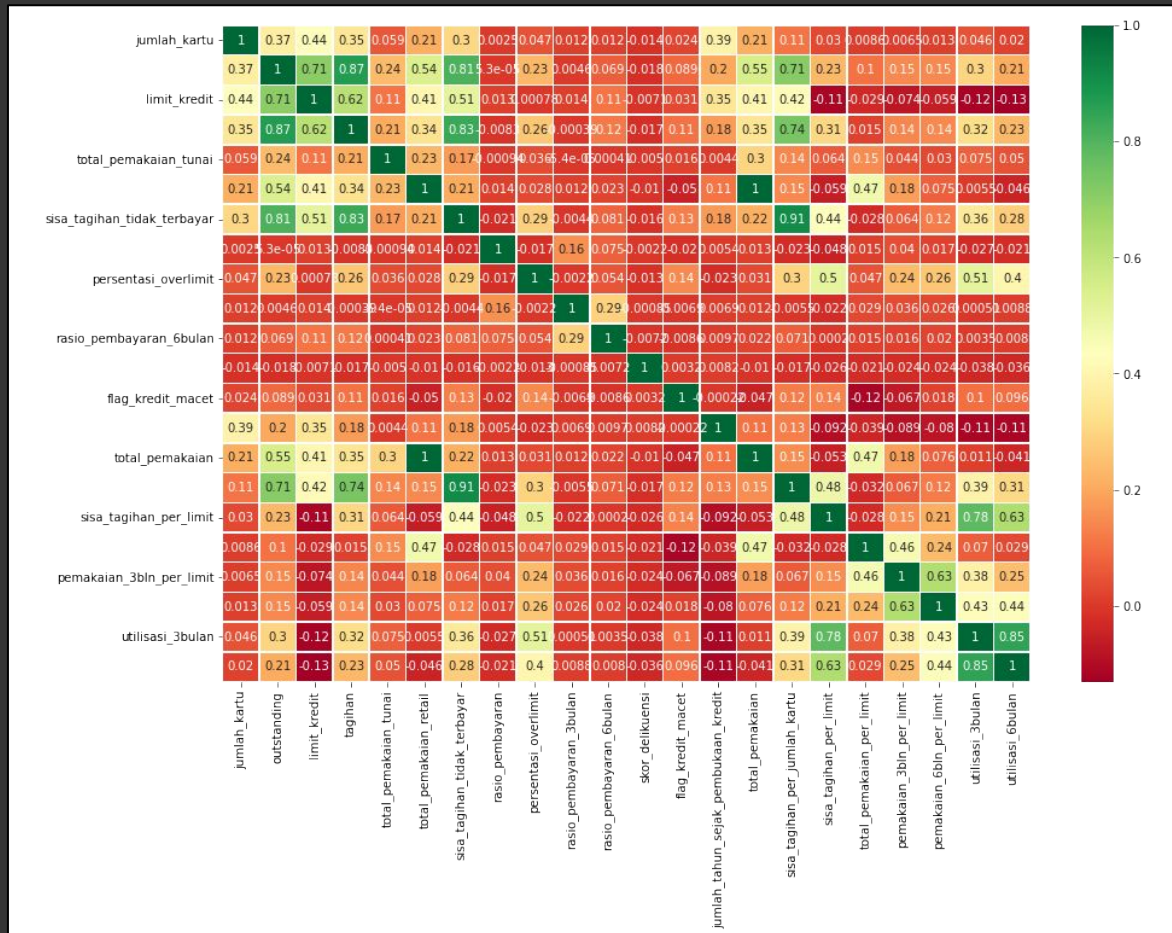
Variabel	Jumlah Missing Value
kode_cabang	100
lainnya	0

