Rakamin Academy X id/x partners

Credit Risk Loan Prediction

Loan Dataset 2004-2017

ID/X Partners - Data Scientist

Presented by :
Alda Fauziah Afifah

Alda Fauziah Afifah

An active student of S-1 Statistics Study Program of Universitas Airlangga who is dedicated and interested in exploring data. I am a detail-oriented and highly analytical Data Analyst experience in interpreting. My expertise lies in data visualization, statistical analysis, and the development of data-driven strategies that enhance operational efficiency and inform decision-making. My passion for uncovering actionable insights from data, coupled with my commitment to continuous learning, allows me to adapt quickly to new challenges and contribute effectively to team success. I am excited to leverage my skills and experience to help your organization achieve its goals through data-driven decision-making.

Surabaya, Indonesia

aldafauziahafifah@gmail.com

linkedin.com/in/aldafauziah

CERTIFICATION

Data Preprocessing and data Classification using Phyton

HIMPUNAN MAHASISWA STATISTIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA

SERTIFIKAT

017/A.3/ARSEN/HIMASTA-UNAIR/XI/2022

Diberikan Kepada :

Alda Fauziah Afifah

Atas partisipasinya sebagai :

PESERTA

Dalam kegiatan Workshop Olah Data (WOD) 2022 dengan tema
"Getting Closer with The Overview of Statistics for Achieving SDGs Targets in Digital Transformation Era"
yang diselenggarakan oleh Himpunan Mahasiswa Statistika Universitas Airlangga pada tanggal 13 November 2022

Dekan	Ketua HIMASTA	Ketua Pelaksana
Fakultas Sains dan Teknologi	Fakultas Sains dan Teknologi	Fakultas Sains dan Tek-
Universitas Airlangga	Universitas Airlangga	Universitas Airla
Prof. Dr. Moh Yasin, M.Si.	Chaerobby Fakhri F.P.	Antonio Nikolas Ma
NIP. 196703121991021001	NIM. 081911833037	NIM. 08191183

PENILAIAN KEGIATAN
WORKSHOP OLAH DATA 2022

No.	Materi	Kriteria	Durasi (Menit)	Nilai
1.	*Data Preprocessing Using Python*	Mengisi *Missing Value*	25	A
2.		Label *Encoder*	25	A
3.		Memilih Variabel Selection	30	A
4.	*Data Classification Analysis Using Python*	*Confusion Matrix*	30	A
5.		Akurasi	25	A
6.		Interpretasi	25	A
Total Durasi dan Predikat			160	A

Penilaian dilakukan secara objektif oleh tim penilai yang telah berkoordinasi dengan pemateri
A : Sangat Menguasai
B : Menguasai
C: Cukup Menguasai

Pemateri 1	Pemateri 2
Elly Pusporani, S.Si., M.Stat.	Sa'idah Zahrotul Jannah, S.Si., M.St
NIP. 199403242020122010	NIP. 199509182020122014

Building a business Presence with Facebook Marketing

coursera project network

PROJECT CERTIFICATE

Nov 25, 2022

ALDA FAUZIAH AFIFAH

has successfully completed

Building a Business Presence With Facebook Marketing

an online non-credit project authorized by Coursera Project Network and offered through Coursera

Kathleen Bergner
Subject Matter Expert

coursera

Introduction to Microsoft Excel

coursera project network

PROJECT CERTIFICATE

Dec 3, 2022

ALDA FAUZIAH AFIFAH

has successfully completed

Introduction to Microsoft Excel

an online non-credit project authorized by Coursera Project Network and offered through Coursera

Sommer Scaggs
Subject Matter Expert
Freedom Learning Group

coursera

Hit on Statistical Software : Discovering True Meaning of Data

Tanggal : 11 Agustus 2023

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS DIPONEGORO
FAKULTAS SAINS DAN MATEMATIKA
HIMPUNAN MAHASISWA STATISTIKA

Sertifikat

Diberikan kepada :

ALDA FAUZIAH AFIFAH

Atas Partisipasinya Sebagai

PESERTA

pada kegiatan

Training Software Statistik 2023 dengan tema
"Hit On Statistical Software: Discovering True Meaning of Data"

Diselenggarakan pada tanggal 19 dan 20 Agustus 2023

Semarang, 20 Agustus 2023

Dekan Fakultas Sains dan Matematika	Ketua Himpunan Mahasiswa Statistika	Ketua Panitia
Universitas Diponegoro	Universitas Diponegoro	
Prof. Dr. Widowati, S.Si., M.Si.	Edmundus Olgih Galih Poetra	Diva Au
NIP 196902141994032002	NIM 24050121130090	NIM 2405

Get To Know My Portofolio :

bitly

id/x partners

About Company

Rakamin Academy

ID/X partners menyediakan layanan konsultasi yang berspesialisasi dalam memanfaatkan solusi data analitik dan pengambilan keputusan (DAD) yang dikombinasikan dengan manajemen risiko dan disiplin pemasaran yang terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis. **Dengan *culture* sebagai berikut :**

C **Customer First**
Memberikan solusi yang tepat & layanan unggul bagi pelanggan untuk mewujudkan nilai bisnis yang maksimal

H **Honorable**
Fokus pada pertumbuhan yang menguntungkan tanpa mengorbankan integritas & kualitas

A **Agile**
Menanggapi dengan cepat dan penuh arti terhadap peluang (atau ancaman) baru

M **Mentorship**
Mengembangkan pertumbuhan pribadi dan profesional tim yang berkelanjutan

P **Proactive**
Mengajukan ide-ide baru untuk meningkatkan pekerjaan kita

I **Innovative**
Menumbuhkan lingkungan kerja yang kreatif untuk memecahkan masalah yang sulit

O **Ownership**
Menjalankan akuntabilitas dan rasa memiliki terhadap pekerjaan dan tanggung jawab kita

N **Numeric**
Membuat keputusan berdasarkan fakta dan angka

Project Description

Rakamin Academy

Sebagai data scientist di ID/X Partners, terdapat sebuah proyek dari perusahaan pemberi pinjaman (multifinance), dimana clien ingin meningkatkan keakuratan dalam menilai dan mengelola risiko kredit, sehingga dapat mengoptimalkan keputusan bisnis dan mengurangi potensi kerugian. **Proyek ini dirancang untuk mengembangkan model machine learning yang dapat memprediksi risiko kredit** berdasarkan data yang diberikan yang mencakup data pinjaman yang disetujui dan ditolak.

Data Understanding
Mengidentifikasi dan mengeksplorasi awal mengenai struktur dataset dan pola umum data.

Data Preparation
Penanganan missing value, mengatasi outlier, melakukan encoding pada variabel kategorikal, melakukan scalling atau normalisasi, lalu membagi data menjadi train set dan test set.

Exploratory Data Analysis (EDA)
Melakukan visualisasi data dan analisis korelasi.

Data Modelling
Memilih model machine learning, melakukan pelatihan model pada training set, dan mengevaluasi kinerja model.

Evaluation

Rakamin Academy

Project Tools

Google Colab for Python

PySeek

More Information :

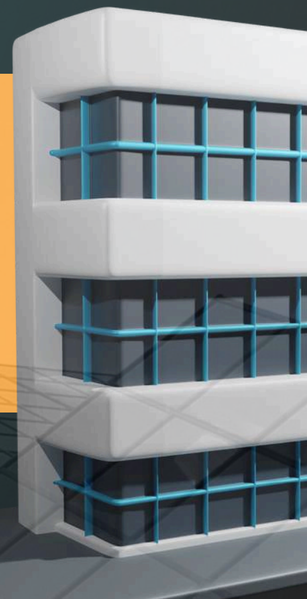
[Link Video Presentasi Project Based Internship]

[Link GitHub Alda Fauziah Afifah]

Introduction

Di era digital yang semakin berkembang, perusahaan pemberi pinjaman, khususnya di sektor multifinance, menghadapi tantangan besar dalam menilai dan mengelola risiko kredit. Risiko kredit (credit risk) merupakan ancaman signifikan yang dapat memengaruhi profitabilitas perusahaan. Risiko kredit yang tidak terkelola dengan baik dapat mengakibatkan kerugian besar akibat gagal bayar oleh peminjam. Dengan meningkatnya volume data historis dan kompleksitas dalam pengambilan keputusan, teknologi machine learning menjadi solusi yang efektif untuk memprediksi dan menganalisis risiko kredit secara akurat.

