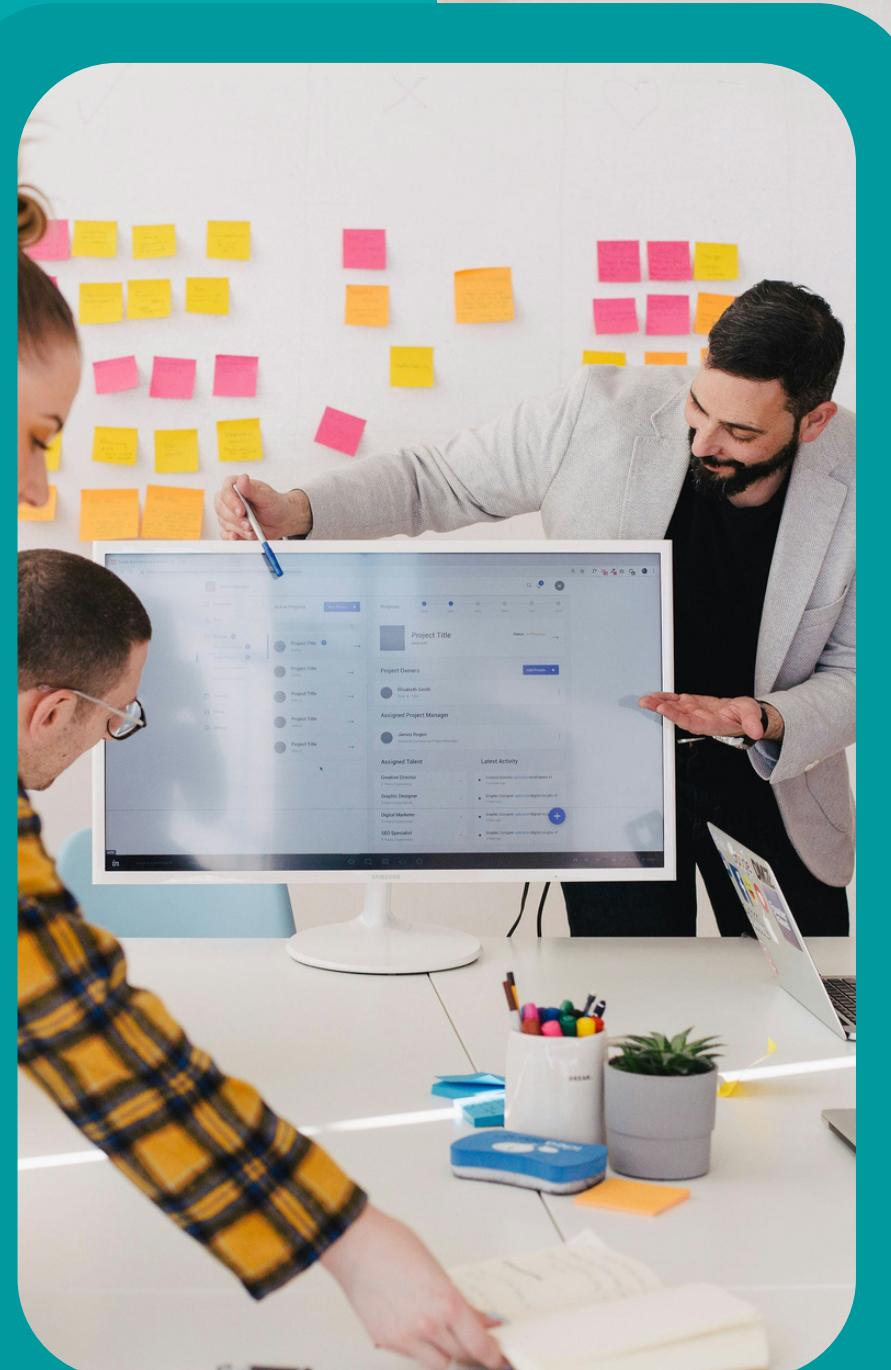


Credit Risk Modelling

Final Task of ID/X Partners Data Scientist Project Based Internship

March 2024 at Rakamin Academy

Developed by **Mochammad Adhi Buchori**



Final Task of ID/X Partners
Data Scientist Project Based Internship

This document contains complete information regarding Credit Risk Modelling.



Welcome to My Final Task Document

Hello, my name is

Mochammad Adhi Buchori

An Informatics student who is highly driven to achieve and learn new skills through practical experience. I am a highly disciplined, creative, and open-minded person who is interested in Information Technology, especially Data Science.

Interested in Data Analytics, Data Visualization, Data Science, and Machine Learning.



Final Task of ID/X Partners

Data Scientist Project Based Internship



Domain Proyek

1

Domain proyek yang diambil untuk proyek *machine learning* ini,
yaitu **Keuangan** dengan judul "**Credit Risk Modeling**".

Latar Belakang



Masalah

1 Kesulitan dalam Mengukur Risiko Kredit

Risiko kredit sulit diukur karena kompleksitas faktor-faktor yang mempengaruhinya, seperti kondisi keuangan peminjam, kualitas manajemen, kondisi industri, dan kondisi ekonomi secara keseluruhan.

2 Dampak Signifikan pada Kesehatan Finansial

Kegagalan peminjam dalam menyelesaikan kewajibannya terkait pembayaran pinjaman dapat menimbulkan konsekuensi finansial yang substansial bagi pemberi pinjaman atau kreditur.



Solusi

1 Pengembangan Model Prediksi

Pengembangan model *machine learning* yang dapat memprediksi risiko kredit untuk membantu perusahaan dalam melakukan evaluasi pinjaman, pengambilan keputusan, dan penetapan kebijakan.





Business Understanding

2

Pemahaman menyeluruh mengenai tujuan bisnis, kebutuhan pengguna, dan konteks pasar yang menjadi dasar bagi pengembangan solusi bisnis.

Business Understanding

Tahukah Kamu Apa itu Credit Risk?

Credit risk adalah probabilitas seorang peminjam akan mengalami kegagalan dalam membayar kembali jumlah pinjaman yang telah diberikan.

Wah, ... Prosesnya Gimana Tuh?

Proses pemberian pinjaman dilakukan berdasarkan analisis kemampuan bisnis atau individu untuk memenuhi kewajiban pembayaran di masa mendatang, termasuk pembayaran pokok dan bunga. Tentunya, proses ini penting banget untuk dilakukan! **Rugi dong** ... kalau sampai kita salah ngasih pinjaman ke klien.

Nah, Kamu Tahu Ga Sih?

Manajemen *credit risk* terdiri dari berbagai proses yang melibatkan beberapa langkah yang umumnya dikategorikan ke dalam 2 (dua) tahap utama, yaitu **measurement** (pengukuran) dan **mitigation** (mitigasi).

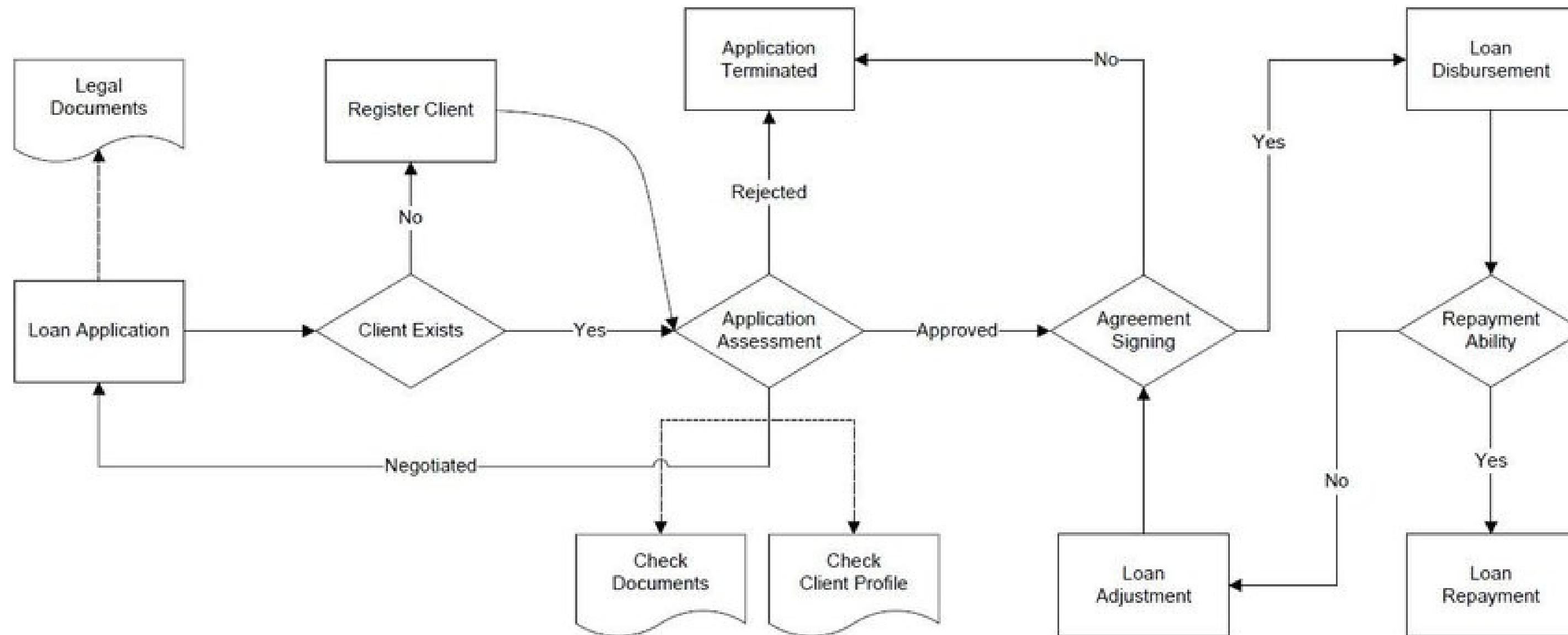
Nah, ... pengukuran ini melibatkan evaluasi keuangan dan profil peminjam untuk menilai risiko kredit, sedangkan mitigasi melibatkan penstrukturkan pinjaman dan pengendalian portofolio untuk mengurangi risiko kredit.