drop kolom

total_pemakaian_per_limit	-0.116207
pemakaian_3bln_per_limit	-0.067138
total_pemakaian_retail	-0.049683
total_pemakaian	-0.047419
rasio_pembayaran	-0.020056
rasio_pembayaran_6bulan	-0.008569
rasio_pembayaran_3bulan	-0.006857
jumlah_tahun_sejak_pembukaan_kredit	-0.000222
skor_delikuenasi	0.003236
total_pemakaian_tunai	0.016374
pemakaian_6bln_per_limit	0.018036
jumlah_kartu	0.023667
limit_kredit	0.030994
outstanding	0.089344
utilisasi_6bulan	0.096429
utilisasi_3bulan	0.103251
tagihan	0.107561
sisa_tagihan_per_jumlah_kartu	0.122347
sisa_tagihan_tidak_terbayar	0.129020
sisa_tagihan_per_limit	0.138163
persentase_overlimit	0.143785
flag_kredit_macet	1.000000

# Feature Engineering

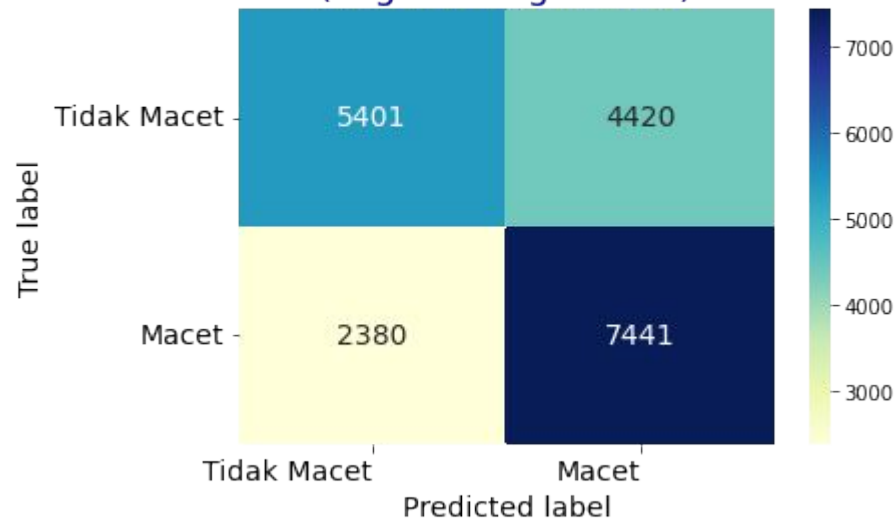
# Feature Selection



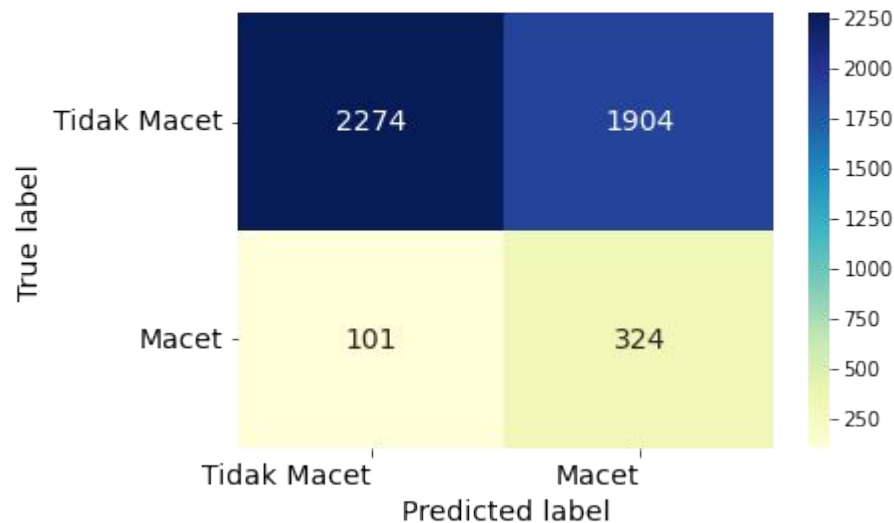
## Korelasi Kuat

Beberapa variabel memiliki **korelasi kuat** ( $>0.7$ ) satu sama lain sehingga dapat dihilangkan dan memilih beberapa variabel saja untuk dilakukan modeling.