Model machine learning memungkinkan perusahaan untuk mengenali pola data yang relevan, mengevaluasi kelayakan peminjam, serta membuat keputusan yang berbasis data, sehingga dapat meningkatkan efisiensi operasional dan mengurangi potensi kerugian. Pendekatan ini tidak hanya memberikan manfaat finansial, tetapi juga membantu membangun ekosistem kredit yang lebih adil dan transparan.



Problem Statements

Bagaimana melakukan preprocessing dan eksplorasi data untuk pemodelan credit risk?

Bagaimana cara mengetahui variabel penting yang akan dipertimbangkan pada data set credit risk?

Bagaimana cara menangani imbalance class dalam dataset credit risk?

Bagaimana cara mengevaluasi kinerja tiap metode machine learning yang akan digunakan untuk pemodelan credit risk?

Apakah model machine learning yang memiliki kinerja terbaik untuk memprediksi credit risk?

Rakamin Academy

Solution Statements

Berdasarkan permasalahan tersebut, peneliti mengembangkan teknik SMOTE dalam penanganan imbalancing pada data dan menggunakan 7 algoritma yang berbeda untuk memprediksi model.

Data Understanding

Ringkasan dataset mengenai struktur dataset, mengidentifikasi setiap atribut, serta eksplorasi data awal mengenai distribusi variabel, statistik deskriptif, dan pola umum dalam data.

Informasi Dataset

Dataset yang digunakan merupakan loan dataset 2004-207 yang merupakan data pinjaman yang telah disediakan oleh Rakamin Academy. Data yang diberikan terdiri dari 466285 baris dengan 75 kolom.

Link Dataset :

CLICK HERE

Data Dictionary :

OPEN HERE

Rakamin Academy

Identifikasi tiap atribut

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Unnamed: 0                  466285 non-null  int64
 1   id                          466285 non-null  int64
 2   member_id                   466285 non-null  int64
 3   loan_amnt                   466285 non-null  int64
 4   funded_amnt                 466285 non-null  int64
 5   funded_amnt_inv             466285 non-null  float64
 6   term                        466285 non-null  object
 7   int_rate                    466285 non-null  float64
 8   installment                 466285 non-null  float64
 9   grade                       466285 non-null  object
 10  sub_grade                   466285 non-null  object
 11  emp_title                   438697 non-null  object
 12  emp_length                  445277 non-null  object
 13  home_ownership              466285 non-null  object
 14  annual_inc                  466281 non-null  float64
 15  verification_status         466285 non-null  object
 16  issue_d                     466285 non-null  object
 17  loan_status                 466285 non-null  object
 18  pymnt_plan                  466285 non-null  object
 19  url                         466285 non-null  object
 20  desc                        125981 non-null  object
 21  purpose                     466285 non-null  object
 22  title                       466264 non-null  object
 23  zip_code                    466285 non-null  object
```

Statistika Deskriptif

	Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	...	total_bal_il	il_util	open_rv_12m	open_rv_24m	max_bal_bc	all
count	466285.000000	4.662850e+05	4.662850e+05	466285.000000	466285.000000	466285.000000	466285.000000	466285.000000	4.662810e+05	466285.000000	...	0.0	0.0	0.0	0.0	0.0	
mean	233142.000000	1.307973e+07	1.459766e+07	14317.277577	14291.801044	14222.329888	13.829236	432.061201	7.327738e+04	17.218758	...	NaN	NaN	NaN	NaN	NaN	
std	134605.029472	1.089371e+07	1.168237e+07	8286.509164	8274.371300	8297.637788	4.357587	243.485550	5.496357e+04	7.851121	...	NaN	NaN	NaN	NaN	NaN	
min	0.000000	5.473400e+04	7.047300e+04	500.000000	500.000000	0.000000	5.420000	15.670000	1.896000e+03	0.000000	...	NaN	NaN	NaN	NaN	NaN	
25%	116571.000000	3.639987e+06	4.379705e+06	8000.000000	8000.000000	8000.000000	10.990000	256.690000	4.500000e+04	11.360000	...	NaN	NaN	NaN	NaN	NaN	
50%	233142.000000	1.010790e+07	1.194108e+07	12000.000000	12000.000000	12000.000000	13.660000	379.890000	6.300000e+04	16.870000	...	NaN	NaN	NaN	NaN	NaN	
75%	349713.000000	2.073121e+07	2.300154e+07	20000.000000	20000.000000	19950.000000	16.490000	566.580000	8.896000e+04	22.780000	...	NaN	NaN	NaN	NaN	NaN	
max	466284.000000	3.809811e+07	4.086083e+07	35000.000000	35000.000000	35000.000000	26.060000	1409.990000	7.500000e+06	39.990000	...	NaN	NaN	NaN	NaN	NaN	

Identifikasi kolom unique dan missing value

	Column	Dtype	null count	null perc.	unique count	unique sample
0	Unnamed: 0	int64	0	0.00	466285	[19007, 122761, 73832, 78360, 269854]
1	id	int64	0	0.00	466285	[12656021, 6375661, 103478, 9036897, 3291007]
2	member_id	int64	0	0.00	466285	[8549867, 30144045, 8327580, 10317706, 1773878]
3	loan_amnt	int64	0	0.00	1352	[9475, 24350, 22225, 10425, 3900]
4	funded_amnt	int64	0	0.00	1354	[7000, 12150, 28250, 23925, 19675]
...	...	...	...	...	...	...
70	all_util	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
71	total_rev_hi_lim	float64	70276	15.07	14612	[33821.0, 56071.0, 191300.0, 4714.0, 33748.0]
72	inq_fi	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
73	total_cu_tl	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
74	inq_last_12m	float64	466285	100.00	0	[nan, nan, nan, nan, nan]

Berdasarkan statistika deskriptif tersebut, pada data frame kolom yang memiliki nilai NaN atau kosong akan di hapus untuk meningkatkan kualitas data dan memastikan relevansi dalam analisis atau pemodelan.

Feature Selection

Identifikasi kolom tidak relevan

```
for column in df_clean.columns:
    value_counts = df_clean[column].value_counts()
    print(f"Value counts for {column}:\n{value_counts}\n")
```

```
Value counts for member_id:
member_id
1296599     1
28653081    1
28692177    1
28702376    1
28763241    1
           ..
4686866     1
4847180     1
4724047     1
4678105     1
11061576    1
Name: count, Length: 466285, dtype: int64

Value counts for loan_amnt:
loan_amnt
10000    33023
12000    25519
15000    23486
20000    22759
35000    16596
         ...
34250        1
33400        1
32150        1
34325        1
33175        1
Name: count, Length: 1352, dtype: int64
```

Penghapusan kolom tidak relevan

```
unused = ['policy_code', 'application_type', 'Unnamed: 0', 'id', 'member_id','issue_d', 'pymnt_
                    'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'co
                    'last_pymnt_d', 'last_pymnt_amnt', 'zip_code', 'title', 'emp_title','loan_st

drop_data = df_clean[unused]

df_clean.drop(columns=unused, axis=1, inplace=True)
```

Berdasarkan hasil identifikasi, maka akan dilakukan penghapusan variabel-variabel yang diasumsikan tidal relevan selama proses pelunasan pinjaman, seperti kolom id, member_id, serta variabel-variabel lain yang tidak relevan dengan karakteristik atau performa pinjaman bank.

Identifikasi tipe Variabel

Identifikasi Variabel Numerik

