CREDIT RISK PREDICTION USING XGBOOST WITH SMOTE AND ADASYN

Yemima Sipayung

LATAR BELAKANG

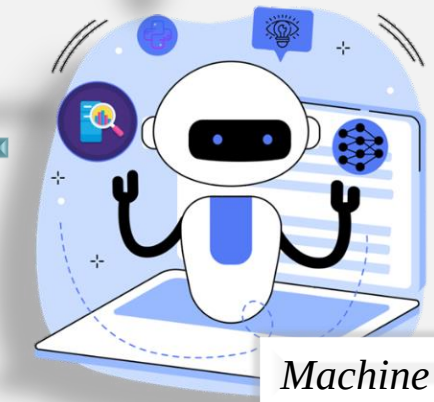
Teknologi finansial adalah salah satu inovasi layanan jasa keuangan yang mulai populer di era digital saat ini.

Peer-to-peer Lending

Risiko pemberian kredit

LATAR BELAKANG

Machine Learning



Permasalahan data dengan kelas tidak seimbang

Evaluasi tingkat risiko kredit calon debitur

LATAR BELAKANG

No.	Judul	Penelitian	Keterangan
1	Credit risk assessment based on gradient boosting decision tree.	Tian dkk. (2020)	Mengaplikasikan teknik resampling ROS, RUS, SMOTE pada beberapa metode ML, yaitu LR, SVM, DT, GBDT, MLP, Adaboost, serta RF. Hasil penelitian ini menyatakan bahwa SMOTE dan GBDT merupakan metode dan teknik resampling yang menghasilkan performa lebih baik daripada algoritma lainnya
2	Extreme learning machine enhanced gradient boosting for credit scoring.	Zou & Gao (2022)	Membuat model prediksi penilaian kredit dengan mengembangkan algoritma GBDT. Zou & Gao menggunakan XGBoost sebagai salah satu base model karena dapat meningkatkan performa model dan lebih efisien dibandingkan dengan GBDT.

LATAR BELAKANG

No.	Judul	Penelitian	Keterangan
3.	Loan default predictive analytics	Owusu dkk. (2022)	Penelitian ini mengaplikasikan teknik resampling ADASYN pada metode ML untuk mengevaluasi kelayakan pinjaman. Penggunaan ADASYN pada penelitian ini berperan dalam menyeimbangkan data dan memberikan kontribusi dalam menghasilkan performa model yang lebih baik dan kurang bias.

RUMUSAN MASALAH

Bagaimana perbandingan performa metode *XGBoost*, SMOTE-*XGBoost*, dan ADASYN-*XGBoost* dalam memprediksi tingkat risiko kredit pada data dengan kelas yang tidak seimbang.

TUJUAN PENELITIAN

Untuk mengetahui perbandingan performa metode *XGBoost*, SMOTE-*XGBoost*, dan ADASYN-*XGBoost* dalam memprediksi tingkat risiko kredit pada data dengan kelas yang tidak seimbang.

RISIKO KREDIT

Risiko kredit disebabkan oleh pihak debitur yang tidak dapat atau tidak mau memenuhi kewajiban untuk membayar kembali dana yang dipinjamnya secara penuh pada saat jatuh tempo atau sesudahnya

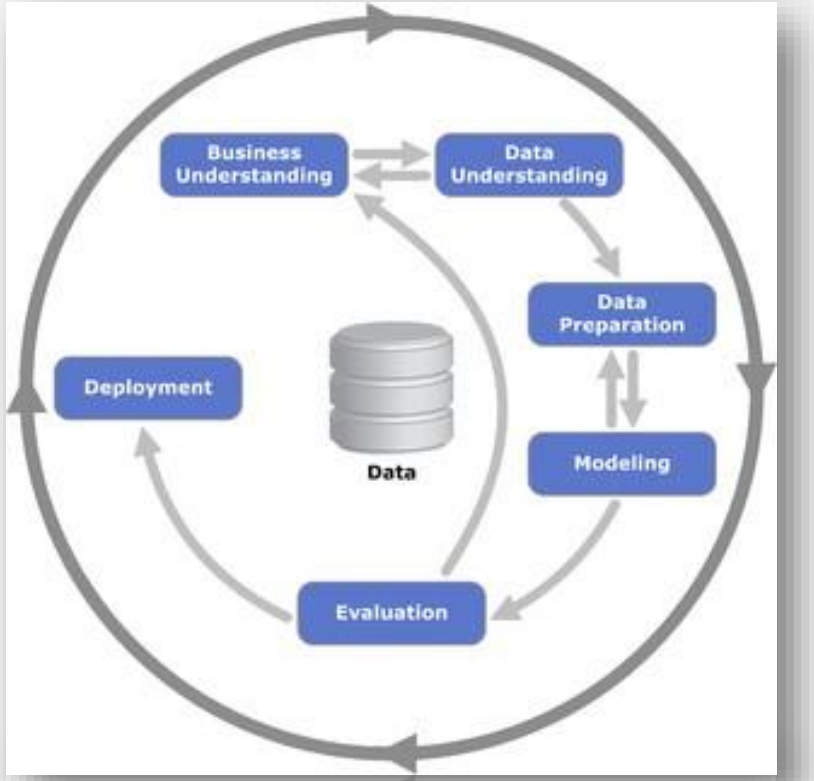
Kredit dikatakan **lancar** apabila debitur selalu membayar pokok dan bunga tepat waktu

Kredit dikatakan **kurang lancar** apabila debitur menunggak pembayaran pokok dan/ atau bunga sampai dengan 120 hari

Kredit dikatakan **macet** apabila debitur tidak mampu melanjutkan pembayaran pokok dan/ atau bunga

DATA MINING

Data mining merupakan kegiatan untuk menemukan suatu pola atau relasi dalam skala besar dengan menggunakan data historis yang telah dikumpulkan.



ONE-HOT ENCODING

ID	Buah
1	Melon
2	Pisang
3	Semangka
4	Pisang

One-hot encoding

ID	Buah_Melon	Buah_Pisang	Buah_Semangka
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

ANALISIS KORELASI

Korelasi rank spearman

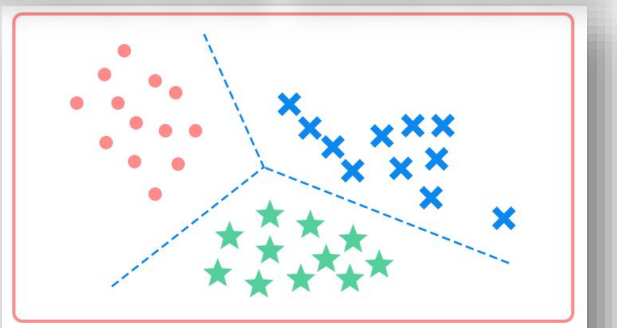
$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Korelasi rank biserial