Confusion Matrix for Training Model  
(Logistic Regression)

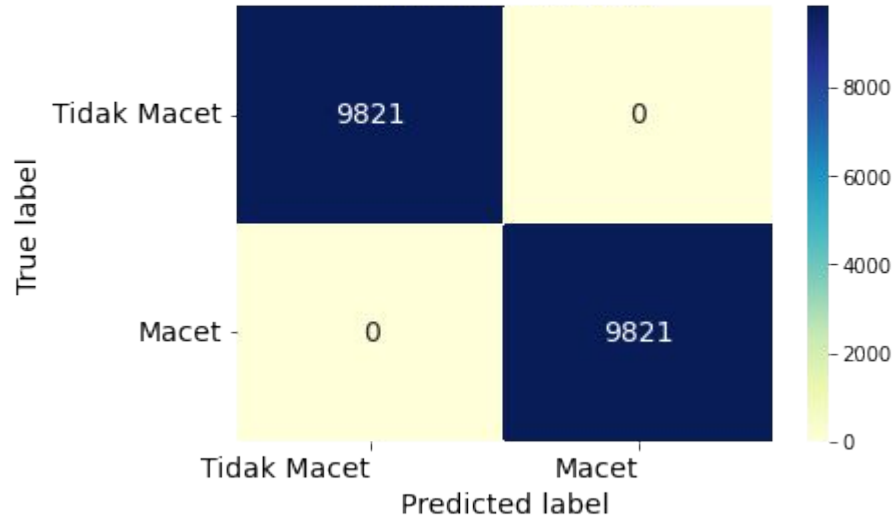


Confusion Matrix for Testing Model  
(Logistic Regression)

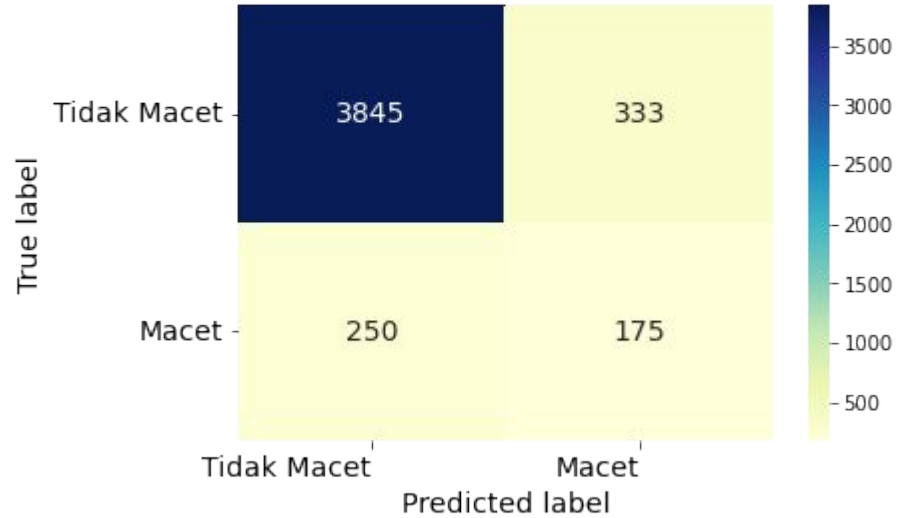


- Data train: akurasi 65% dan recall 76%
- Data testing: akurasi 56% dan recall 76%
- Tidak terjadi peningkatan model score ketika dilakukan cross validation dan tuning parameter.