```
#Menentukan variabel numerik
numerik = [var for var in df_clean.columns if df_clean[var].dtype!='O']
print('Terdapat {} variabel numerik\n'.format(len(numerik)))
print('Berikut adalah yang termasuk variabel numerik :', numerik)
```

Terdapat 19 variabel numerik
Berikut adalah yang termasuk variabel numerik : ['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'collections_12_mths_ex_med', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim']

Rakamin Academy

Identifikasi Variabel Kategorik

```
#Menentukan variabel kategorik
kategorik = [var for var in df_clean.columns if df_clean[var].dtype=='O']
print('Terdapat {} variabel kategorik\n'.format(len(kategorik)))
print('Berikut adalah yang termasuk variabel kategorik :\n\n', kategorik)

#Label pada variabel kategorik
for var in kategorik:
    print(var, ' contains ', len(df_clean[var].unique()), ' labels')

#Deteksi missing value pada variabel kategorik
df_clean[kategorik].isnull().sum()
```

```
Terdapat 13 variabel kategorik

Berikut adalah yang termasuk variabel kategorik :

 ['term', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'verificatio
term  contains  2  labels
grade  contains  7  labels
sub_grade  contains  35  labels
emp_length  contains  12  labels
home_ownership  contains  6  labels
verification_status  contains  3  labels
url  contains  466285  labels
purpose  contains  14  labels
addr_state  contains  50  labels
earliest_cr_line  contains  665  labels
initial_list_status  contains  2  labels
last_credit_pull_d  contains  104  labels
loan_category  contains  2  labels
```

Exploratory Data Analysis

Eksplorasi data melalui visualisasi data menggunakan grafik dan plot serta menganalisis korelasi untuk mendapatkan wawasan yang lebih mengenai data.

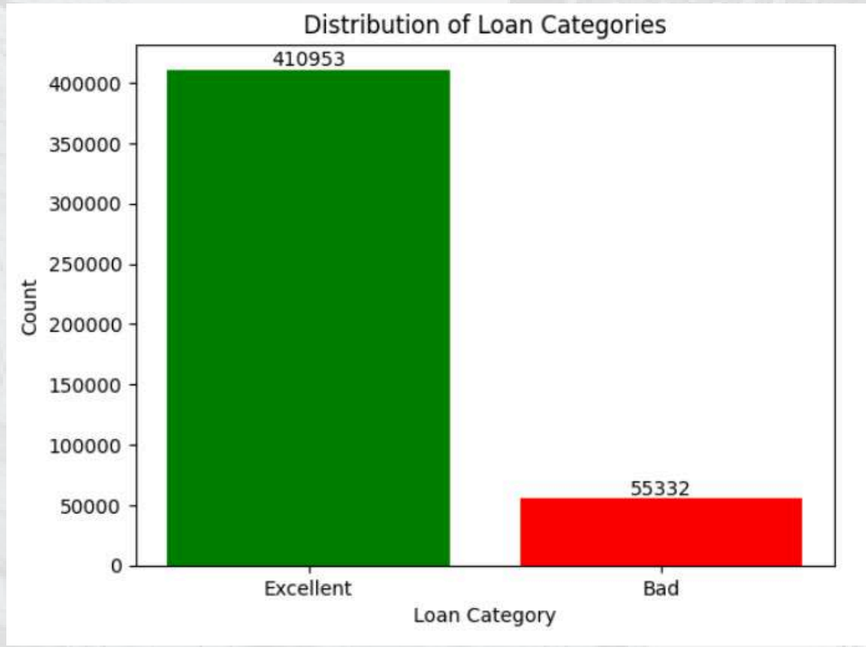
Pembagian Kolom Target

Identifikasi Kolom Loan Status

loan_status	count
Current	224226
Fully Paid	184739
Charged Off	42475
Late (31-120 days)	6900
In Grace Period	3146
Does not meet the credit policy. Status:Fully Paid	1988
Late (16-30 days)	1218
Default	832
Does not meet the credit policy. Status:Charged Off	761

Kita dapat mengklasifikasikan status kredit sebagai berikut :

- **Excellent Loans :** ["Fully Paid", "Does not meet the credit policy. Status:Fully Paid"]
- **Bad Loans :** ["Charged Off", "Late (31-120 days)", "Late (16-30 days)", "Default", "Does not meet the credit policy. Status:Charged Off"]



Personal Records Data

Salah satu poin menarik berdasarkan informasi data di atas adalah data yang terkait dengan catatan pribadi peminjam memiliki banyak nilai kosong (null). Secara khusus, ini mencakup data seperti jumlah bulan sejak terakhir kali terjadi keterlambatan pembayaran, jumlah bulan sejak catatan publik terakhir, dan jumlah bulan sejak pelanggaran besar terakhir.

	mths_since_last_delinq	mths_since_last_record	mths_since_last_major_derog
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	35.0	NaN	NaN
4	38.0	NaN	NaN
...	...	...	...
466280	NaN	NaN	NaN
466281	NaN	116.0	NaN

Data dengan nilai unik terkecil

	0
open_acc_6m	0
open_il_12m	0
open_il_6m	0
inq_last_12m	0
total_bal_il	0
verification_status_joint	0
dti_joint	0
annual_inc_joint	0
open_il_24m	0
mths_since_rcnt_il	0
il_util	0
open_rv_12m	0
open_rv_24m	0
max_bal_bc	0
all_util	0
inq_fi	0

inq_fi	0
total_cu_tl	0
policy_code	1
application_type	1
term	2
initial_list_status	2
pymnt_plan	2
loan_category	2
verification_status	3
home_ownership	6
acc_now_delinq	6
grade	7
collections_12_mths_ex_med	9
loan_status	9

Data Preparation

Penanganan missing value, mengatasi outlier, melakukan encoding pada variabel kategorikal, melakukan scalling atau normalisasi, lalu membagi data menjadi train set dan test set.

Penanganan Missing Value

Jumlah Missing Value tiap Kolom

	0
loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
installment	0
grade	0
sub_grade	0
emp_length	21008
home_ownership	0
annual_inc	4
verification_status	0
url	0
purpose	0
addr_state	0
dti	0
delinq_2yrs	29
earliest_cr_line	29
inq_last_6mths	29
open_acc	29
pub_rec	29
revol_bal	0
revol_util	340
total_acc	29
initial_list_status	0
last_credit_pull_d	42
collections_12_mths_ex_med	145
acc_now_delinq	29
tot_coll_amt	70276
tot_cur_bal	70276
total_rev_hi_lim	70276
loan_category	0

Penanganan missing value dilakukan dengan melakukan imputasi. Untuk kolom yang bertipa kategori akan di imputasi dengan menggunakan **modus**. Sedangkan kolom numerik akan di imputasi dengan menggunakan **median.**

Deteksi Duplikasi Data

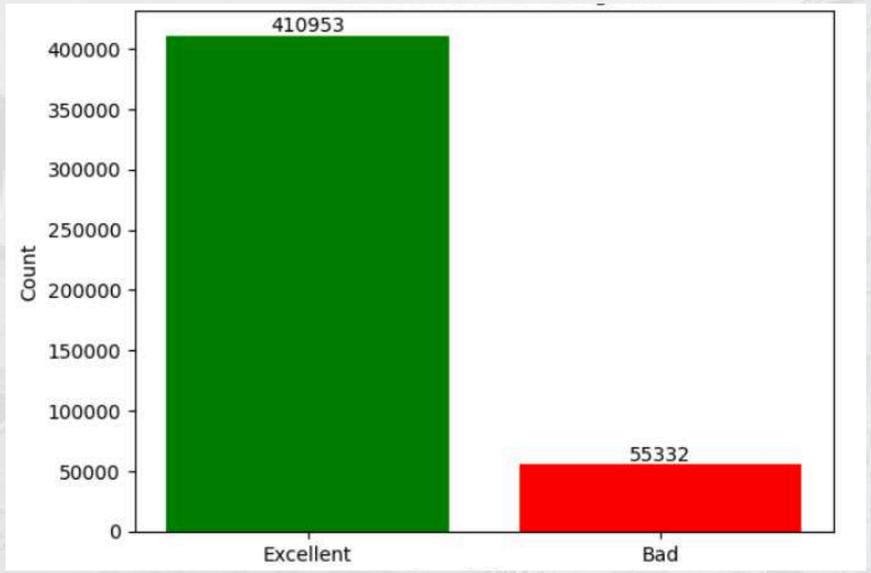
```
[ ]  df_clean.duplicated().sum()