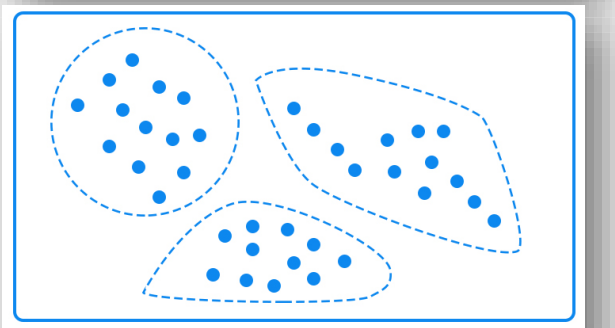
$$r_{rb} = \frac{2 \cdot [M(\boldsymbol{X}_1) - M(\boldsymbol{X}_0)]}{n}$$

dimana $d_i = R(\boldsymbol{X}_i) - R(\boldsymbol{Y}_i)$ dengan $R(\boldsymbol{X}_i)$ merupakan peringkat pada data ke-i fitur pertama, $R(\boldsymbol{Y}_i)$ merupakan peringkat pada data ke-i fitur kedua, $M(\boldsymbol{X}_1)$ adalah rata-rata variabel kontinu untuk kelompok biner dengan nilai 1, $M(\boldsymbol{X}_0)$ adalah rata-rata variabel kontinu untuk kelompok biner dengan nilai 0, dan n adalah jumlah sampel data

MACHINE LEARNING

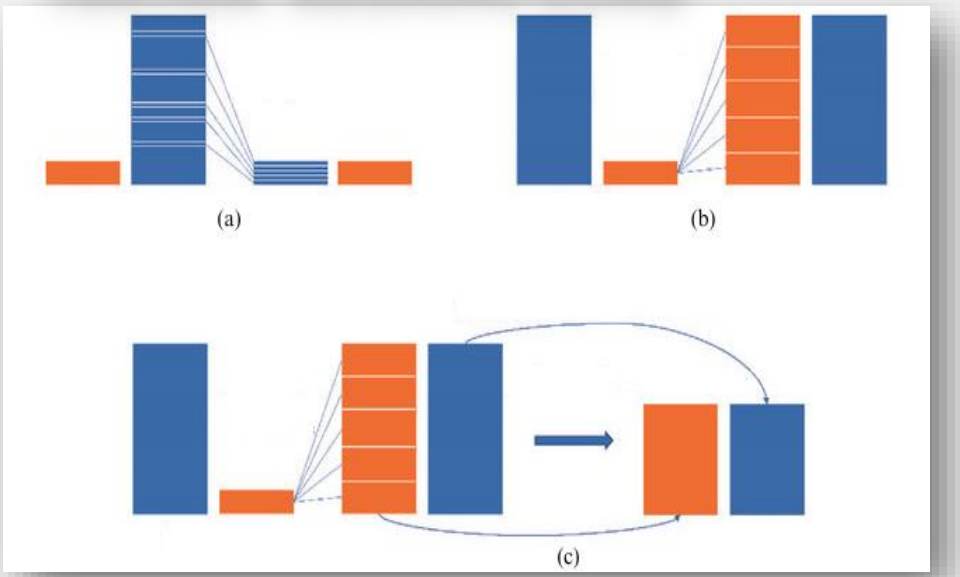


Supervised learning

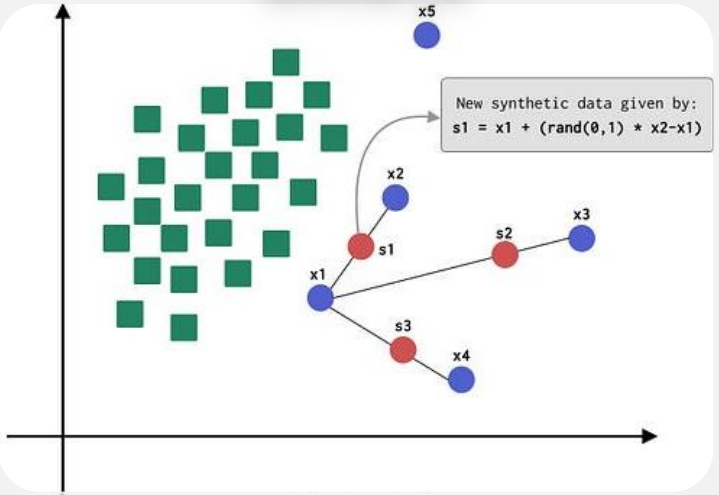


Unsupervised learning

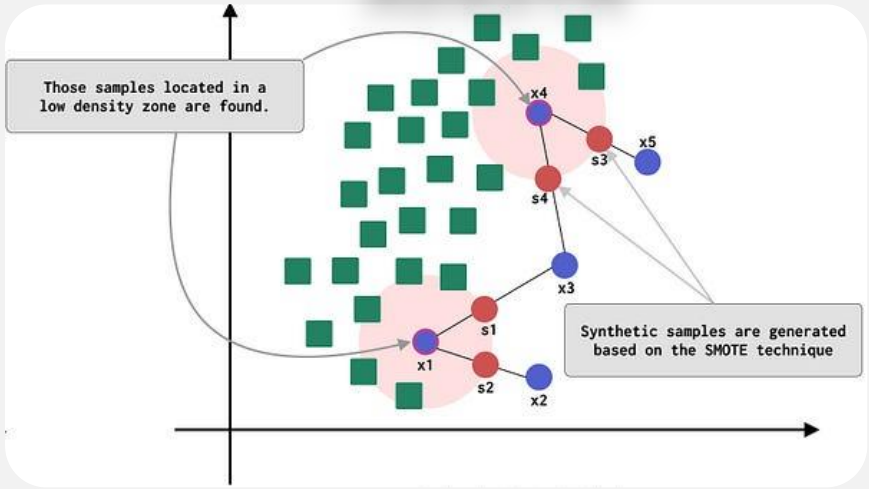
TEKNIK RESAMPLING



SMOTE



ADASYN

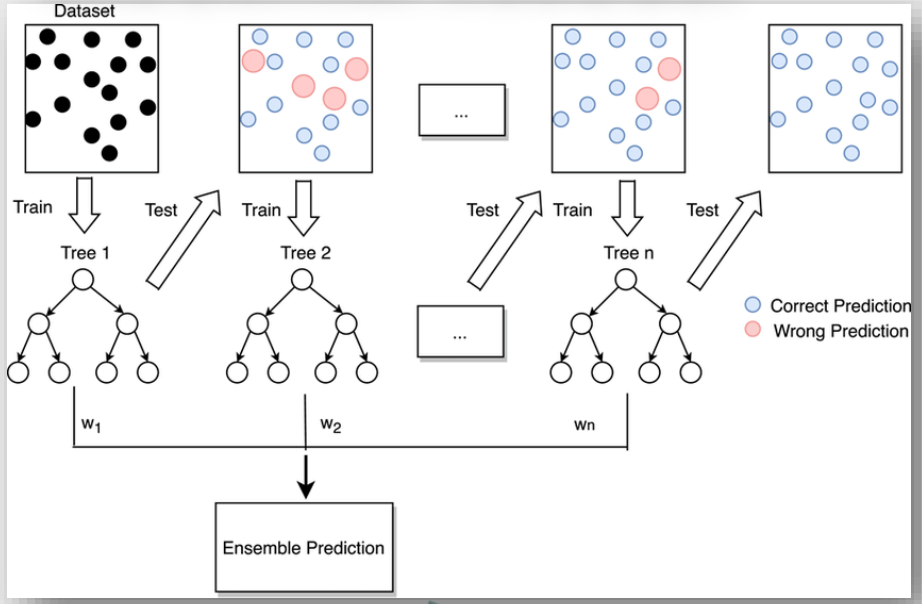


SMOTE menciptakan sampel sintetis di antara titik data minoritas terdekat, sementara ADASYN menggunakan bobot adaptif untuk menghasilkan sampel sintetis di dekat area yang kurang representatif dari kelas minoritas.