Masih Belum Paham, Kak :)

Tenang aja, aku bakal jelasin alurnya pelan-pelan. Yuk, **next slide!**

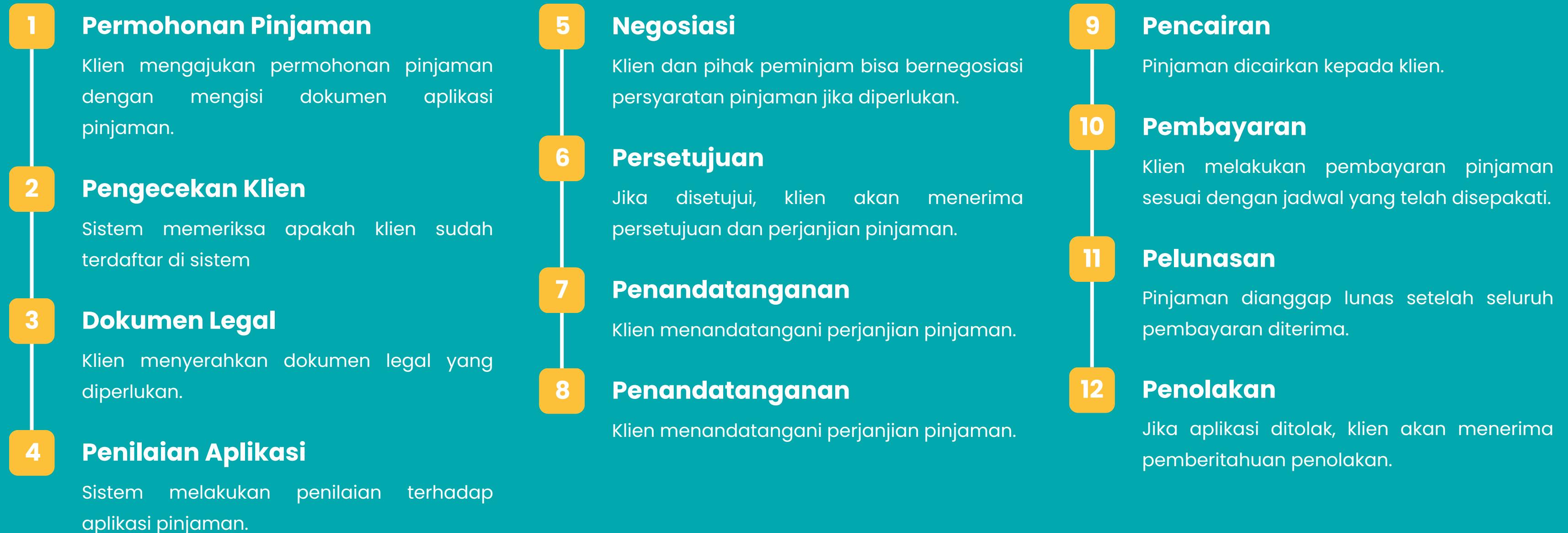
Yuk Simak Diagram Berikut!

Diagram Alur Proses Pengajuan Pinjaman



Simak penjelasannya di slide selanjutnya yaa!

Dapat diketahui bahwa **secara umum proses pengajuan pinjaman** terdiri dari tahapan berikut:

- 
- 1 Permohonan Pinjaman**
Klien mengajukan permohonan pinjaman dengan mengisi dokumen aplikasi pinjaman.
 - 2 Pengecekan Klien**
Sistem memeriksa apakah klien sudah terdaftar di sistem
 - 3 Dokumen Legal**
Klien menyerahkan dokumen legal yang diperlukan.
 - 4 Penilaian Aplikasi**
Sistem melakukan penilaian terhadap aplikasi pinjaman.
 - 5 Negosiasi**
Klien dan pihak peminjam bisa bernegosiasi persyaratan pinjaman jika diperlukan.
 - 6 Persetujuan**
Jika disetujui, klien akan menerima persetujuan dan perjanjian pinjaman.
 - 7 Penandatanganan**
Klien menandatangani perjanjian pinjaman.
 - 8 Penandatanganan**
Klien menandatangani perjanjian pinjaman.
 - 9 Pencairan**
Pinjaman dicairkan kepada klien.
 - 10 Pembayaran**
Klien melakukan pembayaran pinjaman sesuai dengan jadwal yang telah disepakati.
 - 11 Pelunasan**
Pinjaman dianggap lunas setelah seluruh pembayaran diterima.
 - 12 Penolakan**
Jika aplikasi ditolak, klien akan menerima pemberitahuan penolakan.



Problem Statements

- 1** Bagaimana pendekatan dalam membersihkan data dan *preprocessing* untuk pemodelan *credit risk*?
- 2** Apa variabel penting yang akan dipertimbangkan dalam dataset *credit risk* dan bagaimana cara menangani data yang hilang?
- 3** Bagaimana cara menangani ketidakseimbangan kelas dalam dataset saat memilih model *credit risk*?
- 4** Bagaimana cara membandingkan pro dan kontra dari model *machine learning* yang akan digunakan untuk pemodelan *credit risk*?
- 5** Apa model *machine learning* yang paling efektif untuk memprediksi *credit risk*?



Goals

- 1** Membangun model *machine learning* yang dapat digunakan untuk memprediksi *credit risk*.
- 2** Membandingkan beberapa algoritma guna memperoleh akurasi terbaik dalam melakukan prediksi terhadap *credit risk*.
- 3** Mengidentifikasi variabel yang paling efektif dalam menentukan *credit risk*.



Solution Statements

Guna mencapai tujuan, peneliti mengembangkan model prediktif dengan menggunakan 7 algoritma berbeda, meliputi Random Forest, Gradient Boosting, AdaBoost, XGBoost, Logistic Regression, K-Nearest Neighbors, dan Neural Network.



3

Data Loading

Proses mengimpor atau memuat data untuk
analisis atau pemrosesan lebih lanjut.

Data Loading

Tahap ini meliputi proses **mengimpor** atau **memuat data** untuk analisis atau pemrosesan lebih lanjut.

Tujuannya adalah untuk **membuat data yang diperlukan tersedia dalam format yang sesuai** untuk analisis atau pengolahan lebih lanjut.

Library yang digunakan meliputi library untuk manipulasi file dan direktori (os), pengelolaan Google Drive (google.colab), manipulasi data (pandas, numpy), visualisasi data (seaborn, matplotlib.pyplot), penanganan tanggal dan waktu (datetime), pra-pemrosesan data (sklearn.preprocessing, sklearn.compose), pembagian data (sklearn.model_selection), model-model machine learning (sklearn.ensemble, sklearn.linear_model, sklearn.neighbors, xgboost, tensorflow), serta evaluasi model (sklearn.metrics).

Proses yang dilakukan pada tahap ini meliputi:

- 1 **Mengimpor Library yang dibutuhkan**
- 2 **Menghubungkan Google Drive ke Google Colab**
- 3 **Memuat Dataset**



Important Notes



4

Data Understanding

Proses mengimpor atau memuat data untuk
analisis atau pemrosesan lebih lanjut.

Data Understanding

Informasi Dataset

Dataset yang digunakan dalam proyek ini, yaitu **data pinjaman** yang disediakan oleh Rakamin Academy sebagai bagian dari program Project Based Internship.

Jumlah Dataset

Data tersebut terdiri dari **466.285 entri** dengan **75 kolom**.

Link Dataset



atau download melalui
Source : [**Click Here**](#)

Proses yang dilakukan pada tahap ini meliputi:

- 1** Menampilkan **informasi** terperinci tentang **struktur data** pada DataFrame.
- 2** Menghitung **jumlah nilai NaN** dalam setiap kolom.
- 3** Menghapus **kolom indeks** dan kolom yang hanya berisi nilai NaN.

Important Inferences

Pada DataFrame, kolom indeks dihapus karena tidak relevan untuk analisis atau pembuatan model. Selain itu, kolom dengan nilai NaN atau kosong juga dihapus untuk meningkatkan kualitas data dan memastikan relevansi dalam analisis atau pemodelan.

Pada tahap ini, kamus data disusun untuk membantu memahami konteks setiap variabel dalam *loan dataset*.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Penyusunan Data Dictionary
- 2 Identifikasi variabel untuk menghapus fitur yang tidak relevan dengan kebutuhan pemodelan
- 3 Menghapus variabel yang tidak relevan.

Link Data Dictionary



atau akses melalui
[Source : Click Here](#)

Important Notes

```
irrelevant_features = [  
    'id', 'member_id', 'issue_d', 'url', 'delinq_2yrs', 'inq_last_6mths',  
    'mths_since_last_delinq', 'mths_since_last_record', 'out_prncp',  
    'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp',  
    'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee',  
    'last_pymnt_d', 'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d',  
    'collections_12_mths_ex_med', 'mths_since_last_major_derog', 'tot_coll_amt',  
    'tot_cur_bal', 'total_rev_hi_lim'  
]  
  
df = df.drop(irrelevant_features, axis=1)
```

Dilakukan penghapusan variabel-variabel yang diasumsikan diperoleh selama proses pelunasan pinjaman dan tidak relevan dengan fokus analisis, seperti **id**, **member_id**, **url**, serta variabel-variabel lain yang tidak memberikan informasi yang berguna tentang karakteristik atau performa pinjaman.