Confusion Matrix for Training Model  
(Random Forest)

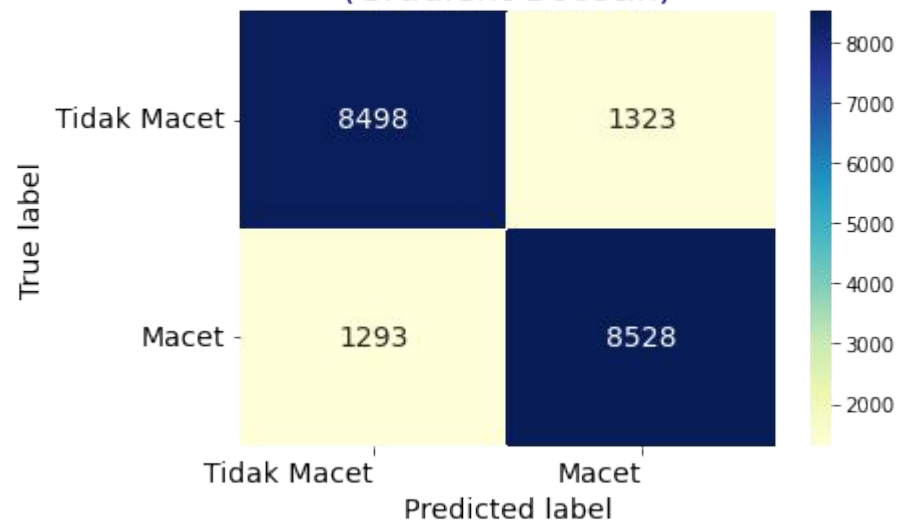


Confusion Matrix for Testing Model  
(Random Forest)

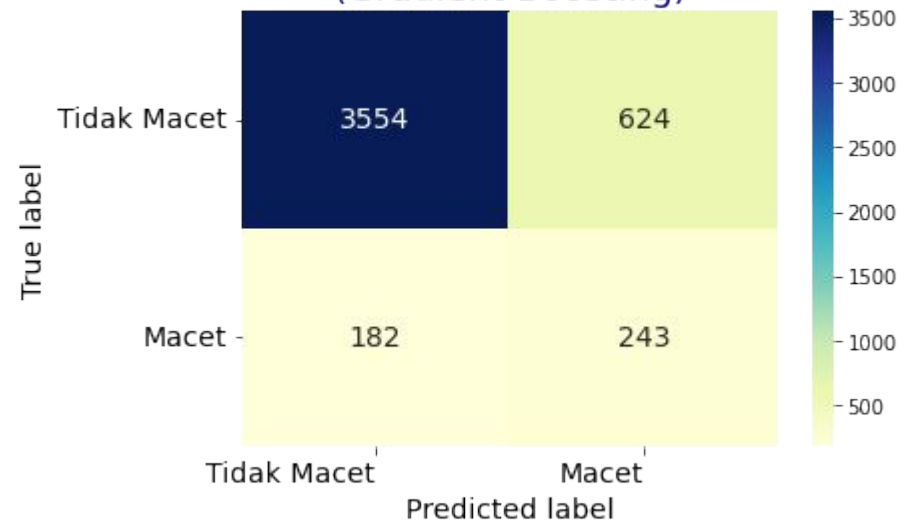


- Data train: akurasi dan recall 100%
- Data testing: akurasi 87% dan recall 41%

Confusion Matrix for Training Model  
(Gradient Boostin)



Confusion Matrix for Testing Model  
(Gradient Boosting)