$$d_{euclid}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_{1,i} - \boldsymbol{x}_{2,i})^2}$$

$$d_{modi\_euclid}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_{1,i} - \boldsymbol{x}_{2,i})^2 + \sum_{1}^{q} Med^2}$$

ALGORITMA XGBOOST



$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i), f_k \in F$$

Fungsi objektif algoritma *XGBoost* didefinisikan dengan:

$$\mathcal{L}^{(t)}\left(y, \hat{y}^{(t)}\right) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k),$$

dimana $\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} \omega_{kj}^2$

Cross entropy loss dan fungsi *softmax* digunakan dalam loss function $l\left(y_i, \hat{y}_i^{(t)}\right)$, sehingga diperoleh:

$$l(y,p) = -\sum_{c=1}^{C} y_c \log(p_c), \quad p_c = \frac{e^{\hat{y}_i^{(t)}}}{\sum_{c=1}^{C} e^{\hat{y}_i^{(t)}}}$$

Untuk setiap simpul daun, simpul fitur dengan nilai *gain* yang terbesar dipilih sebagai *split point*

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]$$

Keterangan

$l\left(y_i, \hat{y}_i^{(t)}\right)$: loss function

$\hat{y}_i$: hasil prediksi

y_i : nilai actual

f_k : hasil prediksi pohon keputusan ke-k

$\Omega(f_k)$: kompleksitas model

F : himpunan pohon klasifikasi atau CART

T_k : jumlah simpul daun dari pohon ke-k

γ : pengurangan kerugian minimum yang diperlukan untuk membuat partisi lebih lanjut pada simpul daun T

λ : koefisien penalty

ω_{kj} : skor dari simpul daun ke-j dan pohon ke-k

p_s : nilai probabilitas untuk kelas ke-s

$\hat{y}_s^{(t)}$: hasil prediksi model untuk kelas ke-s

y_s : label kelas ke-s dalam bentuk *one-hot encoding* dan merupakan representasi label dalam bentuk vektor yang berisi semua nol kecuali pada indeks yang sesuai dengan kelas, yaitu bernilai 1.

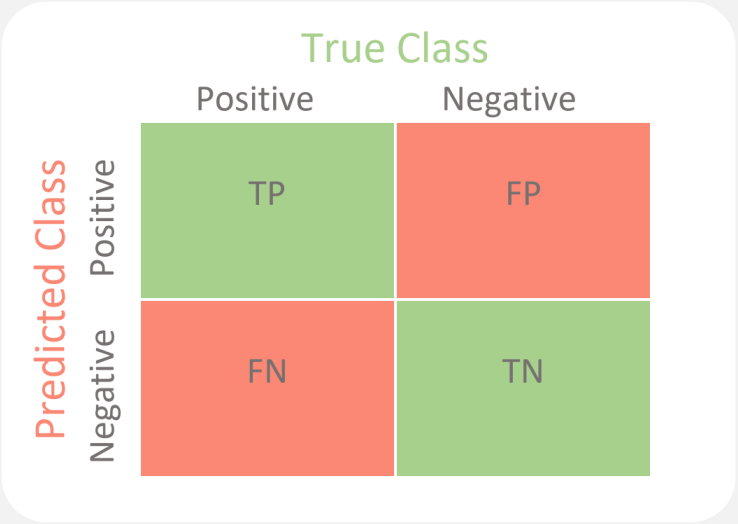
G_L : Gradien loss function dari data latih untuk node sisi kiri

H_L : Hessian loss function dari data latih untuk node sisi kiri

G_R : Gradien loss function dari data latih untuk node sisi kanan

H_R : Hessian loss function dari data latih untuk node sisi kanan

CONFUSION MATRIX



$$Accuracy_c = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}$$

$$Precision_c = \frac{TP_c}{TP_c + TN_c}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

$$F1 - Score_c = \frac{2 \cdot (Precision_c \cdot Recall_c)}{Precision_c + Recall_c}$$

dengan $c = \{1,2, \dots, C\}$ merepresentasikan kelas yang sedang dievaluasi. Pada kasus multikelas, evaluasi model menggunakan rata-rata dari metrik evaluasi dengan *macro average*. Contoh perhitungan metrik evaluasi dengan *macro average* adalah