0
```

Penanganan Imbalancing

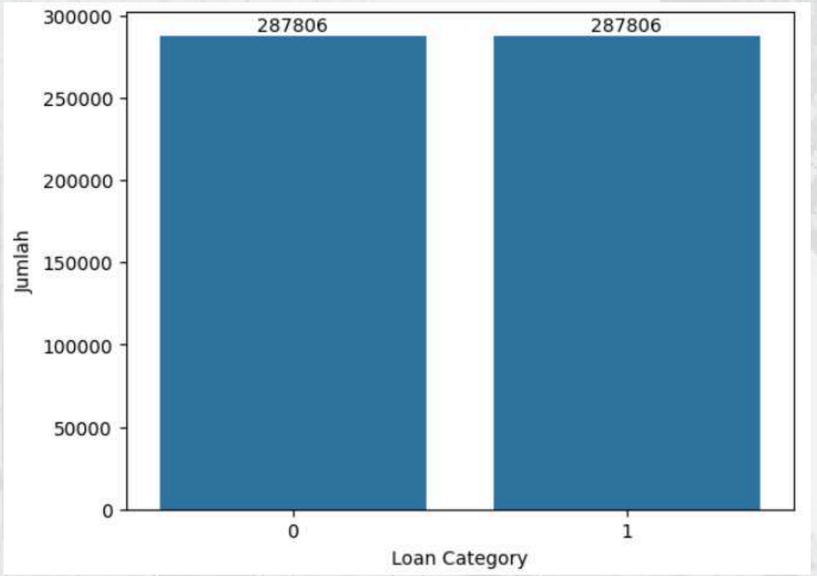
Rakamin Academy

Data tidak seimbang atau sering disebut dengan imbalanced data merupakan data yang proporsinya tidak seimbang antara satu kelas dengan kelas yang lain, sehingga kelas mayoritas (data yang banyak) dan kelas minoritas (data yang sedikit). Untuk mengatasi masalah ini, digunakan teknik oversampling dengan metode SMOTE (Synthetic Minority Oversampling Technique), yang secara efektik menyeimbangkan data dengan menghasilkan data sintetis pada kelas minoritas. Berikut adalah visualisasi data sebelum dan sesudah penangan imbalanced data dengan metode SMOTE pada data training.

Data sebelum Balancing



Data setelah Balancing



Encode Data Labels

```
encoded_verification = pd.get_dummies(df_clean2['verification_status'], prefix='verification', drop_first=True)
encoded_home_ownership = pd.get_dummies(df_clean2['home_ownership'], prefix='home', drop_first=True)
encoded_grade = pd.get_dummies(df_clean2['grade'], prefix='grade', drop_first=True)

# Combine the encoded features
encoded_categorical = pd.concat([encoded_verification,encoded_home_ownership, encoded_grade], axis=1)
df_clean2 = pd.concat([df_clean2, encoded_categorical], axis=1)

# Drop the original columns
df_clean2.drop(['verification_status','home_ownership', 'grade','term','emp_length'], axis=1, inplace=True)
```