Exploratory Data Analysis

5

Proses investigasi awal yang dilakukan pada dataset untuk memahami dan menganalisis karakteristik utama dalam dataset.

Deskripsi Variabel

Tahap ini merujuk pada proses analisis yang bertujuan untuk memahami struktur, karakteristik, dan informasi yang terkandung dalam variabel-variabel yang digunakan dalam suatu dataset.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Menampilkan informasi terperinci tentang struktur data pada DataFrame.
- 2 Menghasilkan ringkasan statistik deskriptif dari dataset.
- 3 Melihat informasi tentang kolom-kolom yang memiliki tipe data objek (string) dan tipe data numerik.
- 4 Menghitung jumlah nilai unik dalam setiap kolom yang memiliki tipe data objek (string) dan tipe data numerik.

Important Inferences

Setelah melakukan analisis variabel, informasi yang diperoleh meliputi jumlah kolom dan baris, tipe data yang digunakan, keberadaan missing value, variasi data (baik kategorikal maupun numerik), serta statistik deskriptif untuk kolom-kolom numerik dalam dataset.

Setelah menghapus kolom indeks, menghapus kolom yang hanya berisi nilai NaN atau kosong, dan menghapus *irrelevant feature*, dataset terdiri dari **31 kolom** yang terdiri dari 17 kolom dengan tipe data object, 4 kolom numerik dengan tipe data integer, dan 10 kolom numerik dengan tipe data float.

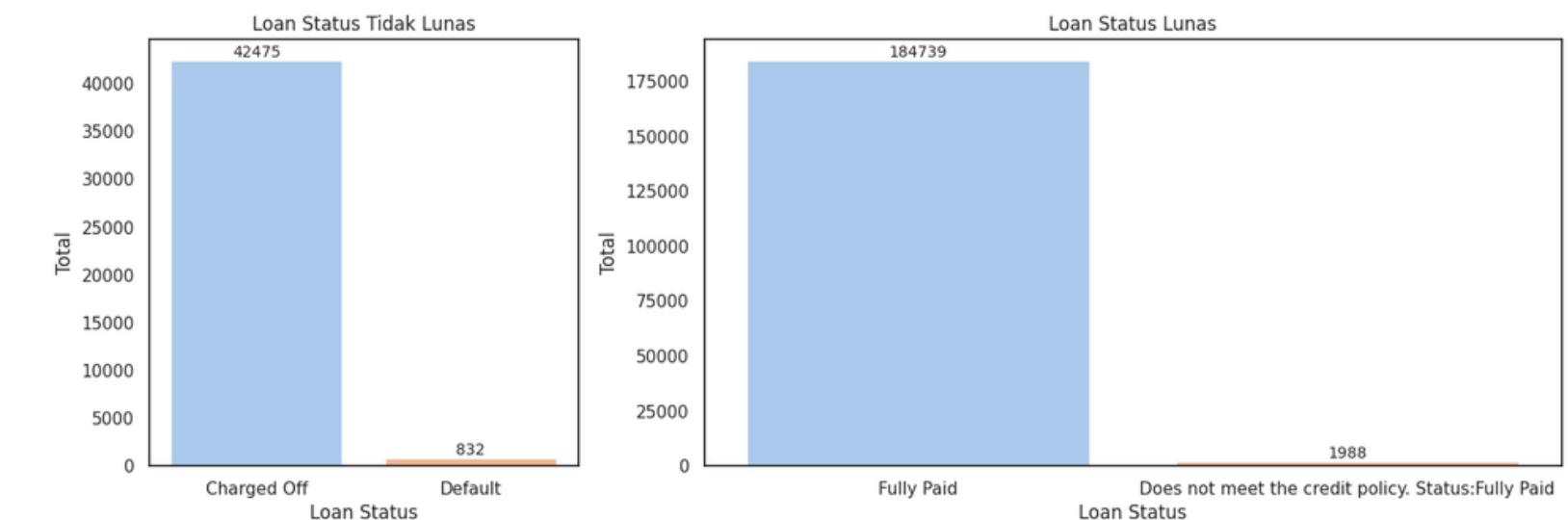
Feature Engineering

Pada tahap ini, **fitur-fitur baru diciptakan dari data yang sudah ada** untuk meningkatkan performa model machine learning.

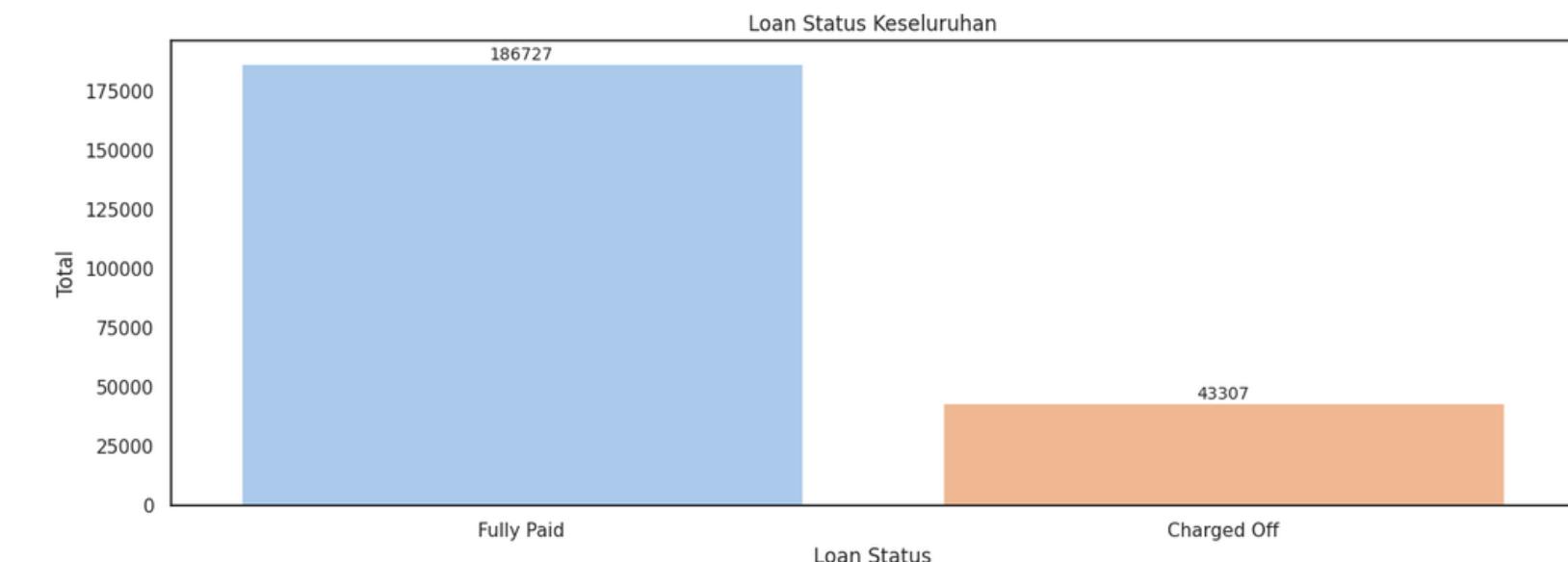
Proses yang dilakukan pada tahap ini meliputi:

- 1** Mendapatkan nilai unik dari setiap kolom dalam DataFrame.
- 2** Melakukan analisis terhadap isi kolom-kolom yang bersifat unik.
- 3** Memilih kolom **loan_status** karena berisikan informasi mengenai status peminjaman.
- 4** Menghitung frekuensi kemunculan setiap nilai unik dalam kolom **loan_status**.
- 5** Menggabungkan kelas **Fully Paid** dan **Does not meet the credit policy**. **Status:Fully Paid** menjadi satu kategori yang **positif (Lunas)** dengan nama kelas **Fully Paid**, serta menggabungkan kelas **Charged Off** dan **Default** menjadi satu kategori yang **negatif (Tidak Lunas)** dengan nama kelas **Charged Off**. Lalu, menghapus fitur yang tidak digunakan.

Hasil Plotting Jumlah Pinjaman untuk **Status Lunas dan Tidak Lunas**.



Hasil Plotting Jumlah Pinjaman untuk **Status Keseluruhan**.



Univariate Analysis

Data Analysis

Pada tahap ini, dilakukan analisis univariat untuk memahami karakteristik dari satu variabel tunggal dalam dataset. Tujuan utama analisis ini adalah merangkum data, mengidentifikasi distribusi, pola, dan sifat-sifat statistik dari variabel tersebut.