$$Accuracy_{macro} = \frac{\sum_{c=1}^{C} Accuracy_c}{C}$$

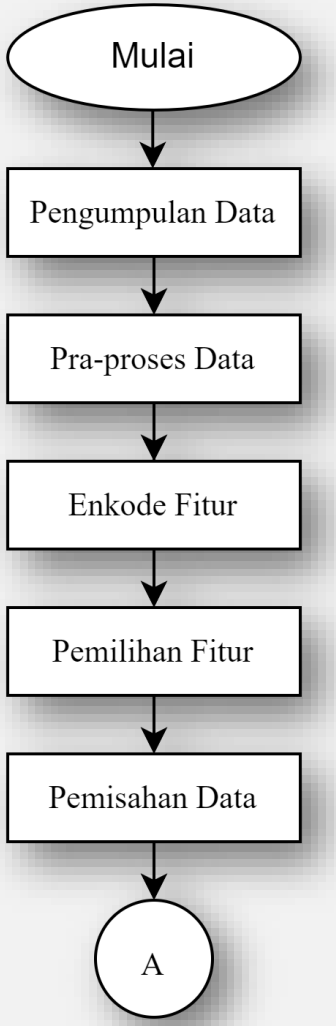
SUMBER DATA

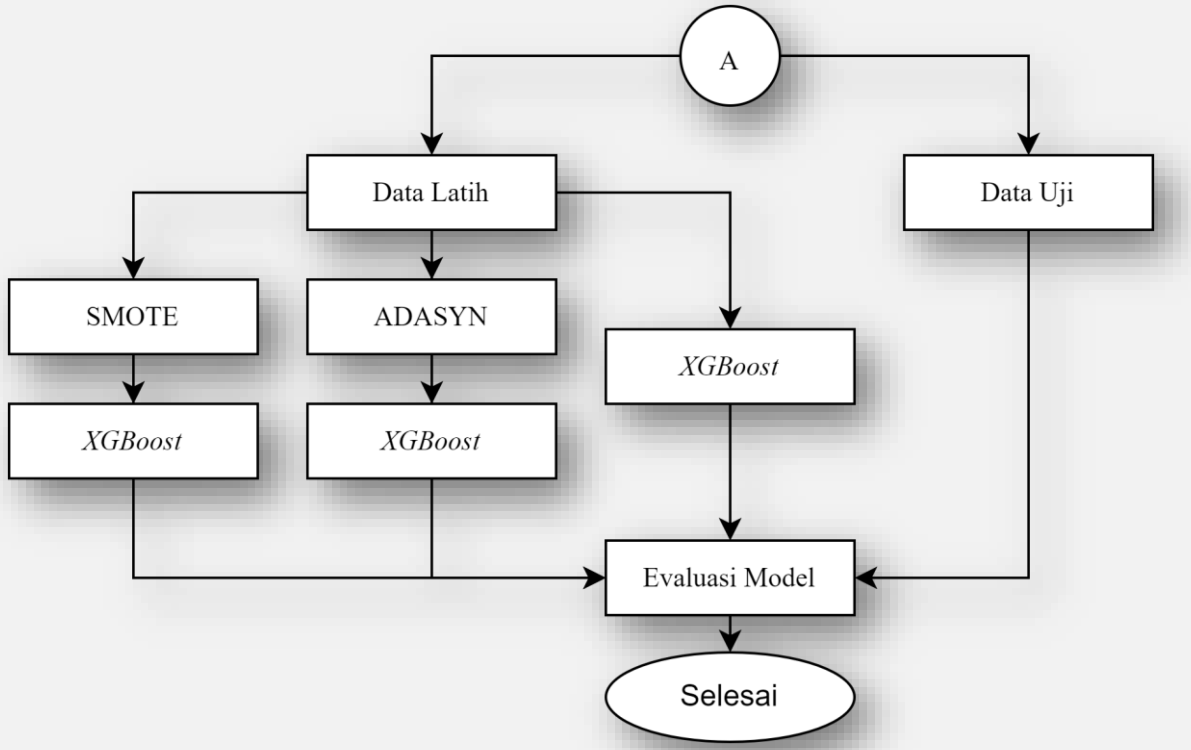
Data yang digunakan : Pengajuan pinjaman yang dilakukan oleh calon debitur melalui *Lending Club* tahun 2014 – 2021
Sumber : Kaggle (2023)
Jumlah sampel : 242059
Jumlah fitur : 74

Status pinjaman dikelompokkan menjadi tiga jenis kredit, yaitu *Good*, *Poor* dan *Bad*.

Kelas	Status Pinjaman
Good	Fully Paid
Poor	In Grace Period, Late (16-30 days), dan Late (31-120 days)
Bad	Charged Off, Default, Does not meet the credit policy. Status: Charged Off, Does not meet the credit policy. Status: Fully Paid

TAHAPAN PENELITIAN





TOOLS

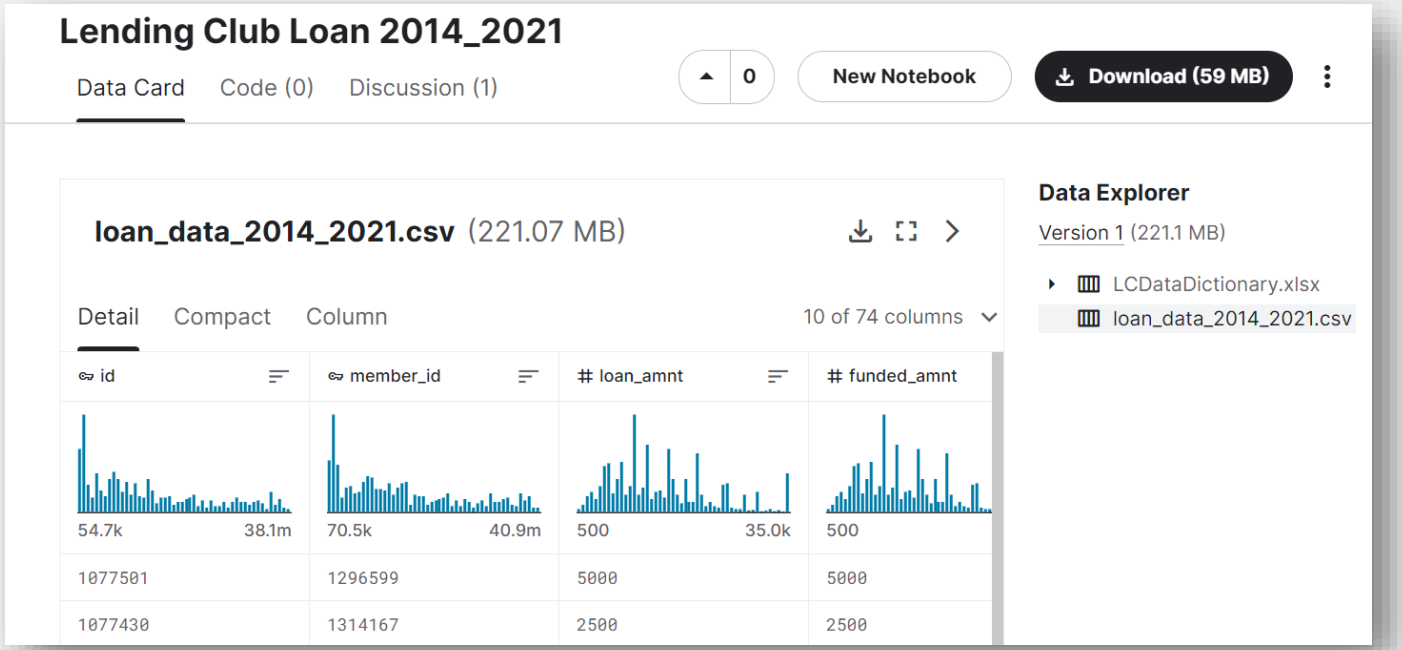
Software

Python

Jupyter Notebook

Hardware atau perangkat keras yang digunakan pada penelitian ini adalah prosesor AMD Ryzen 3 4300U with Radeon Graphics 2.70 GHz, RAM 8.00 GB. Sistem operasi yang digunakan adalah Windows 11 *Home Single Language* 64-bit.