```
df_clean2.head()
```

	loan_amnt	int_rate	annual_inc	delinq_2yrs	inq_last_6mths	open_acc	revol_util	collections_12_mths_ex_med	acc_now_delinq	loan_category
0	5000	10.65	24000.0	0.0	1.0	3.0	83.7	0.0	0.0	Excellent
1	2500	15.27	30000.0	0.0	5.0	3.0	9.4	0.0	0.0	Bad
2	2400	15.96	12252.0	0.0	2.0	2.0	98.5	0.0	0.0	Excellent
3	10000	13.49	49200.0	0.0	1.0	10.0	21.0	0.0	0.0	Excellent
4	3000	12.69	80000.0	0.0	0.0	15.0	53.9	0.0	0.0	Excellent

Mengubah kedua kategori pada loan_category, **excellent menjadi 1** dan **bad menjadi 0** untuk mempermudah proses klasifikasi dan prediksi.

Train Test Split

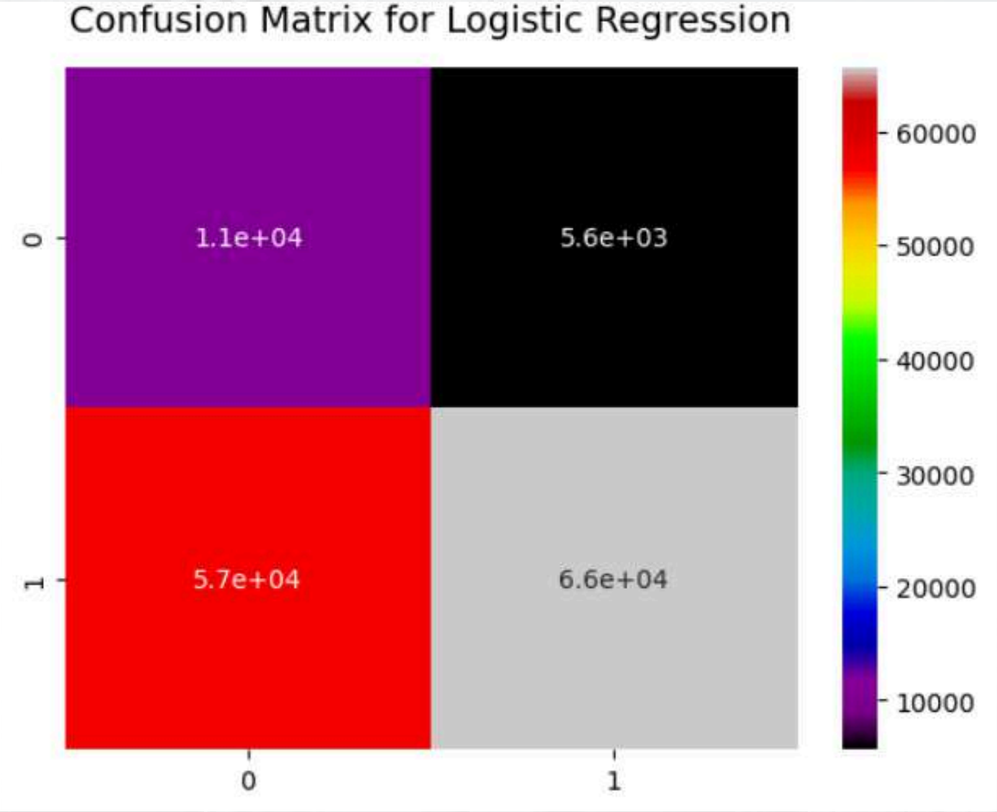
Setelah diperoleh penanganan imbalance pada data, maka langkah selanjutnya adalah membagi menjadi data training dan data testing. Dalam pemilihan proporsi data training dan data testing, pada penelitian ini dilakukan beberapa percobaan dengan membagi proporsi yaitu 90:10, 80:20, dan 70:30. Berdasarkan nilai akurasi, didapatkan bahwa proporsi **pembagian 70:30** menunjukkan akurasi tertinggi dibandingkan proporsi lainnya

Data Modeling

Memilih model machine learning, melakukan pelatihan model pada training set, dan mengevaluasi kinerja model.

Regresi Logistik

Regresi logistik adalah metode analisis statistik untuk memodelkan hubungan antara variabel dependen kualitatif dengan satu atau lebih variabel independen, yang dapat bersifat kategorikal atau kontinu

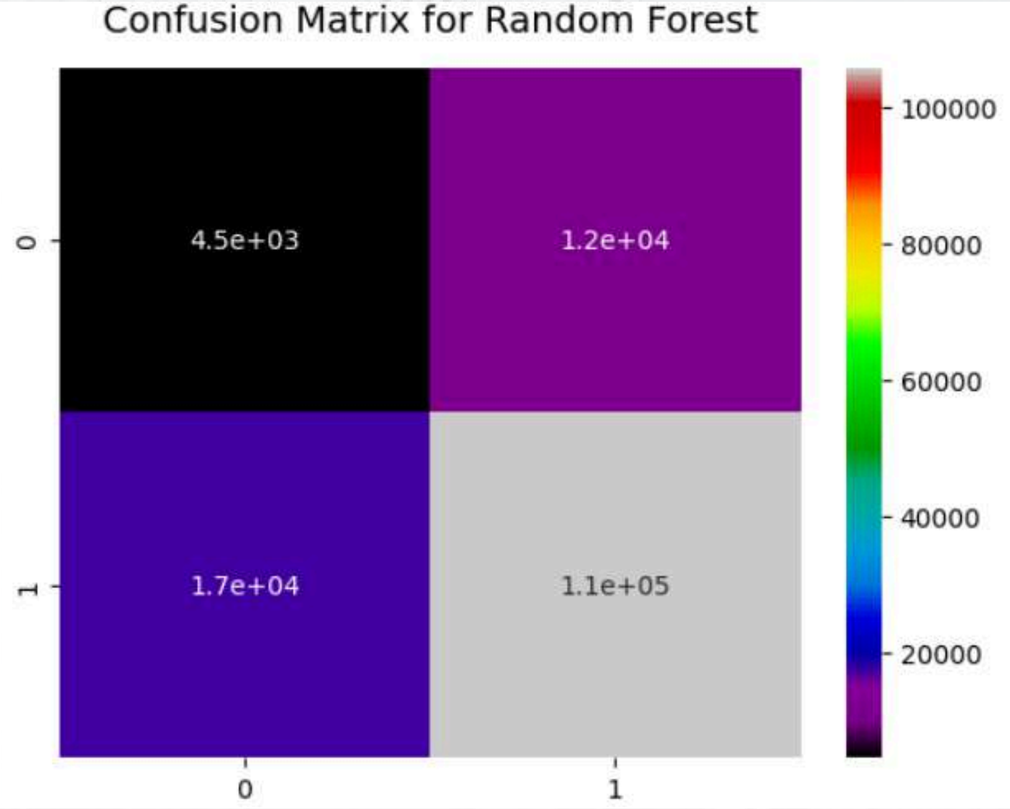


Accuracy Testing	54.89%
Accuracy Training	61.71%
Precision	92,1%
Recall	53,33%
AUC	63,6%

Hasil analisis regresi logistik menunjukkan performa model yang moderat, dengan accuracy testing 54.89% dan training 61.71% yang menandakan sedikit overfitting. Dari confusion matrix terlihat model berhasil mengidentifikasi sekitar 66,000 true positive dan 11,000 true negative, namun juga menghasilkan cukup banyak false positive dan false negative, yang menunjukkan masih ada ruang untuk peningkatan performa model terutama dalam hal mengurangi kesalahan klasifikasi.

Random Forest

Random Forest adalah metode atau algoritma yang digunakan untuk klasifikasi atau regresi. Metode ini merupakan sebuah kumpulan metode pembelajaran dengan menggunakandecision tree sebagai base classifier.