Categorical Features

Proses yang dilakukan pada tahap ini meliputi analisis, penghapusan, hingga penyesuaian format data terhadap fitur yang memiliki tipe data kategorikal (*object*) yang terdiri dari:

application_type, emp_title, earliest_cr_line

- 1
- 2
- 3

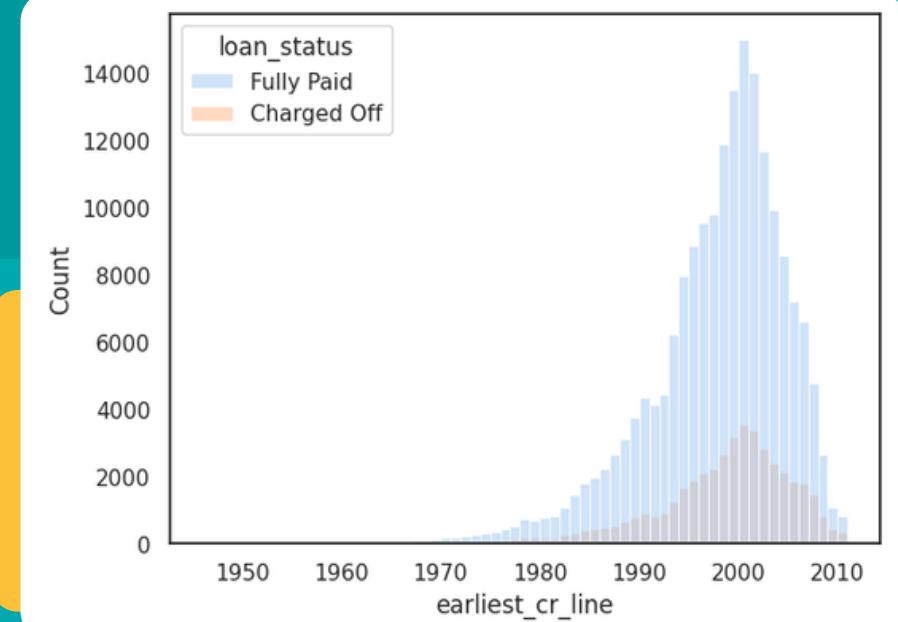
Fitur **application_type** dihapus karena **hanya memiliki 1 (satu) nilai unik**.

Fitur **emp_title** dihapus karena **memiliki terlalu banyak nilai unik**, yakni sebesar **12.867**.

Data pada fitur **earliest_cr_line** diformat ke dalam format **datetime**, disesuaikan **agar tidak melebihi tahun 2030**, dan kemudian **diubah menjadi format tahun**.

Important Inferences

Berikut merupakan hasil plotting distribusi data pada fitur **earliest_cr_line**

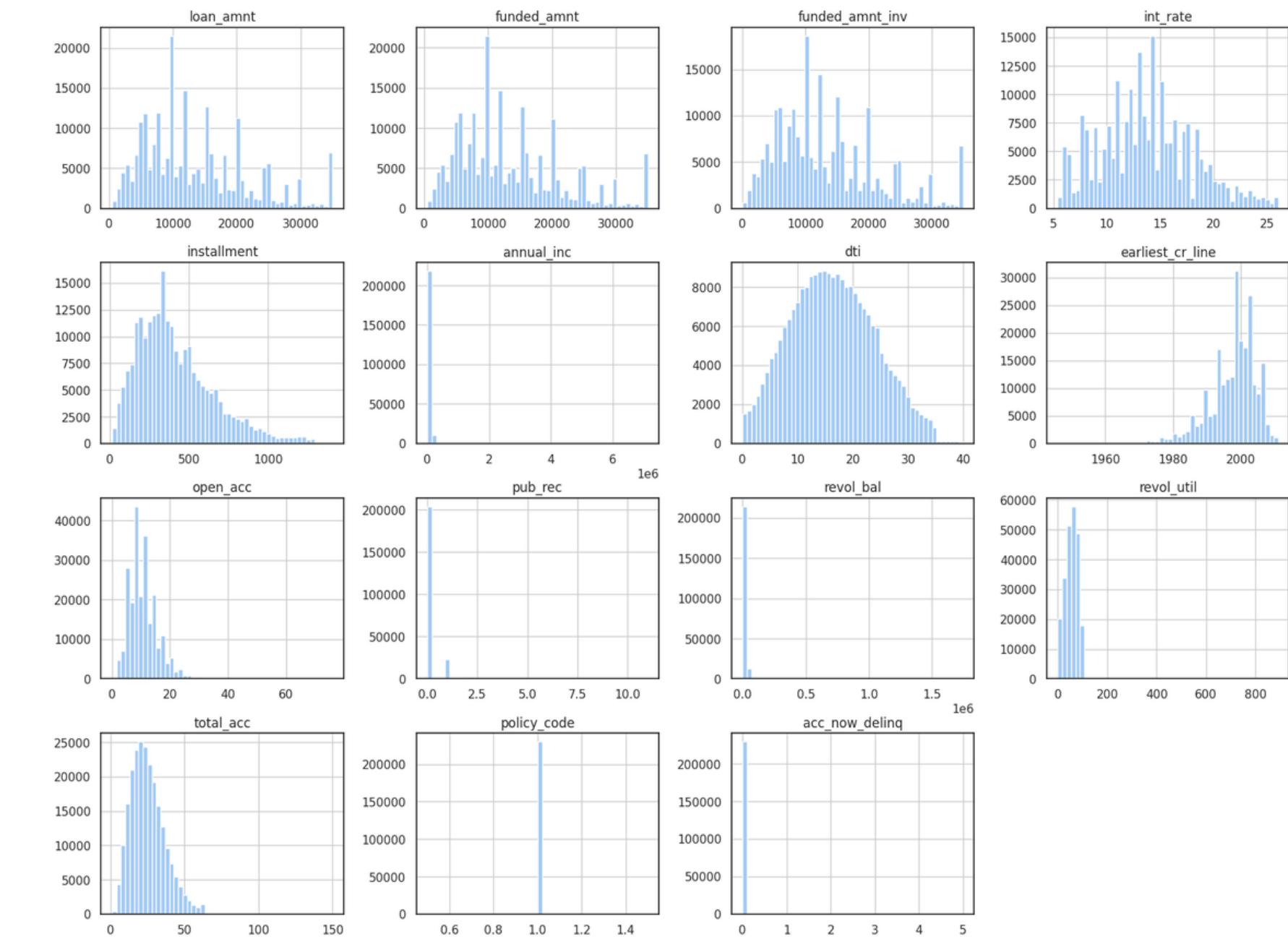


Numerical Features

Proses yang dilakukan pada tahap ini meliputi analisis menyeleruh terhadap fitur yang memiliki tipe data numerik dan berfokus pada fitur **loan_amnt** karena berisikan informasi tentang perubahan dalam jumlah pinjaman.

Important Inferences

- 1 Rata-rata pinjaman adalah Rp14.317.277,57 dengan standar deviasi Rp. 8.286.509,16.
- 2 Sebagian besar pinjaman berada di antara Rp8.000.000 dan Rp20.000.000.
- 3 Nilai maksimum jumlah pinjaman, yakni sebesar Rp35.000.000, sedangkan nilai minimum jumlah pinjaman, yakni sebesar Rp500.000.
- 4 Distribusi pinjaman tidak simetris.



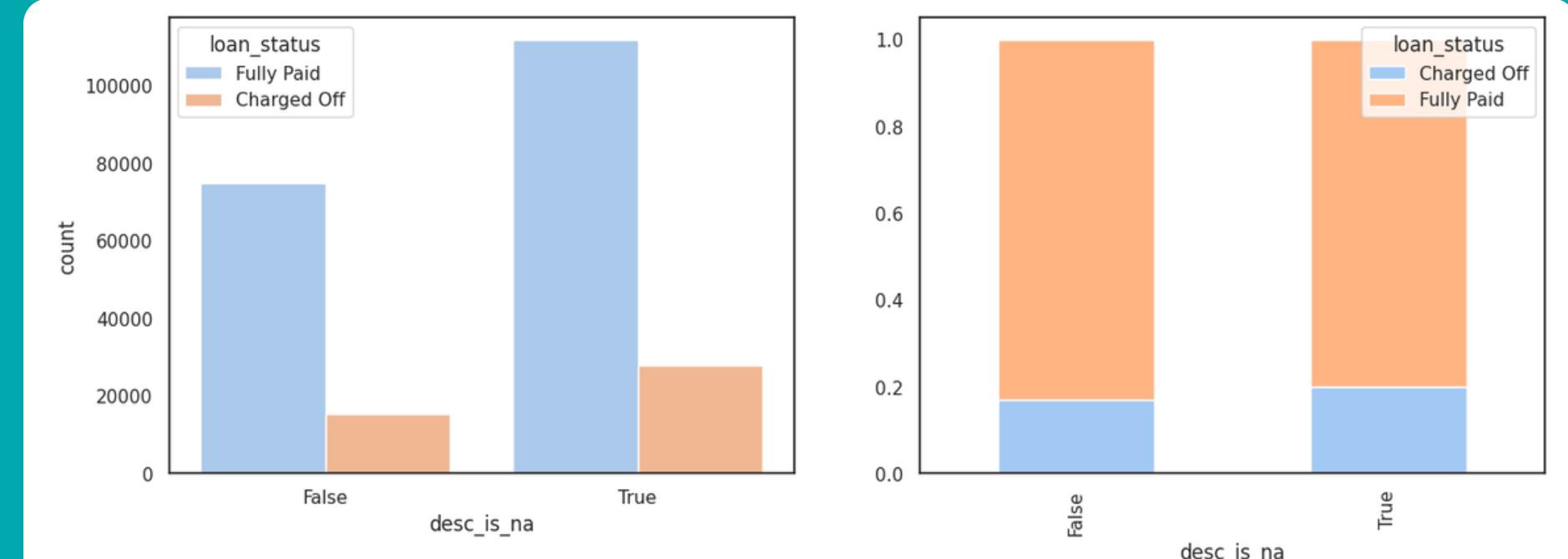
Pada tahap ini, dilakukan analisis multivariat untuk memahami hubungan antara dua atau lebih variabel dalam sebuah dataset. Analisis ini memungkinkan eksplorasi korelasi, pola, dan struktur kompleks antara variabel-variabel tersebut.