PENGUMPULAN DATA



PRA-PROSES DATA

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik
id	int64	0	466179
member_id	int64	0	466178
loan_amnt	int64	0	1352
funded_amnt	int64	0	1354
funded_amnt_inv	float64	0	9854
term	object	0	2
int_rate	float64	0	506
installment	float64	0	55622
grade	object	0	7
sub_grade	object	0	35
emp_title	object	27588	205475
emp_length	object	21008	11
home_ownership	object	0	6

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik
annual_inc	float64	4	31896
verification_status	object	0	3
issue_d	object	0	91
loan_status	object	0	9
pymnt_plan	object	0	2
url	object	0	466179
desc	object	340302	124436
purpose	object	0	14
title	object	20	63099
zip_code	object	0	888
addr_state	object	0	50
dti	float64	0	3997
delinq_2yrs	float64	29	24
earliest_cr_line	object	29	664
inq_last_6mths	float64	29	28
mths_since_last_delinq	float64	250351	145
mths_since_last_record	float64	403647	123
open_acc	float64	29	62
pub_rec	float64	29	26

PRA-PROSES DATA

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik
revol_bal	int64	0	58038
revol_util	float64	340	1269
total_acc	float64	29	112
initial_list_status	object	0	2
out_prncp	float64	0	135647
out_prncp_inv	float64	0	141173
total_pymnt	float64	0	351576
total_pymnt_inv	float64	0	347602
total_rec_prncp	float64	0	172697
total_rec_int	float64	0	270039
total_rec_late_fee	float64	0	5808
recoveries	float64	0	22773
collection_recovery_fee	float64	0	20275
last_pymnt_d	object	376	85
last_pymnt_amnt	float64	0	198182
next_pymnt_d	object	227214	85
last_credit_pull_d	object	42	92
collections_12_mths_ex_med	float64	145	9
mths_since_last_major_derog	float64	367311	162
policy_code	int64	0	1
application_type	object	0	1

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik
annual_inc_joint	float64	466285	0
dti_joint	float64	466285	0
verification_status_joint	float64	466285	0
acc_now_delinq	float64	29	6
tot_coll_amt	float64	70276	6315
tot_cur_bal	float64	70276	220327
open_acc_6m	float64	466285	0
open_il_6m	float64	466285	0
open_il_12m	float64	466285	0
open_il_24m	float64	466285	0
mths_since_rcnt_il	float64	466285	0
total_bal_il	float64	466285	0
il_util	float64	466285	0
open_rv_12m	float64	466285	0
open_rv_24m	float64	466285	0
max_bal_bc	float64	466285	0
all_util	float64	466285	0
total_rev_hi_lim	float64	70276	14600
inq_fi	float64	466285	0
total_cu_tl	float64	466285	0
inq_last_12m	float64	466285	0

PRA-PROSES DATA

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik	Penanganan
loan_amnt	int64	0	1352	
funded_amnt	int64	0	1354	
funded_amnt_inv	float64	0	9854	
term	object	0	2	Konversi nilai menjadi tipe numerik
grade	object	0	7	
int_rate	float64	0	506	
installment	float64	0	55622	
home_ownership	object	0	6	
verification_status	object	0	3	
loan_status	object	0	9	Mengelompokkan nilai menjadi tiga kelas (good, poor, dan bad)
pymnt_plan	object	0	2	
purpose	object	0	14	
dti	float64	0	3997	
emp_length	object	21008	11	Konversi nilai menjadi tipe numerik, dan imputasi null values dengan mean
annual_inc	float64	4	31896	Data dengan null values terhapus
delinq_2yrs	float64	29	24	
inq_last_6mths	float64	29	28	

PRA-PROSES DATA

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik	Penanganan
initial_list_status	object	0	2	
revol_bal	int64	0	58038	
out_prncp	float64	0	135647	
out_prncp_inv	float64	0	141173	
total_pymnt	float64	0	351576	
total_pymnt_inv	float64	0	347602	
total_rec_prncp	float64	0	172697	
total_rec_int	float64	0	270039	
total_rec_late_fee	float64	0	5808	
recoveries	float64	0	22773	
collection_recovery_fee	float64	0	20275	
last_pymnt_amnt	float64	0	198182	
tot_coll_amt	float64	70276	6315	Menghapus sampel data dengan null values
tot_cur_bal	float64	70276	220327	
total_rev_hi_lim	float64	70276	14600	
last_pymnt_d	object	376	85	Menghapus null values karena memiliki jumlah null values yang sedikit **setelah** dilakukan penanganan dari tiga kolom sebelumnya
last_credit_pull_d	object	42	92	
issue_d	object	0	91	

PRA-PROSES DATA

Variabel	Tipe data	Jumlah null values	Jumlah nilai unik	Penanganan
mths_since_last_delinq	float64	250351	145	Menghapus kolom
mths_since_last_record	float64	403647	123	
mths_since_last_major_derog	float64	367311	162	
open_acc	float64	29	62	Null values terhapus bersamaan dengan penanganan yang telah dilakukan dari tiga kolom tot_coll_amt, tot_cur_bal, dan total_rev_hi_lim
pub_rec	float64	29	26	
total_acc	float64	29	112	
collections_12_mths_ex_med	float64	145	9	
acc_now_delinq	float64	29	6	
revol_util	float64	340	1269	Imputasi null values dengan mean

ENKODE FITUR

Kolom
home_ownership
Verification_status
pymnt_plan
purpose
initial_list_status

Kolom
home_ownership_ANY
home_ownership_MORTGAGE
home_ownership_NONE
home_ownership_OTHER
home_ownership_OWN
home_ownership_RENT
verification_status_Not Verified
verification_status_Source Verified
verification_status_Verified
pymnt_plan_n
pymnt_plan_y
purpose_car
purpose_credit_card
purpose_debt_consolidation
purpose_home_improvement
purpose_house
purpose_major_purchase
purpose_medical
purpose_moving
purpose_other
purpose_renewable_energy
purpose_small_business
purpose_vacation
purpose_wedding
initial_list_status_f
initial_list_status_w