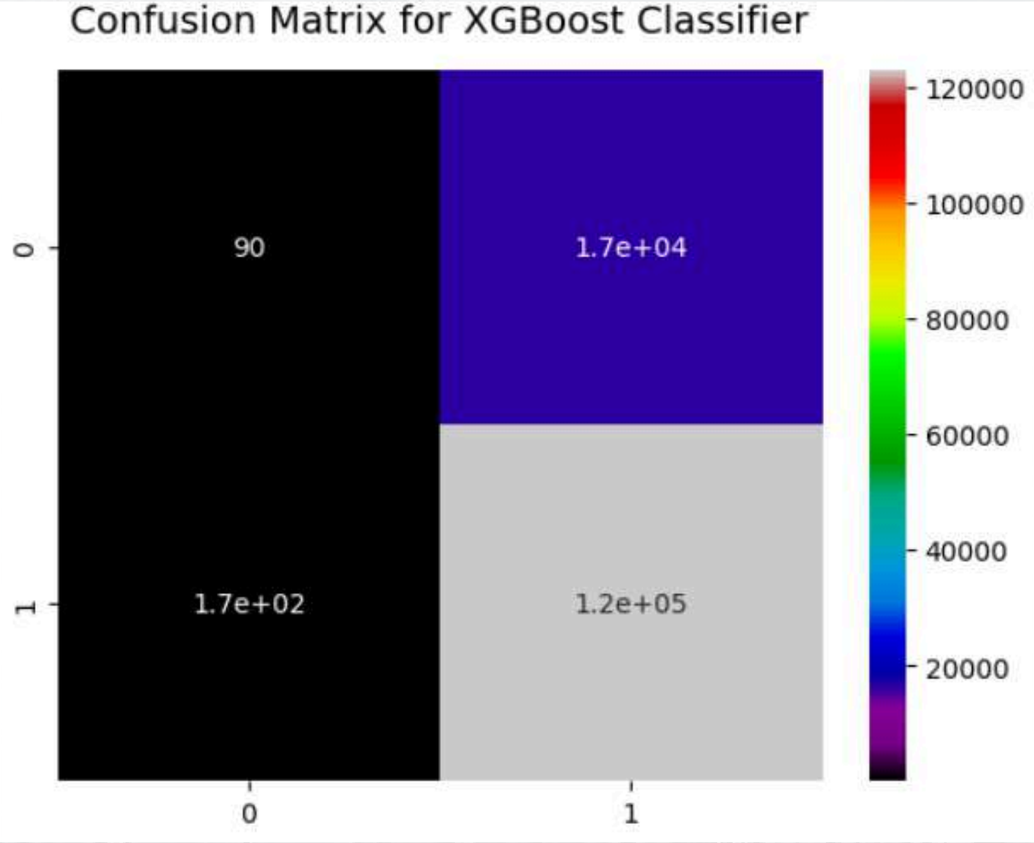


Accuracy Testing	78.78%
Accuracy Training	83.84%
Precision	89.62%
Recall	85.84%
AUC	64.62%

Model Random Forest menunjukkan performa yang lebih baik dibandingkan regresi logistik, dengan accuracy testing 78.78% dan training 83.84% yang menunjukkan model cukup stabil dengan overfitting yang minimal. Model ini memiliki keseimbangan yang baik antara precision (89.62%) dan recall (85.84%), mengindikasikan kemampuan yang baik dalam mengidentifikasi baik kasus positif maupun negatif.

XGBoost Classifier

XGBoost (Extreme Gradient Boosting) classifier adalah algoritma machine learning berbasis pohon keputusan yang sangat efisien. Algoritma ini menggunakan teknik boosting untuk menggabungkan banyak pohon keputusan secara iteratif, sehingga setiap pohon baru berusaha memperbaiki kesalahan dari pohon sebelumnya.

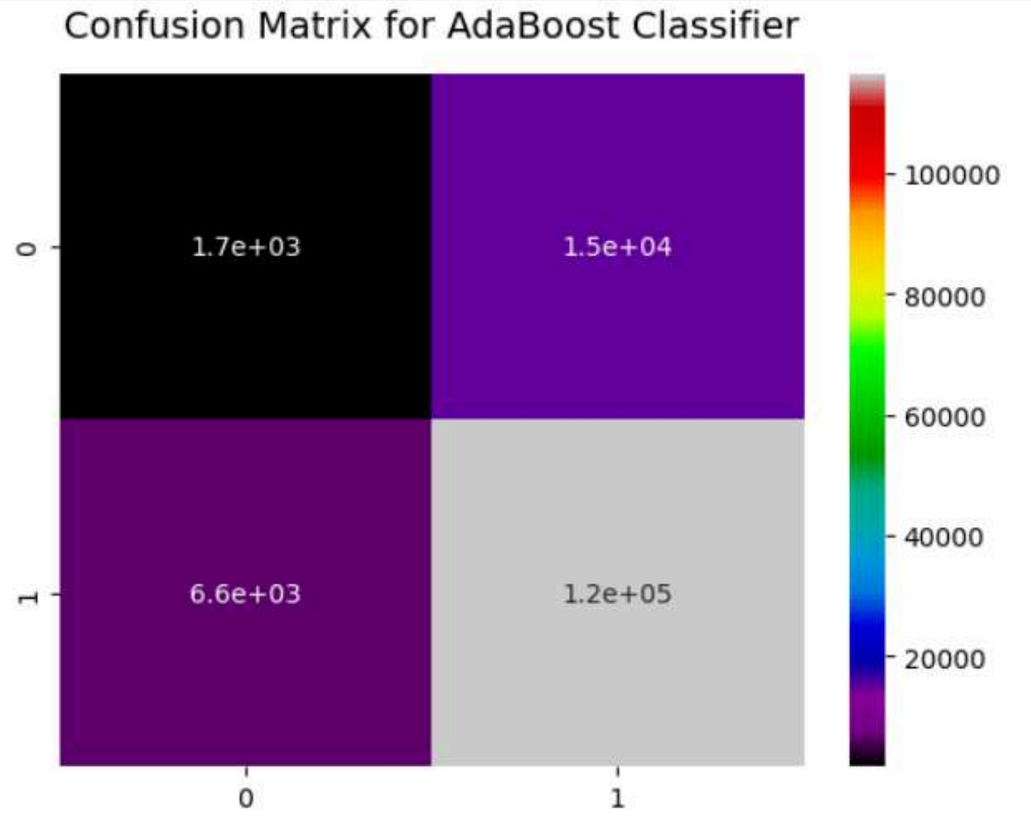


Accuracy Testing	87.97%
Precision	88.07%
Recall	99.86%

Model XGBoost menunjukkan performa yang paling unggul dibandingkan model sebelumnya dengan accuracy testing mencapai 87.97%, menandakan kemampuan prediksi yang sangat baik. Model ini memiliki precision 88.07% dan recall yang sangat tinggi yaitu 99.98%, yang berarti hampir sempurna dalam mengidentifikasi kasus positif yang sebenarnya.

ADABoost Classifier

Algoritma ini bekerja dengan menyesuaikan bobot data, di mana data yang salah diklasifikasikan diberi bobot lebih besar pada iterasi berikutnya, sehingga model berikutnya lebih fokus pada kesalahan tersebut. Proses ini berlanjut hingga jumlah iterasi tertentu atau hingga akurasi yang diinginkan tercapai.

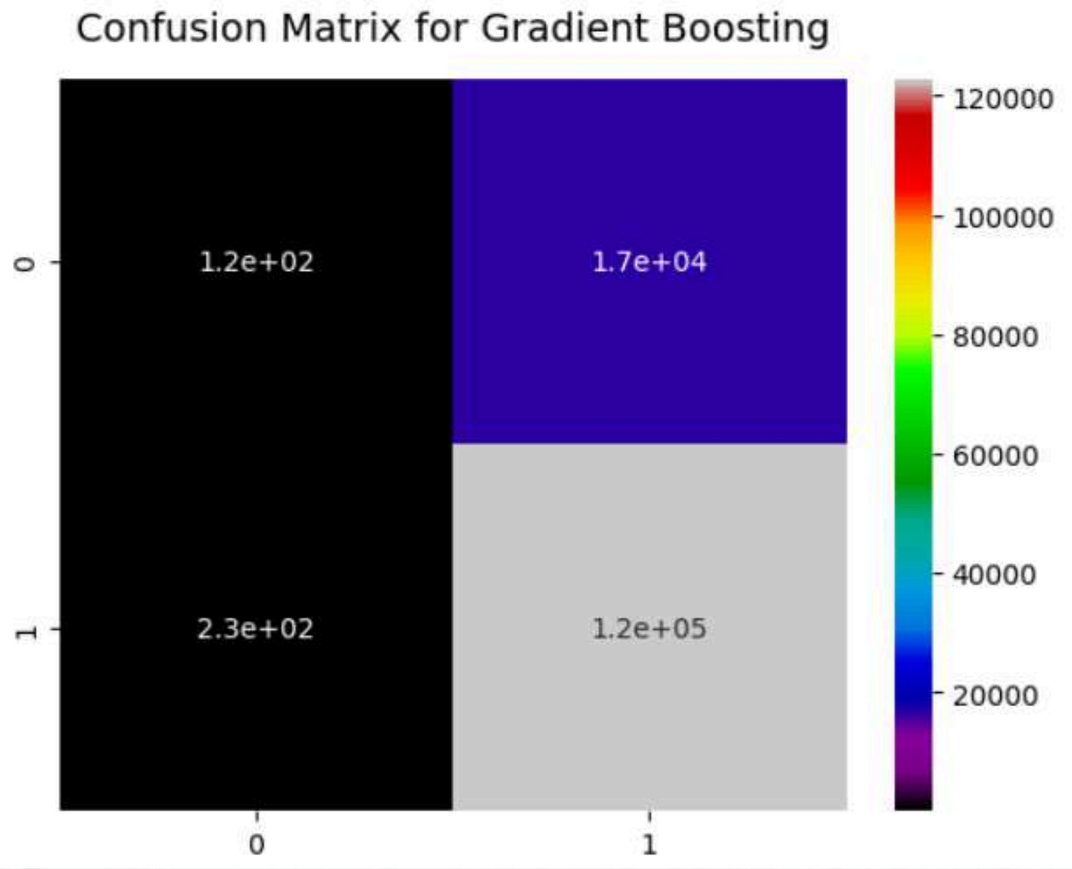


Accuracy Testing	84.52%
Precision	88.57%
Recall	94.63%

Model AdaBoost menunjukkan performa yang sangat baik dengan accuracy testing 84.52%, serta keseimbangan yang baik antara precision (88.57%) dan recall (94.63%). Dari confusion matrix, menunjukkan kemampuan model yang kuat dalam klasifikasi terutama untuk kasus positif meskipun tidak sebaik XGBoost dalam hal recall, namun masih lebih baik dibandingkan dengan model Random Forest dan Regresi Logistik.

Gradient Boosting Classifier