Categorical Features

Proses yang dilakukan pada tahap ini meliputi analisis mengenai hubungan antara dua variabel atau lebih dan penghapusan fitur yang memiliki tipe data kategorikal (*object*) yang terdiri dari:

desc, purpose, and title **zip_code dan addr_state**
grade and sub_grade

Berikut merupakan hasil *plotting* analisis pada fitur **desc, purpose, and title**



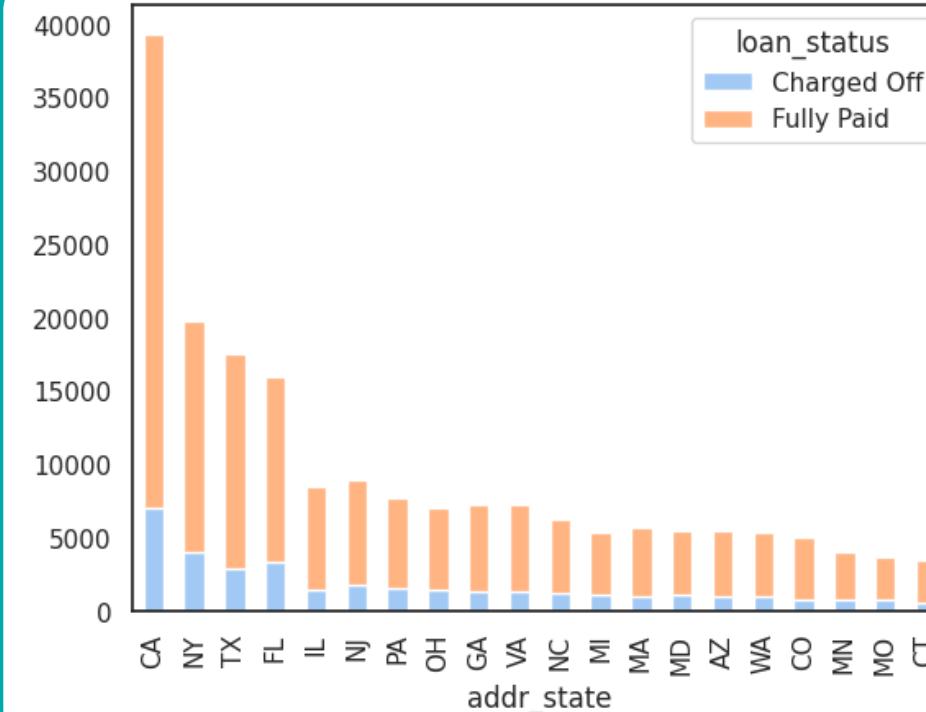
```
df['desc_is_na'].value_counts()
```

True	139844
False	90190
Name:	desc_is_na, dtype: int64

Plot count menunjukkan bahwa proporsi status pinjaman yang lunas (Fully Paid) lebih tinggi untuk deskripsi yang telah diisi (False). Sebagian besar entri tidak memiliki deskripsi (139.844), sementara beberapa entri memiliki deskripsi (90.190). Karena data dalam kolom **desc** tidak lengkap dan tidak relevan dengan pembuatan model, fitur **desc** dihapus.

Berikut merupakan hasil *plotting* analisis pada fitur

zip_code dan addr_state



```
df[['zip_code', 'addr_state']].value_counts()[:20]
```

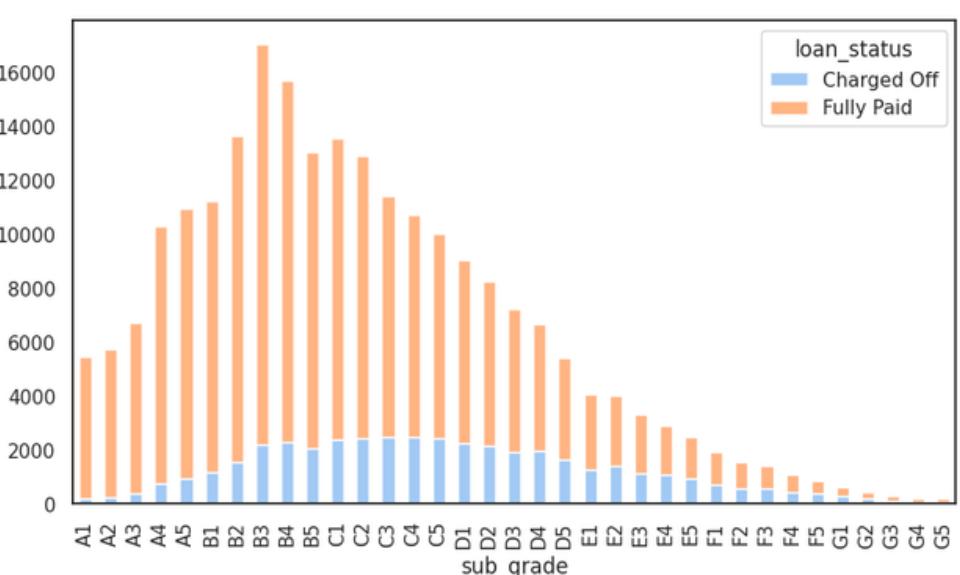
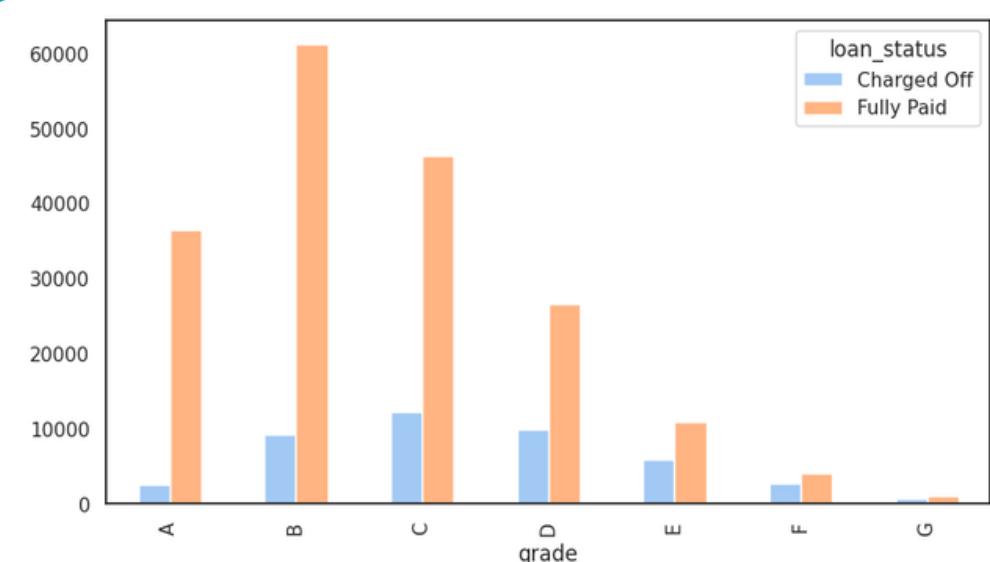
zip_code	addr_state	
945xx	CA	2968
112xx	NY	2646
750xx	TX	2522
100xx	NY	2370
900xx	CA	2290
606xx	IL	2282
331xx	FL	2082
300xx	GA	2070
070xx	NJ	2041
917xx	CA	1913
921xx	CA	1794
770xx	TX	1792
891xx	NV	1740
926xx	CA	1706
330xx	FL	1681
104xx	NY	1527
913xx	CA	1527
117xx	NY	1513
852xx	AZ	1493
925xx	CA	1442

Kesimpulan

Dapat diketahui bahwa 3 (tiga) angka pertama menunjukkan *sectional center* (pusat seksional) dan juga mencakup informasi negara bagian. Karena variabel **addr_state** hanya menjelaskan negara bagian dari tiga digit pertama dalam kode pos, variabel **zip_code** dihapus.

Berikut merupakan hasil plotting analisis pada fitur

grade and sub_grade



Kesimpulan

Dengan demikian, dapat disimpulkan bahwa setiap **sub_grade** mencerminkan **grade** sehingga fitur **grade** dapat dihapus untuk mengurangi redundansi informasi.

```
pd.crosstab(df['grade'], df['loan_status'], normalize='index').round(2)
```

loan_status	Charged Off	Fully Paid
grade		
A	0.07	0.93
B	0.13	0.87
C	0.21	0.79
D	0.27	0.73
E	0.35	0.65
F	0.40	0.60
G	0.42	0.58

Membuat tabel silang antara kolom **sub_grade** dan **loan_status** dan menormalisasi berdasarkan baris (index).

```
pd.crosstab(df['sub_grade'], df['loan_status'], normalize='index').round(2)
```

loan_status	Charged Off	Fully Paid
sub_grade		
A1	0.03	0.97
A2	0.05	0.95
A3	0.06	0.94
A4	0.07	0.93
A5	0.09	0.91
B1	0.11	0.89
B2	0.11	0.89

Membuat tabel silang antara kolom **sub_grade** dan **loan_status** dan menormalisasi berdasarkan baris (index).