PEMILIHAN FITUR

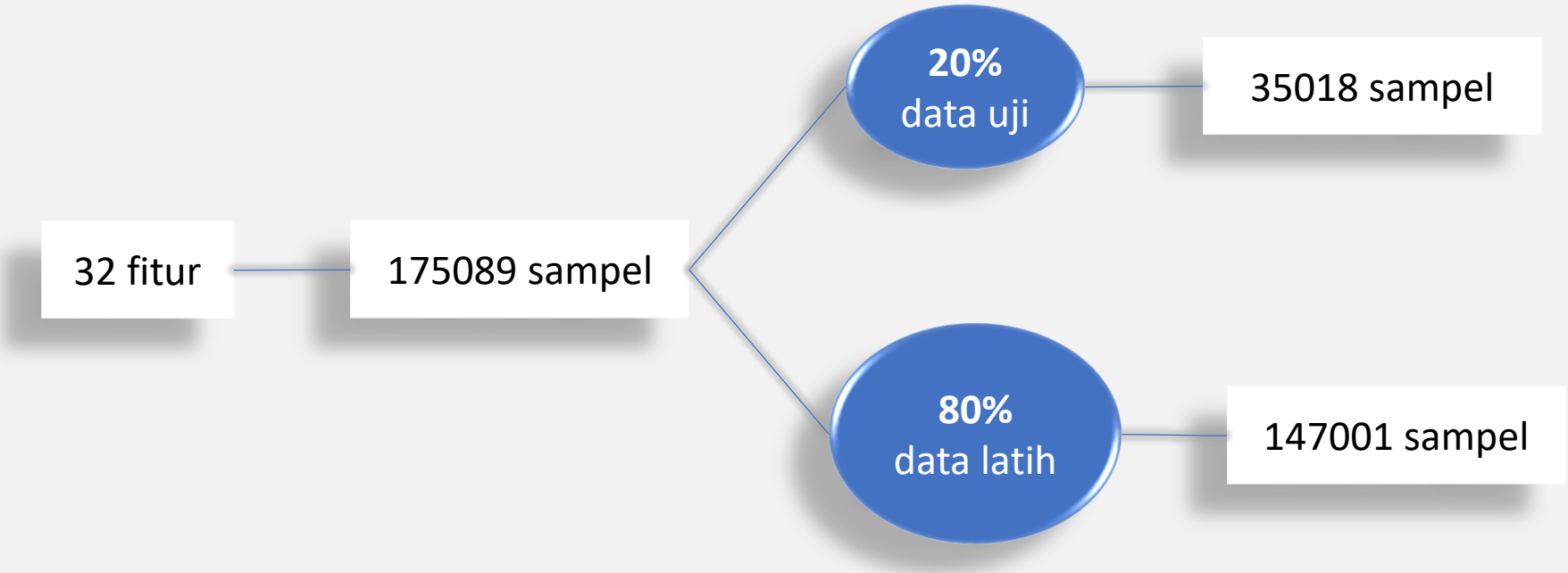
Fitur Kontinu

Fitur	
last_pymnt_amnt	acc_now_delinq
total_rec_prncp	collections_12_mths_ex_med
total_pymnt	open_acc
grade	tot_coll_amt
annual_inc	delinq_2yrs
mths_last_credit_pull_d	inq_last_6mths
mths_issue_d	installment
tot_cur_bal	funded_amnt
total_rev_hi_lim	total_rec_int
total_acc	revol_util
mths_last_pymnt_d	dti
emp_length	term
pub_rec	total_rec_late_fee
collection_recovery_fee	out_prncp

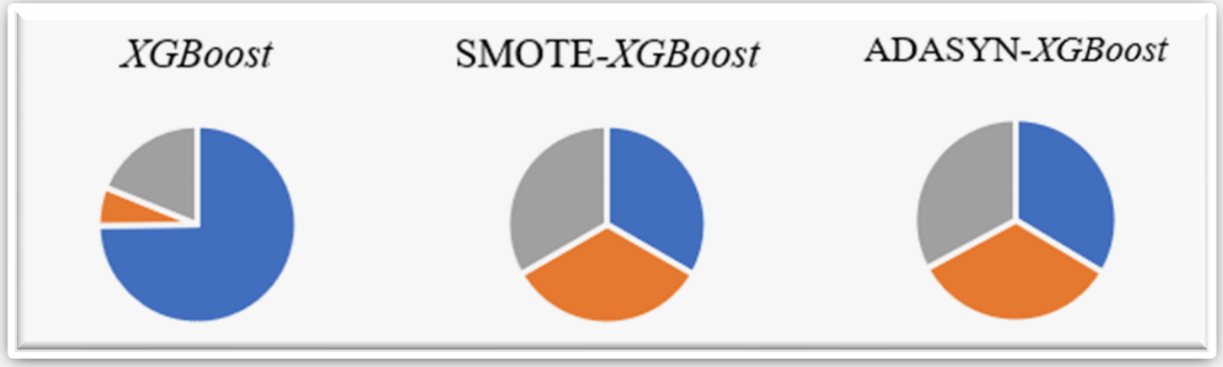
Fitur Biner

Fitur
pymnt_plan_n
initial_list_status_f
purpose_debt_consolidation
home_ownership_MORTGAGE

PEMISAHAN DATA



PEMODELAN



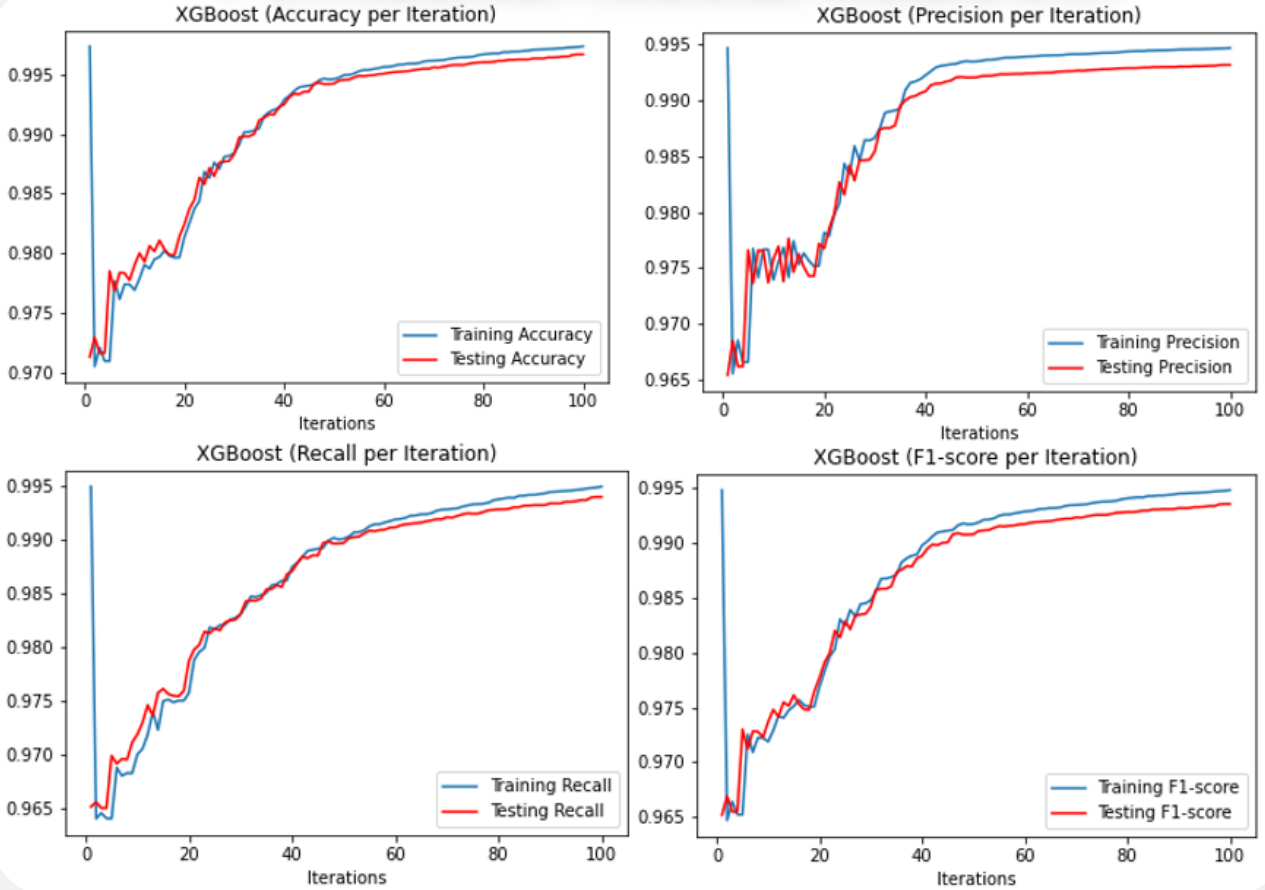
	XGBoost	SMOTE-XGBoost	ADASYN-XGBoost
Bad	26406	104802	102534
Poor	8863	104802	104721
Good	104802	104802	104802
Total	140071	314406	312057