Gradient Boosting Classifier adalah algoritma ensemble yang menggabungkan banyak model lemah, seperti pohon keputusan, dengan mengoptimalkan kesalahan secara bertahap melalui pendekatan gradient descent untuk menghasilkan model klasifikasi yang kuat dan akurat.

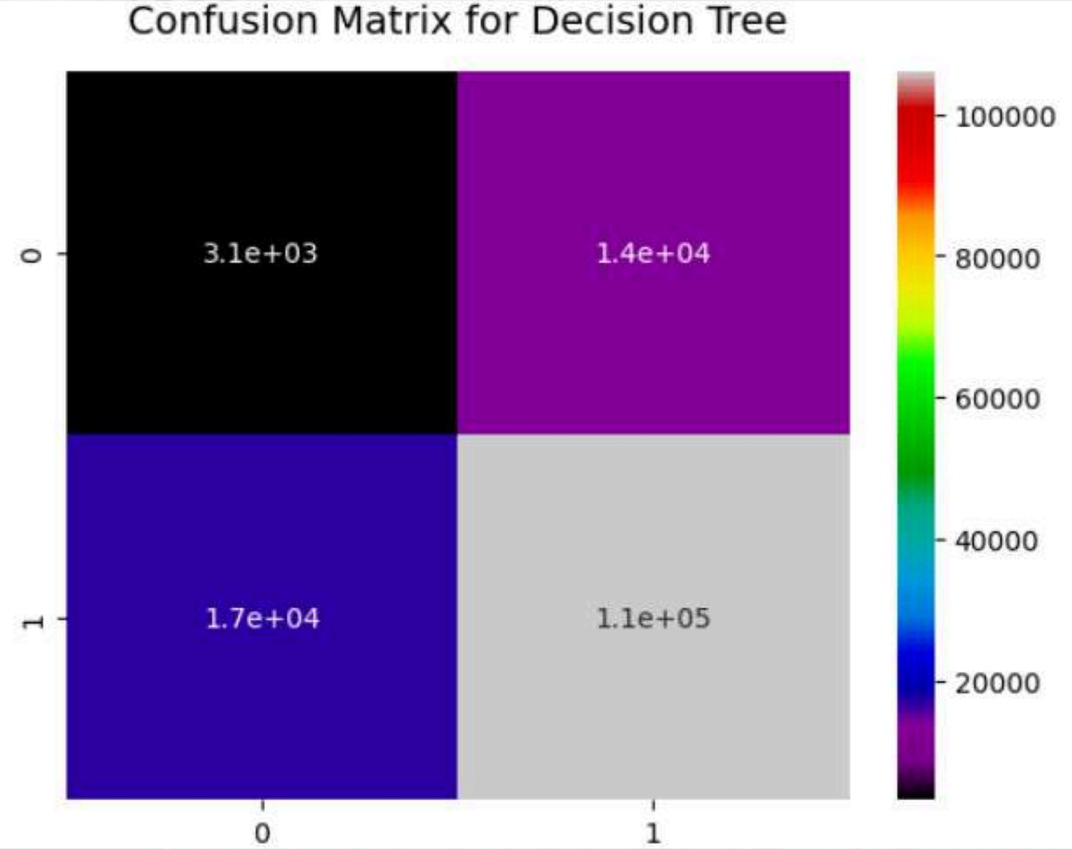


Accuracy Testing	87.95%
Precision	88.09%
Recall	99.81%

Model Gradient Boosting menunjukkan performa yang sangat baik dan hampir setara dengan XGBoost, dengan accuracy testing mencapai 87.95%. menunjukkan kemampuan yang hampir sempurna dalam mengenali kasus positif. Dari confusion matrix terlihat model berhasil mengidentifikasi 120,000 true positive dan hanya 126 true negative, dengan false positive sekitar 17,000 dan false negative yang sangat rendah yaitu 230, yang mengindikasikan model ini sangat efektif dalam klasifikasi terutama untuk kasus positif, mirip dengan performa XGBoost.

Decision Tree

Decision Tree adalah algoritma machine learning yang menggunakan struktur pohon untuk membuat keputusan berdasarkan fitur data, dengan setiap simpul mewakili tes pada fitur, cabang sebagai hasil tes, dan daun sebagai keputusan akhir atau klasifikasi.

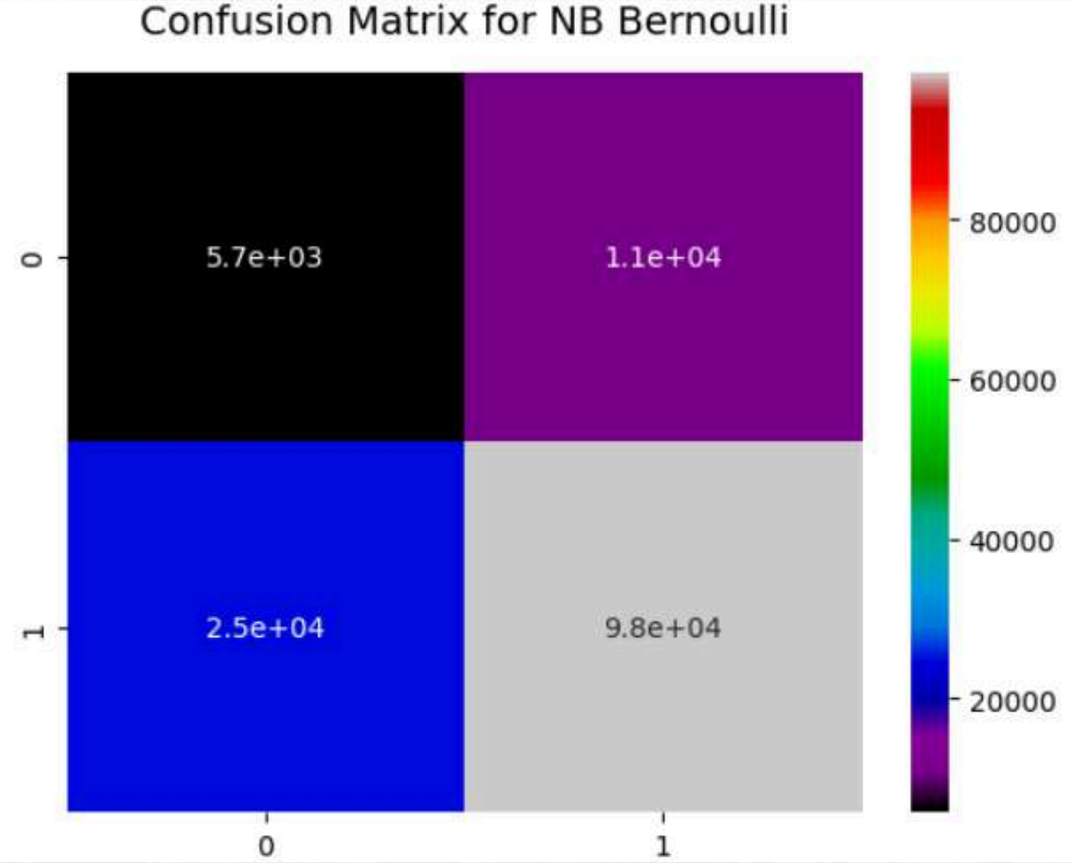


Accuracy Testing	78.10%
Precision	88.63%
Recall	86.18%

Model Decision Tree menunjukkan performa yang cukup baik dengan accuracy testing 78.10%, serta keseimbangan yang baik antara precision (88.63%) dan recall (86.18%). Dari confusion matrix terlihat model berhasil mengidentifikasi sekitar 110,000 true positive dan 3,100 true negative, dengan false positive sekitar 14,000 dan false negative 17,000, menunjukkan performa yang sebanding dengan Random Forest namun masih di bawah model-model boosting seperti XGBoost, Gradient Boosting, dan AdaBoost dalam hal accuracy dan recall.

Naive Bayes

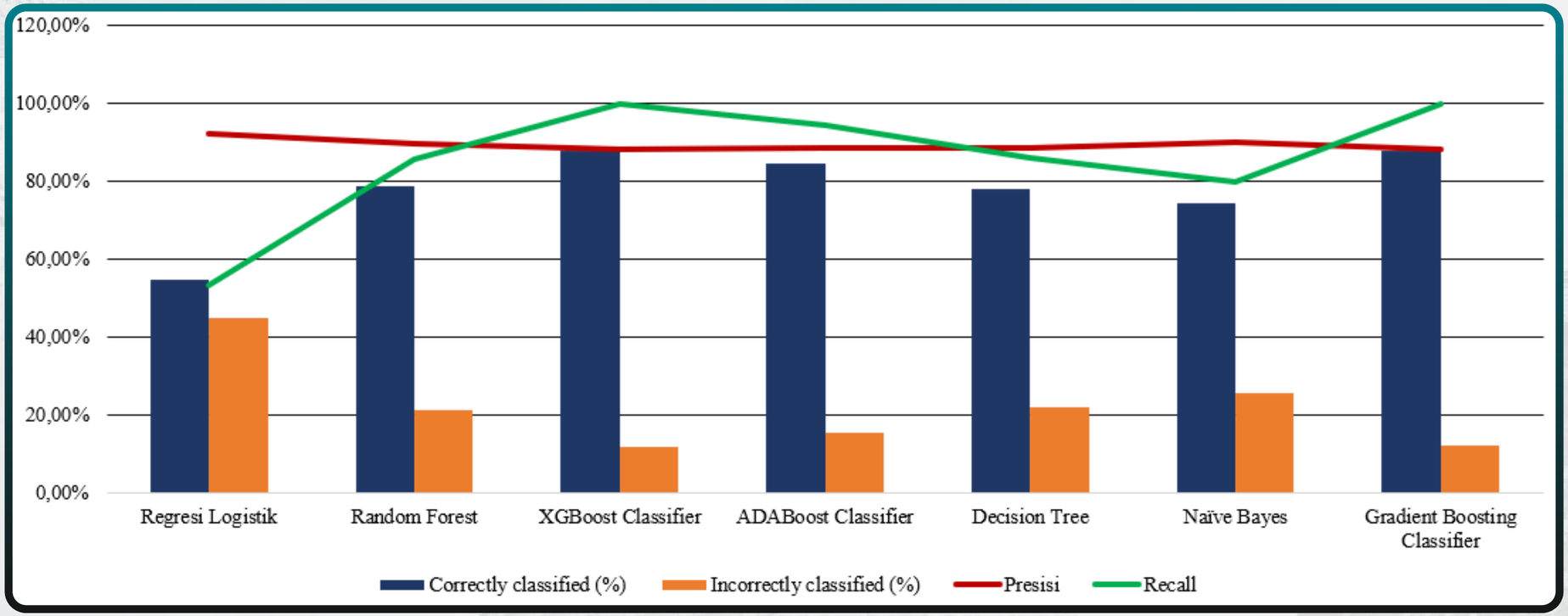
Naive Bayes adalah algoritma klasifikasi berbasis teorema Bayes yang mengasumsikan bahwa setiap fitur bersifat independen, meskipun dalam kenyataannya mungkin tidak, sehingga sederhana namun efektif untuk tugas klasifikasi teks dan data berlabel.



Accuracy Testing	74.46%
Precision	89.91%
Recall	79.97%