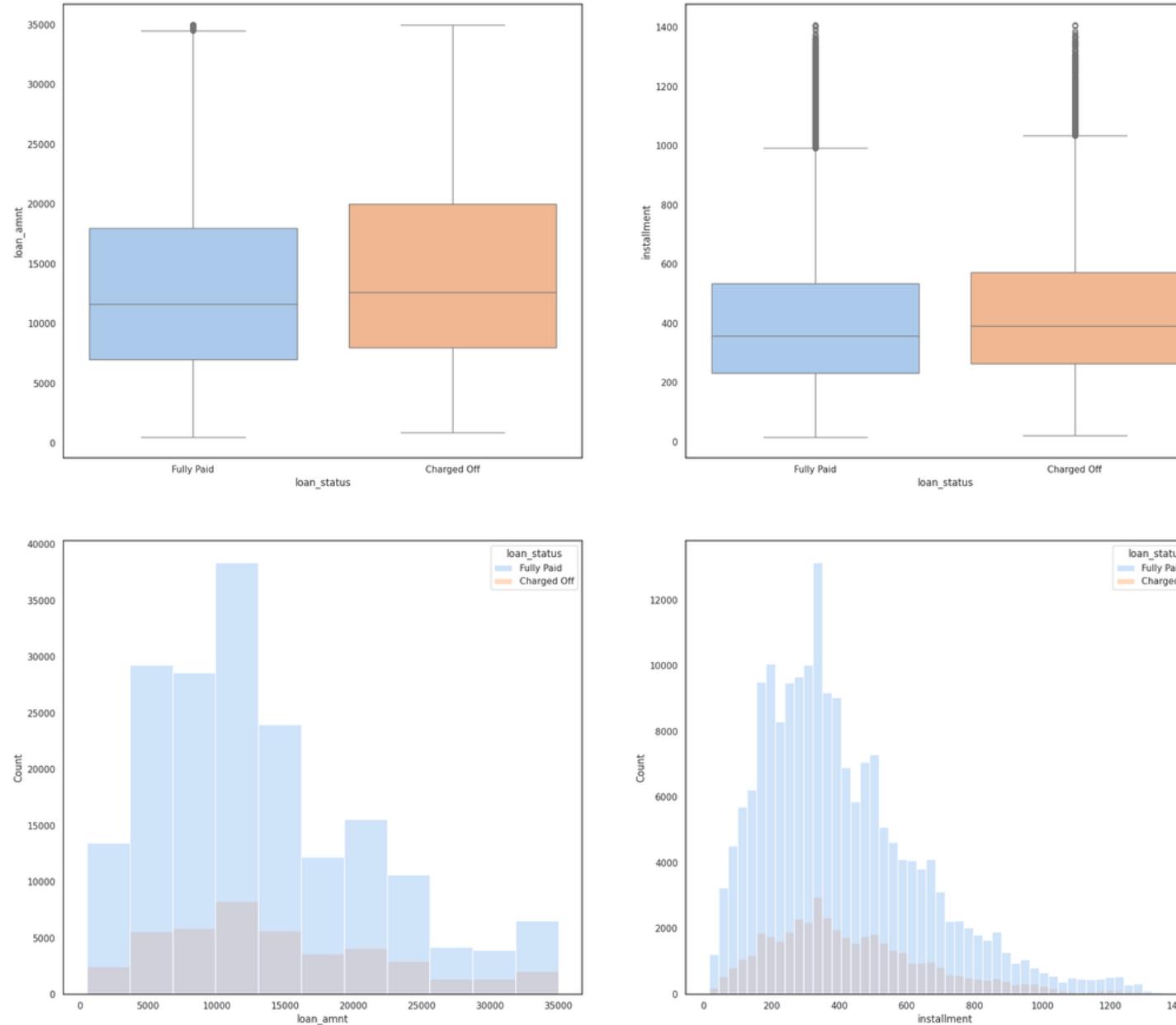
Output lebih lengkap dapat dilihat melalui file .ipynb terlampir

Numerical Features

Proses yang dilakukan pada tahap ini meliputi analisis menyeluruh terhadap fitur yang memiliki tipe data numerik dan berfokus hubungan antara fitur **loan_amnt** dengan fitur lainnya karena berisikan informasi tentang perubahan dalam jumlah pinjaman.

Berikut merupakan hasil *plotting* analisis pada fitur

loan_amnt dan **installment**



Kesimpulan

Tidak ada redundansi antara **loan_amnt** dan **installment** karena fitur **installment** menunjukkan jumlah pembayaran bulanan berdasarkan **loan_amnt**, term, dan **int_rate**. Dengan demikian, fitur **loan_amnt** dan **installment** tetap dipertahankan.

Berikut merupakan hasil *plotting* analisis pada fitur **loan_amnt** dan **installment**

Kesimpulan

Variabel **loan_amnt**, **funded_amnt**, dan **funded_amnt_inv** memiliki korelasi sempurna dan redundan. Selain itu, variabel **policy_code** hanya memiliki satu nilai dan tidak relevan dalam konteks korelasi.

Karena fitur **loan_amnt** merupakan fitur yang berisikan perubahan dalam jumlah pinjaman, maka fitur yang dihapus, yaitu fitur **funded_amnt** dan **funded_amnt_inv** karena redundan, serta **policy_code** karena tidak relevan.





Data Preparation

6

Proses persiapan data sebelum data tersebut dapat
digunakan untuk analisis atau pemodelan.

Data Preparation

Pada tahap ini, proses persiapan data sebelum data dapat digunakan untuk analisis atau pemodelan. Tujuannya adalah untuk memastikan data siap digunakan dalam analisis atau pemodelan.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Handle Missing Data
- 2 Convert Numeric Variable
- 3 Remove Outliers
- 4 Encode Data
- 5 Train Test Split

1 Handle Missing Data

Pada tahap ini, dilakukan identifikasi, analisis, dan pembersihan nilai kosong dalam dataset untuk memastikan konsistensi dan keandalan data yang digunakan dalam analisis atau pemodelan.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Menghapus kolom emp_length dari DataFrame untuk menghilangkan nilai NaN tanpa kehilangan informasi penting.
- 2 Menghapus nilai NaN pada kolom annual_inc, earliest_cr_line, open_acc, pub_rec, revol_util, total_acc, acc_now_delinq dari DataFrame.

Data Preparation

2 Convert Numeric Variable

Pada tahap ini, dilakukan pemeriksaan dan penyesuaian tipe data kolom-kolom dalam DataFrame df untuk memastikan data dapat diproses sesuai dengan kebutuhan analisis atau pemodelan yang akan dilakukan.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Memberikan informasi tentang kolom-kolom dalam DataFrame yang memiliki tipe data numerik dan boolean.
- 2 Mengubah tipe data kolom tertentu dalam DataFrame df menjadi tipe data yang sesuai dengan kebutuhan pemodelan.

```
col_to_int = ['earliest_cr_line', 'open_acc',
              'pub_rec', 'total_acc', 'acc_now_delinq']
col_to_float = ['revol_bal', 'loan_amnt', 'desc_is_na']

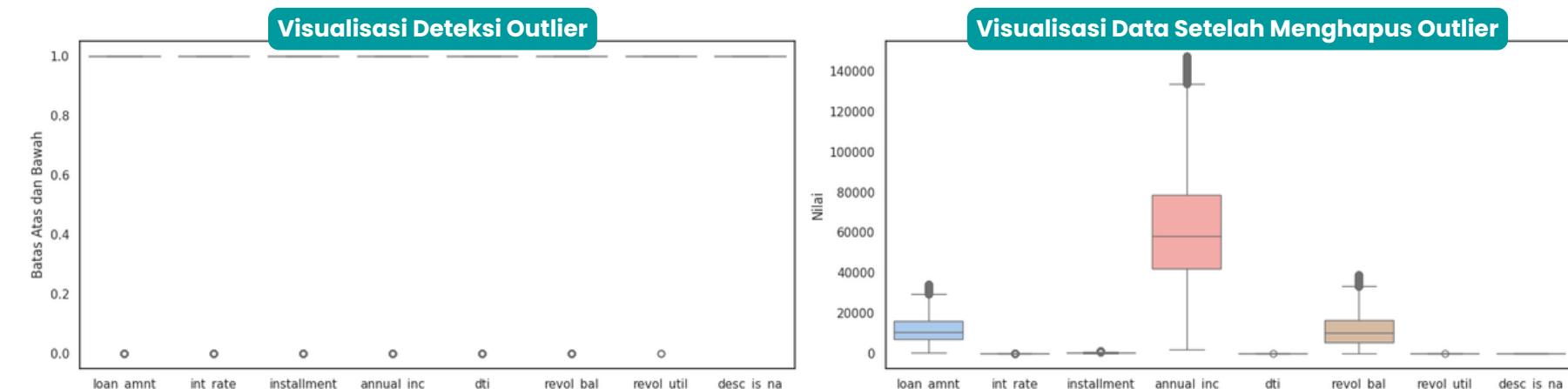
df[col_to_int] = df[col_to_int].astype(int)
df[col_to_float] = df[col_to_float].astype(float)
```

3 Remove Outliers

Pada tahap ini, dilakukan deteksi outlier dalam dataset numerik, diikuti dengan penggantian nilai outlier dengan NaN, dan penghapusan baris yang mengandung nilai NaN.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Mendeteksi outlier dalam DataFrame dengan menggunakan boxplot.
- 2 Membersihkan outlier dari DataFrame.



Data Preparation

4 Encode Data

Pada tahap ini, dilakukan konversi variabel kategorikal dalam dataset menjadi representasi numerik yang sesuai agar dapat dimengerti dan diproses oleh algoritma machine learning.

Proses yang dilakukan pada tahap ini meliputi:

- 1 **Mendefinisikan target** variabel (`loan_status`) dalam dataset, mempersiapkannya, dan mengonversinya dari representasi teks menjadi representasi numerik menggunakan **LabelEncoder**.
- 2 **Mengubah variabel kategorikal** dalam DataFrame menjadi representasi numerik.
- 3 **Mentransformasikan data** pada DataFrame `df` dan menyatukannya dengan variabel target **`loan_status`** menggunakan **OneHotEncoder** dan **MinMaxScaler**.