Iterasi	XGBoost	SMOTE-XGBoost	ADASYN-XGBoost
1	0.97233	0.97855	0.97762
2	0.86695	0.87381	0.87649
3	0.77756	0.78508	0.79095
4	0.7005	0.71209	0.71516
5	0.63353	0.64581	0.64989
⋮	⋮	⋮	⋮
96	0.01528	0.01879	0.02079
97	0.01501	0.01846	0.0205
98	0.01483	0.01825	0.02015
99	0.01461	0.01788	0.01997
100	0.01443	0.01754	0.01982

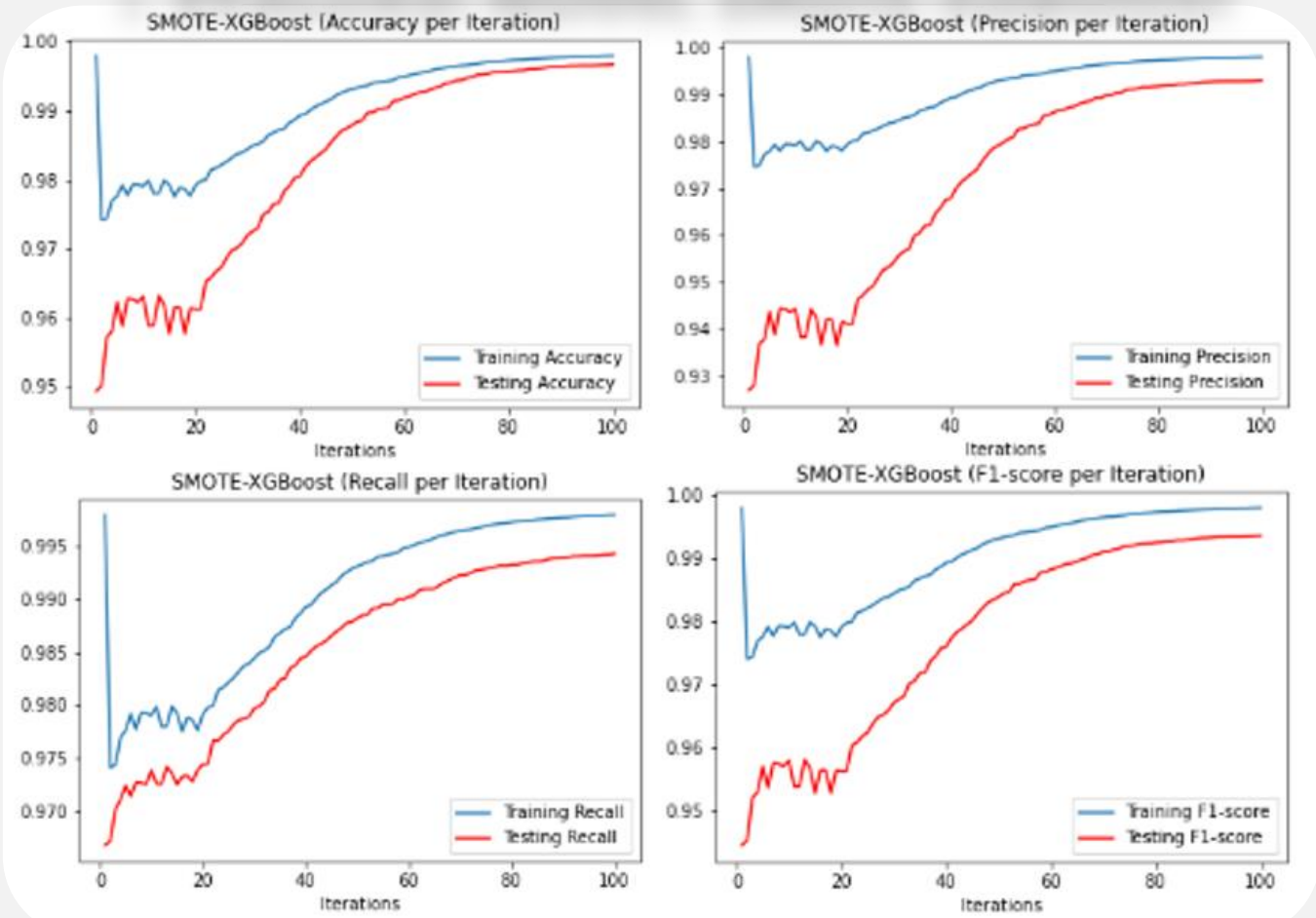
EVALUASI MODEL

XGBoost		Prediksi		
		Bad	Poor	Good
Aktual	Bad	6482	39	76
	Poor	0	2186	1
	Good	0	0	26234
SMOTE-XGBoost		**Prediksi**		
		Bad	Poor	Good
Aktual	Bad	6489	39	69
	Poor	0	2186	1
	Good	6	0	26228
ADASYN-XGBoost		**Prediksi**		
		Bad	Poor	Good
Aktual	Bad	6509	39	49
	Poor	0	2186	1
	Good	28	0	26206