Model Naive Bayes (Bernoulli) menunjukkan performa yang moderat dengan accuracy testing 74.46%, dengan precision yang sangat baik (89.91%) namun recall yang lebih rendah (79.97%). Dari confusion matrix terlihat model berhasil mengidentifikasi sekitar 98,000 true positive dan 5,700 true negative, dengan false positive sekitar 11,000 dan false negative yang cukup tinggi yaitu 25,000, menunjukkan model ini memiliki performa yang lebih rendah dibandingkan model-model sebelumnya terutama dalam hal mengidentifikasi kasus positif yang sebenarnya.

Perbandingan Metode



XGBoost dan Gradient Boosting menunjukkan performa terbaik dengan tingkat correctly classified sekitar 88-90% dan incorrectly classified yang rendah. Kedua model ini juga menunjukkan keseimbangan yang sangat baik antara precision dan recall yang mencapai hampir 100%. AdaBoost menyusul di posisi ketiga dengan performa yang juga baik, diikuti oleh Random Forest dan Decision Tree yang menunjukkan performa moderat. Naive Bayes dan Regresi Logistik menunjukkan performa yang paling rendah di antara semua metode

Secara keseluruhan, metodeberbasis boosting (XGBoost, Gradient Boosting, AdaBoost) menunjukkan performa yang lebih unggul dibandingkan metode klasifikasi lainnya.

Kesimpulan dan Rekomendasi

Berdasarkan grafik, Gradient Boosting Classifier dan XGBoost Classifier menunjukkan performa terbaik dalam memprediksi risiko kredit, dengan akurasi, precision, dan recall mendekati 100%. Random Forest dan AdaBoost juga memberikan hasil yang baik, meskipun sedikit lebih rendah dibandingkan kedua model terbaik. Sebaliknya, Logistic Regression dan Naïve Bayes memiliki performa yang jauh lebih rendah, terutama dalam hal recall.

Untuk kasus loan credit risk, disarankan menggunakan Gradient Boosting atau XGBoost karena mampu secara efektif mendeteksi calon peminjam yang berisiko tinggi. Selain itu, perlu dilakukan optimasi hyperparameter untuk meningkatkan akurasi, serta evaluasi trade-off antara precision dan recall sesuai prioritas bisnis, misalnya, meminimalkan false negatives untuk mengurangi risiko kerugian. Monitoring berkala juga penting untuk memastikan model tetap efektif pada data baru.

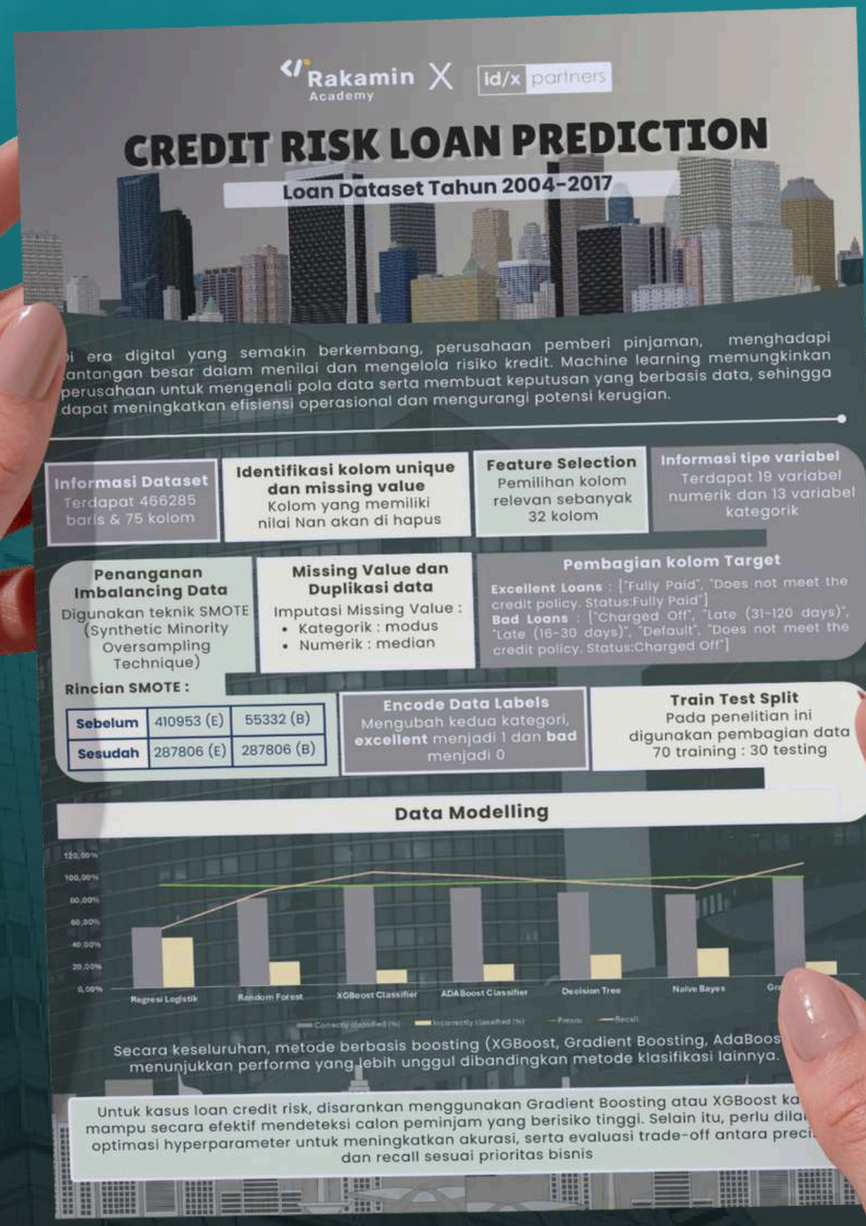
Infographic

For More Details :

bitly

CLICK HERE

Rakamin Academy X id/x partners



Thank You

Rakamin Academy X id/x partners