5 Train Test Split

Pada tahap ini, data dibagi menjadi subset latih dan uji, distribusi kelas dalam variabel target diperiksa, dan fitur serta target disiapkan untuk pelatihan dan pengujian model machine learning.

Proses yang dilakukan pada tahap ini meliputi:

- 1 Membagi DataFrame menjadi dua subset dengan proporsi 20% untuk data uji dan 80% untuk data latih.
- 2 Membagi dataset latih (train) dan dataset uji (test) menjadi dua bagian terpisah yang terdiri dari fitur (`X_train` dan `X_test`) dan variabel target (`y_train` dan `y_test`).



Modelling

7

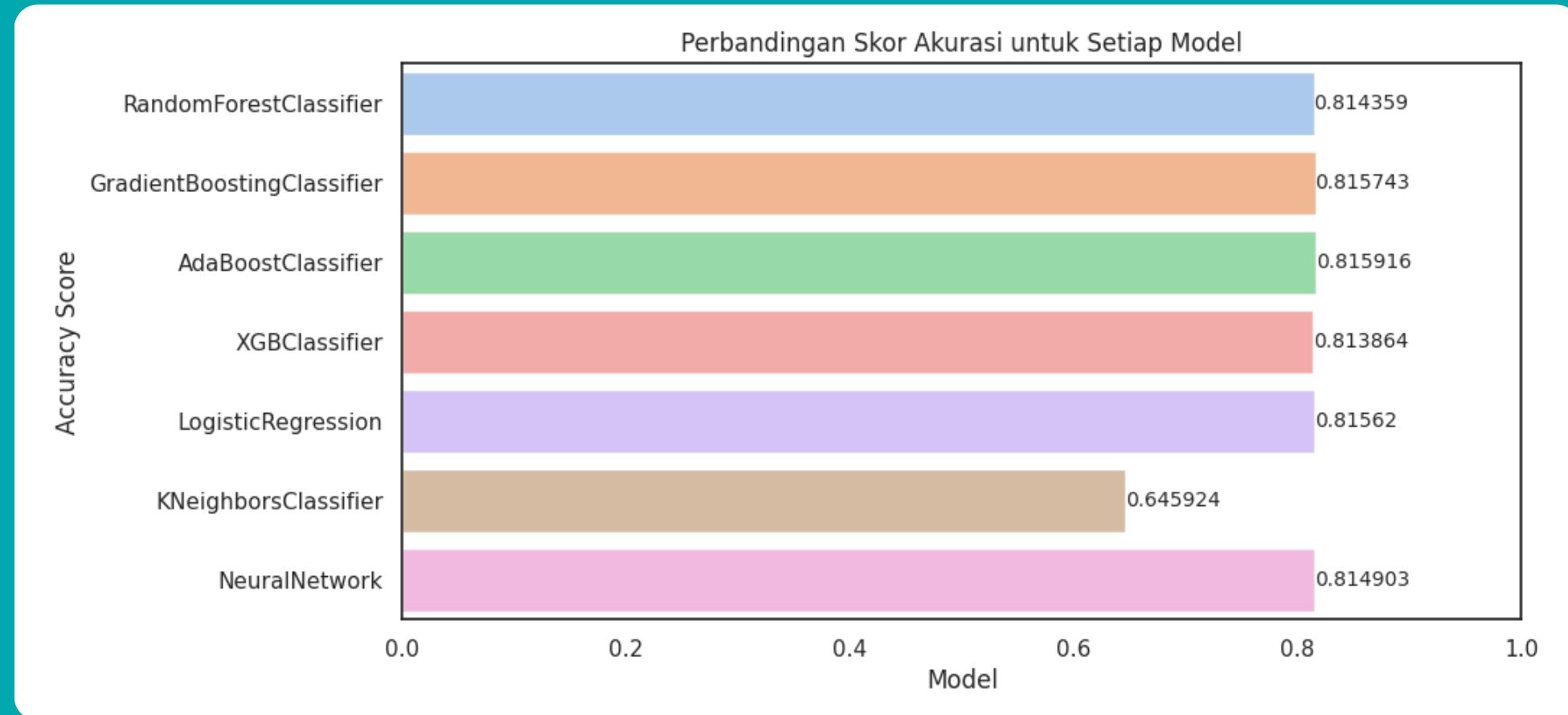
Proses pembangunan dan penyesuaian model berdasarkan data yang tersedia untuk tujuan analisis, prediksi, atau pengambilan keputusan

Modelling

Pada tahap ini, proses pembangunan dan penyesuaian model dilakukan berdasarkan data yang tersedia untuk tujuan analisis, prediksi, atau pengambilan keputusan.

Proses yang dilakukan pada tahap ini meliputi:

- 1** Menetapkan algoritma untuk pelatihan model.
- 2** Melakukan proses pelatihan.
- 3** Menampilkan hasil visualisasi perbandingan skor akurasi untuk setiap model.



Important Inferences

Mayoritas algoritma berhasil mencapai tingkat akurasi sebesar **81%**, dengan pencapaian tertinggi terdapat pada algoritma **AdaBoost** yang mencapai **81.59%**. Hal ini menunjukkan bahwa mayoritas dari model-model tersebut mampu mengklasifikasikan dengan benar sekitar 81% kasus pinjaman.



Evaluation

8

Proses pengukuran kinerja dan akurasi model yang telah dibangun berdasarkan data yang digunakan untuk pelatihan.

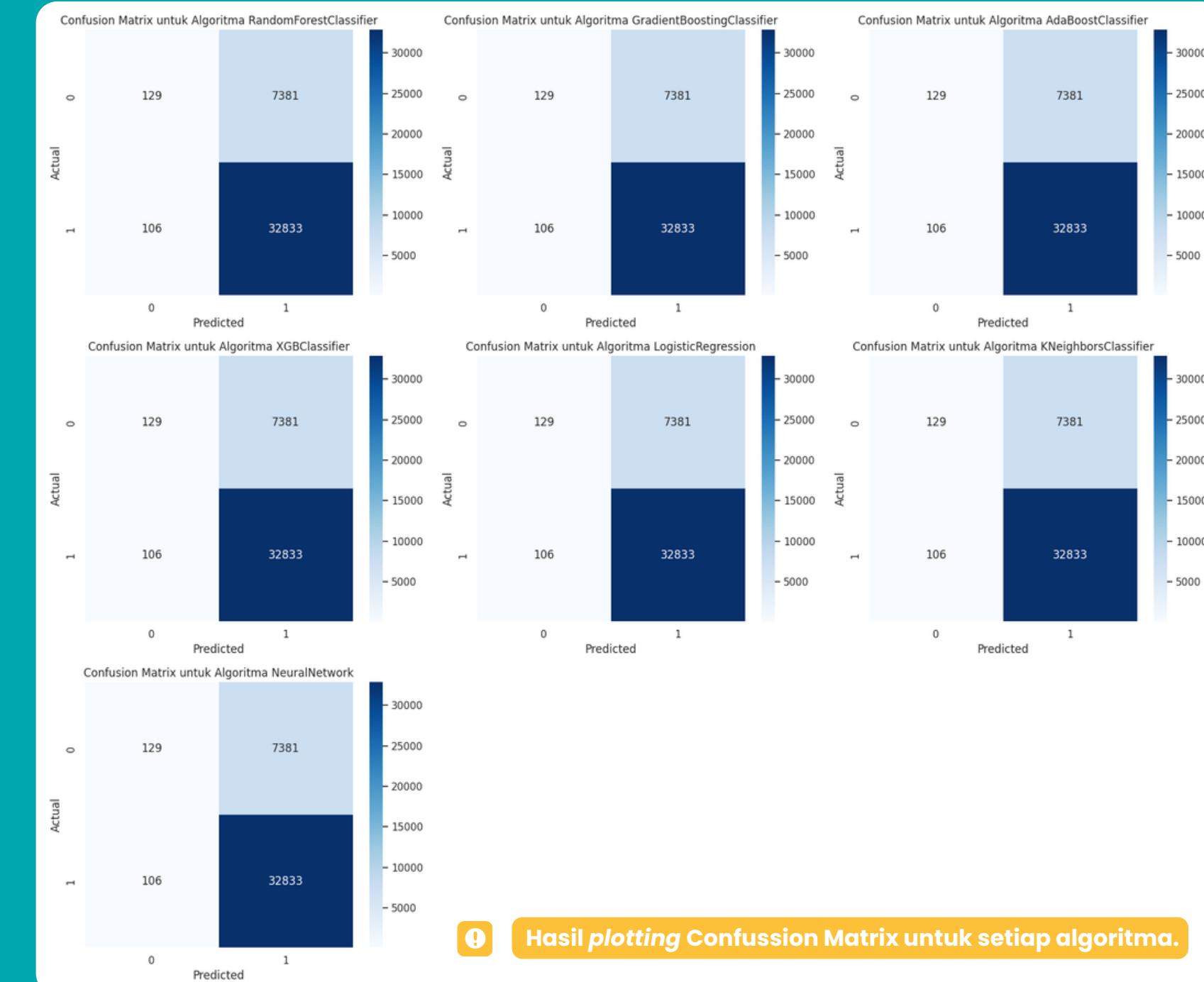
Confusion Matrix

Evaluation

Pada tahap ini, proses pembangunan dan penyesuaian model dilakukan berdasarkan data yang tersedia untuk tujuan analisis, prediksi, atau pengambilan keputusan.

Proses yang dilakukan pada tahap ini meliputi:

- 1** Melakukan evaluasi model dengan menggunakan Confusion Matrix.
- 2** Menampilkan laporan klasifikasi untuk setiap algoritma.
- 3** Melakukan evaluasi model dengan menggunakan kurva ROC (Receiver Operating Characteristic).
- 4** Menampilkan nilai AUC (Area Under the ROC Curve) dari setiap model yang dievaluasi.