- Data train: akurasi dan recall 87%
- Data testing: akurasi 82% dan recall 52%

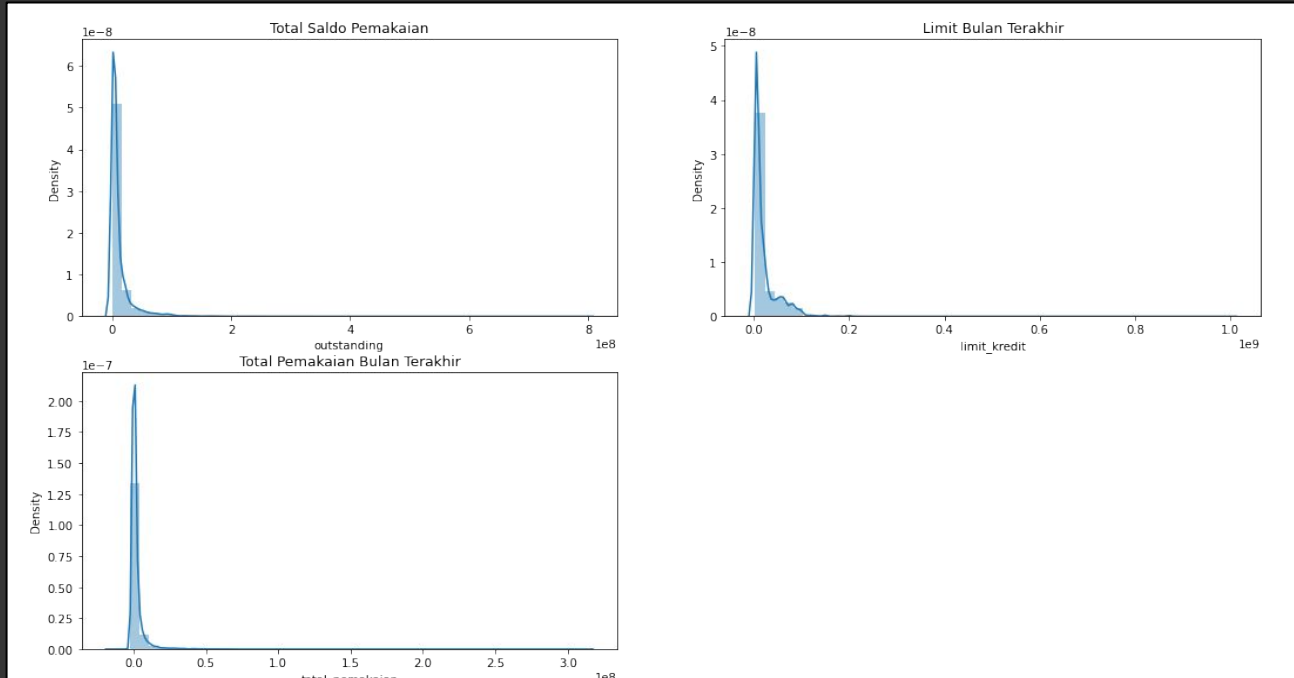
Modeling

Perbandingan Hasil Metrics

Algoritma	Akurasi Data Training	Akurasi Data Testing	Recall Data Training	Recall Data Testing	Performa
Logistic Regression	65%	56%	76%	76%	Kurang baik memprediksi di fase training dan testing
Random Forest Classifier	100%	87%	56%	56%	Over-Fitting
Gradient Boosting Classifier	87%	82%	87%	52%	Appropriate-Fitting

## Load and Predict New Data

- Jumlah kartu terbanyak adalah 16 kartu.
- Dari beberapa kolom seperti total\_pemakaian, rasio\_pembayaran, dan total\_pemakaian\_per\_limit, nilai minimum bernilai negatif.



### Outlier

Terdapat perbedaan yang signifikan antara mean dan median dari kolom outstanding, limit\_kredit, dan total\_pemakaian yang mengindikasikan **outlier** atau **skewed distribution**.





# Conclusion

- Untuk memprediksi kredit macet dari nasabah dengan menggunakan dataset ini model terbaiknya adalah menggunakan algoritma **Gradient Boosting Classifier**.
- Hal ini dikarenakan performa dari model Gradient Boosting Classifier cenderung mampu memprediksi sama baiknya di fase training maupun testing (akurasi training 87%, akurasi testing 82%)
- Di lain sisi algoritma Logistic Regression kurang baik memprediksi di fase testing dan training serta algoritma Random Forest cenderung Under-Fitting.
- Akan tetapi hal ini tidak menjadikan kita untuk menarik kesimpulan bahwasannya jika untuk melakukan pemodelan apapun maka digunakan Gradient Boosting Classifier, kita tetap harus melakukan banyak percobaan model untuk menentukan mana yang terbaik.