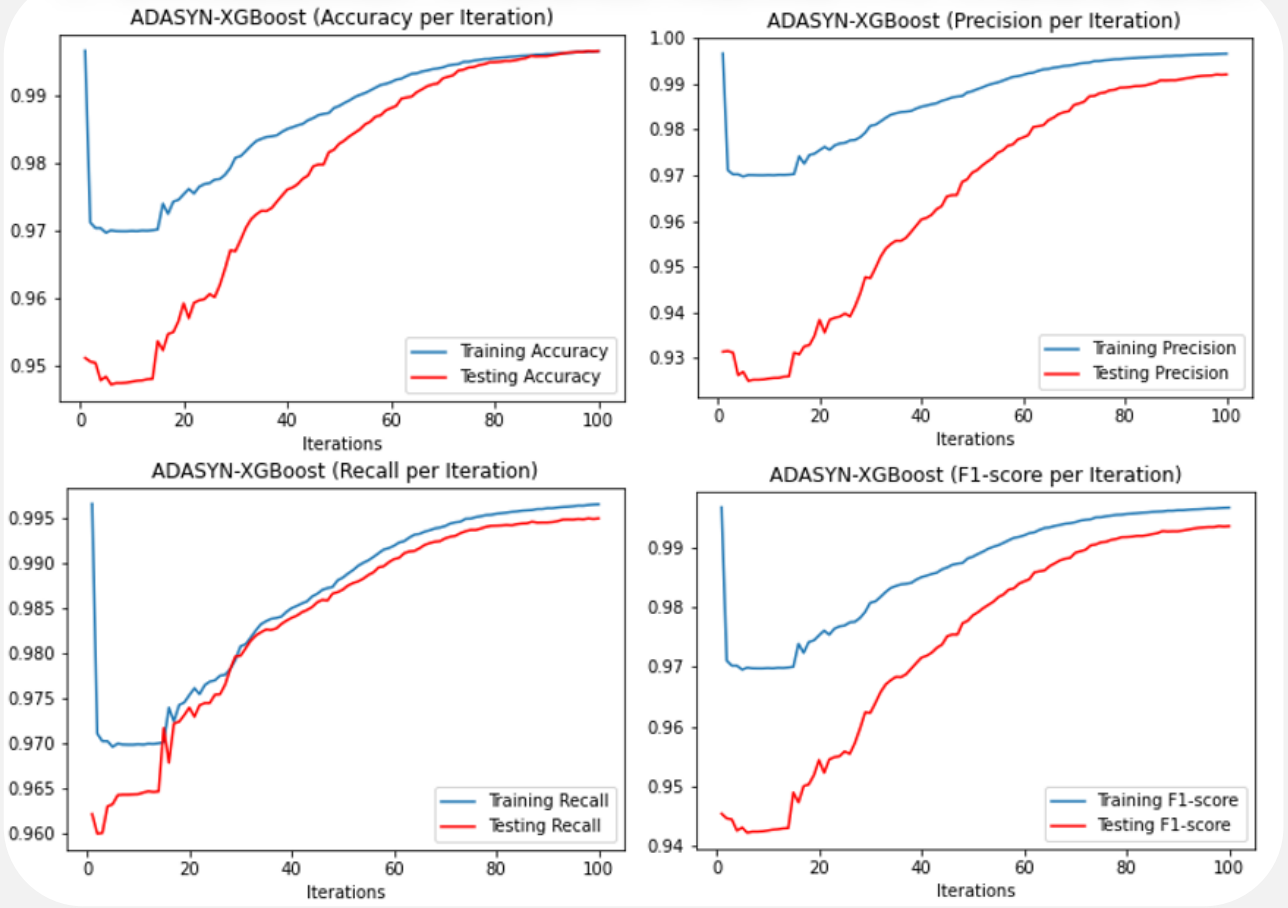
1. EVALUASI MODEL XGBOOST



2. EVALUASI MODEL SMOTE-XGBOOST



3. EVALUASI MODEL ADASYN-XGBOOST



4. PERBANDINGAN HASIL EVALUASI MODEL

	XGBoost		SMOTE-XGBoost		ADASYN-XGBoost	
	Latih	Uji	Latih	Uji	Latih	Uji
Accuracy (%)	99.74	99.67	99.8	99.67	99.67	99.67
Precision (%)	99.47	99.32	99.8	99.3	99.67	99.21
Recall (%)	99.5	99.4	99.8	99.43	99.67	99.5
F1-score (%)	99.48	99.35	99.8	99.36	99.67	99.35
Waktu (s)	14.58		47.3		64.31	

Model *XGBoost*, SMOTE-*XGBoost*, dan ADASYN-*XGBoost* memberikan performa yang **sangat baik** dan **tidak memiliki perbedaan yang signifikan**

KESIMPULAN

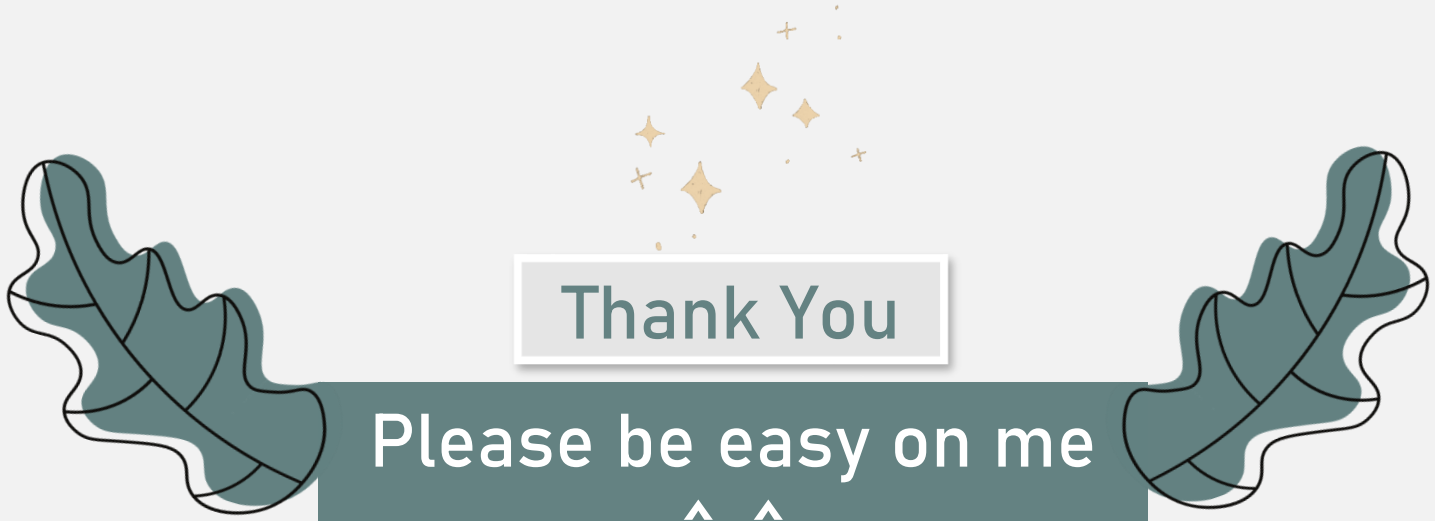
Kombinasi teknik *resampling* SMOTE dan ADASYN dengan *XGBoost* memberikan **performa yang baik** dalam memprediksi data dengan kelas yang tidak seimbang pada kasus tingkat risiko kredit. SMOTE dan ADASYN memberikan pengaruh dalam meningkatkan performa model meskipun **tidak signifikan**.

ADASYN-*XGBoost* sebagai model yang membangkitkan sampel sintetis secara adaptif memberikan performa waktu 4 kali lebih lama, sedangkan SMOTE-*XGBoost* menghasilkan waktu 3 kali lebih lama dari *XGBoost*. Di sisi lain, *XGBoost* sebagai algoritma *ensemble* memberikan kemampuan prediksi yang kuat dan pembelajaran yang sangat baik pada data dengan kelas yang seimbang maupun tidak seimbang. Hal ini menunjukkan bahwa model *XGBoost* merupakan **algoritma yang kuat** untuk klasifikasi. Selain itu, model ini dapat bekerja dengan baik pada data yang telah melalui proses pembersihan dan pemilihan fitur yang penting.

SARAN

Berdasarkan hasil penelitian ini, saran yang dapat diambil untuk penelitian selanjutnya adalah dengan **menambahkan iterasi** dan **optimalisasi *hyperparameter*** pada model untuk mendapatkan hasil prediksi yang lebih optimal.

Penambahan iterasi akan meningkatkan waktu pelatihan model dan memungkinkan model belajar lebih banyak dari data, terutama pada data yang besar dan kompleks. Selain itu, *XGBoost* memiliki sejumlah *hyperparameter* yang dapat diatur. Penelitian selanjutnya dapat fokus pada optimasi *hyperparameter* untuk masing-masing model dengan menggunakan metode optimasi dalam menemukan kombinasi *hyperparameter* yang optimal untuk meningkatkan performa model.



Thank You

Please be easy on me

^_^