Hasil plotting Confusion Matrix untuk setiap algoritma.

Evaluation

Classification Report

Classification Report untuk Algoritma RandomForestClassifier:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma GradientBoostingClassifier:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma AdaBoostClassifier:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma XGBClassifier:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma LogisticRegression:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma KNeighborsClassifier:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Classification Report untuk Algoritma NeuralNetwork:				
	precision	recall	f1-score	support
0	0.55	0.02	0.03	7510
1	0.82	1.00	0.90	32939
accuracy			0.81	40449
macro avg	0.68	0.51	0.47	40449
weighted avg	0.77	0.81	0.74	40449

Important Inferences

1

Terdapat ketidakseimbangan kelas di mana terdapat lebih banyak kasus Fully Paid (kelas 1) daripada kasus Charged Off (kelas 0).

2

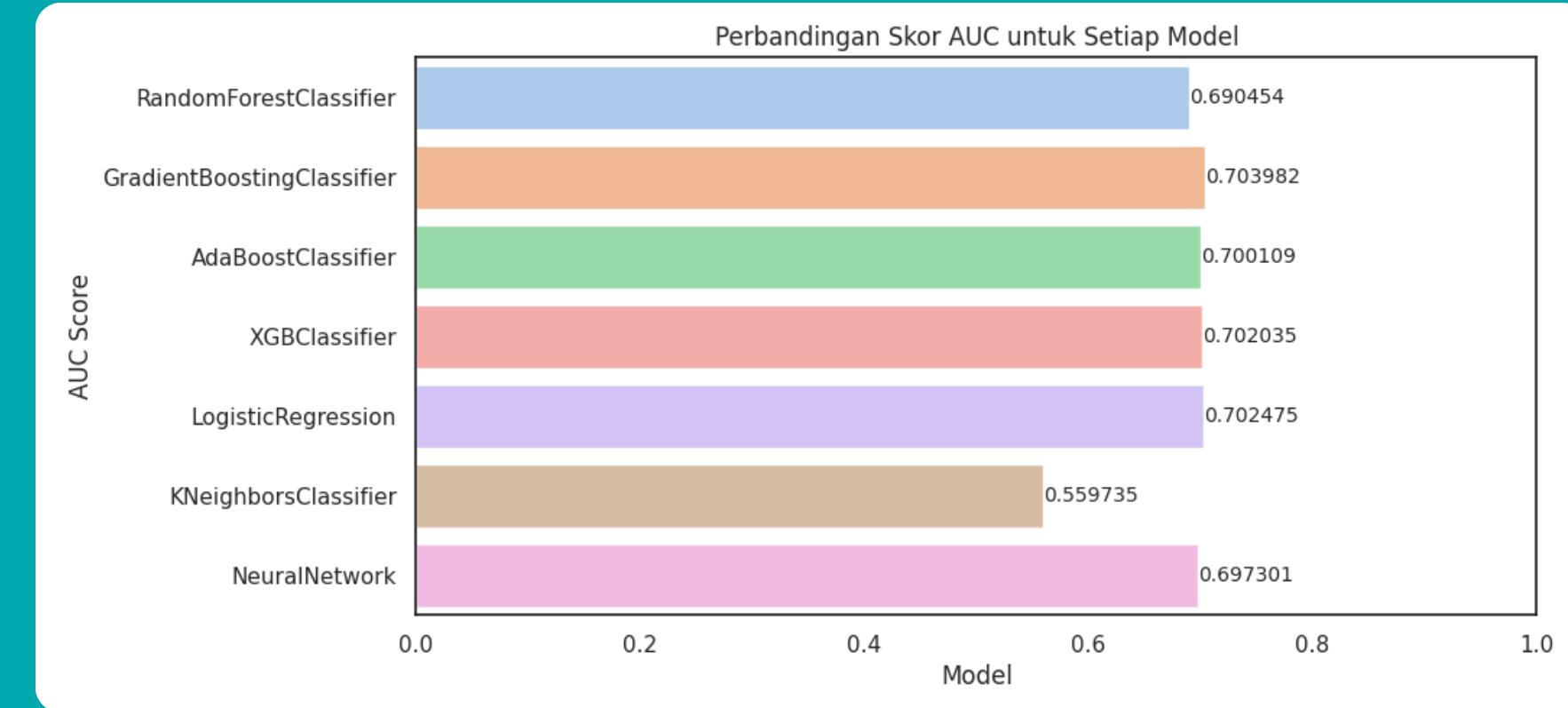
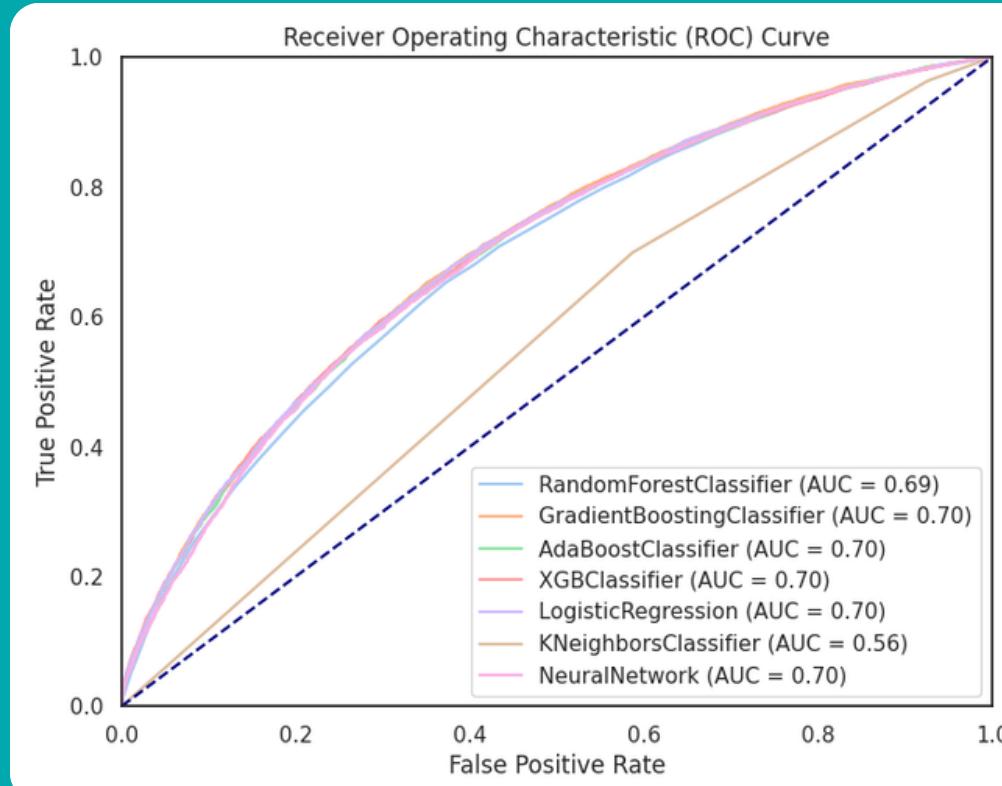
Seluruh model menunjukkan recall yang sangat tinggi (1.00) untuk kelas Fully Paid (kelas 1), yang menunjukkan bahwa seluruh model hampir secara sempurna mengidentifikasi pinjaman yang baik.

3

Presisi untuk kelas Charged Off rendah (sekitar 0.55) untuk semua model. Ini menunjukkan bahwa model mungkin salah mengklasifikasikan banyak pinjaman yang buruk (charged off) sebagai pinjaman yang baik (fully paid).

ROC Curve and AUC

Evaluation



Important Inferences

- 1** Seluruh model kecuali KNeighborsClassifier memiliki skor AUC yang serupa sekitar 0.7. Ini menunjukkan kemampuan moderat untuk membedakan antara pinjaman yang baik dan buruk.
- 2** LogisticRegression memiliki AUC tertinggi (0.702475). Ini menunjukkan kinerja yang sedikit lebih baik dalam membedakan jenis pinjaman.
- 3** KNeighborsClassifier memiliki AUC yang jauh lebih rendah (0.559735). Ini menunjukkan kinerja yang buruk dalam membedakan jenis pinjaman.

Final Task of ID/X Partners

Data Scientist Project Based Internship



Kesimpulan

8

Apa yang dapat disimpulkan dari proyek ini?

Kesimpulan Proyek



Kesimpulan

1

Kinerja Model

Model risiko kredit lebih baik dalam mengidentifikasi pinjaman yang baik daripada yang buruk karena ketidakseimbangan kelas data pelatihan.

2

Skor AUC

Skor AUC menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam menilai risiko kredit.



Saran

1

Data Balancing

Teknik pengelolaan ketidakseimbangan data dapat diterapkan untuk meratakan proporsi kasus positif dan negatif dalam dataset pelatihan.

2

Tuning Model

Proses penyetelan model dapat dilakukan untuk menyesuaikan hiperparameter algoritma guna meningkatkan kinerja terutama pada kelas minoritas, seperti kasus **Charged Off**.

3

Business Recommendation

Dapat dilakukan pembuatan *business recommendation* berdasarkan data yang tersedia untuk mengembangkan potensi bisnis.



Final Task of ID/X Partners

Data Scientist Project Based Internship



Thank You

For Your Attention



Get In Touch

Contact Me For More Information



Mochammad Adhi Buchori



adhi.buchori@gmail.com



adhibuchori



(+62) 822 2526 3810

- Make it **Shine More** -