

# PREDICTION MODEL

## IDX/PARTNERS - DATA SCIENTIST

PRESENTED BY : NETANIA PANGALINAN



# NETANIA PANGALINAN

## FRESH GRADUATE UNM (DATA ENTHUSIAST)

Perkenalkan nama saya Netania, seorang data enthusiast. Dengan latar belakang di Bisnis Digital, saya tertarik pada analisis data, machine learning, dan data visualization.

Memiliki pengalaman internship di BRI sebagai Document Management, serta menjadi awardee beasiswa Job Connector Okupasi Data Analysis di Luar Sekolah.

Saat ini, sedang fokus mendalami Python, SQL, dan machine learning frameworks untuk terus berkembang di dunia data.



 Toraja Utara, Sulawesi Selatan

 netaniapangalinan@gmail.com

 netania pangalinan



# COURSE AND CERTIFICATION

-  [Belajar Dasar Data Science - Dicoding](#)
-  [Belajar Analisis Data dengan Python - Dicoding](#)
-  [Introduction to Data Science with Python - DqLab](#)
-  [Belajar Dasar Structure Query Languange - Dicoding](#)

# ABOUT COMPANY



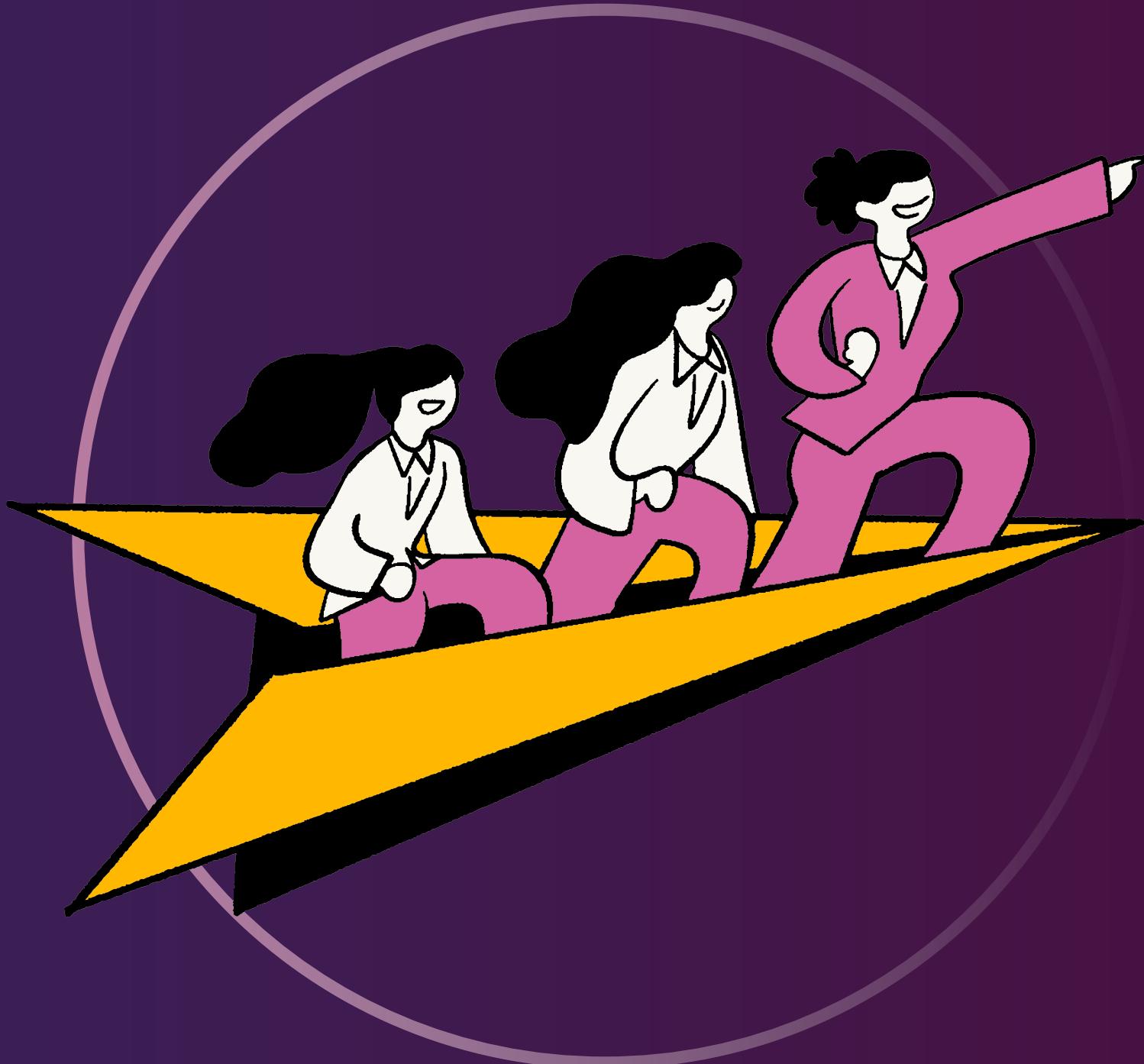
IDX Partners, atau dikenal juga sebagai id/x partners, adalah perusahaan konsultan yang berfokus pada solusi data analytics dan decisioning di Indonesia. Sejak berdiri pada tahun 2006, perusahaan ini telah membantu institusi jasa keuangan untuk tumbuh secara profitabel dan meningkatkan pangsa pasar melalui penerapan AI/ML, otomatisasi pengambilan keputusan, optimasi profit, dan analytics-CRM.

IDX Partners bekerja sama dengan Rakamin Academy dalam Project-Based Internship (PBI) untuk mengembangkan talenta data melalui proyek nyata. Salah satu proyeknya adalah Loan Risk Analysis, yang menganalisis risiko pinjaman berdasarkan dataset 2007-2014 menggunakan teknik analitik dan machine learning untuk membantu mitigasi risiko kredit.



# PROJECT PORTOFOLIO

Proyek Loan Risk Analysis ini bertujuan untuk menganalisis dan memprediksi risiko pinjaman menggunakan dataset historis dari 2007–2014 yang mencakup lebih dari 466.285 entri dan 75 fitur. Model terbaik yang digunakan adalah Gradient Boosting Classifier dengan akurasi 89.04%, yang mampu mengklasifikasikan ulang 88.9% pinjaman buruk menjadi pinjaman baik, sehingga berpotensi mengurangi risiko gagal bayar secara signifikan. Proyek ini menunjukkan bagaimana data analytics dan machine learning dapat membantu lembaga keuangan dalam mengoptimalkan strategi mitigasi risiko dan meningkatkan profitabilitas kredit.



[Link Code Here](#)

[Link Github Here](#)

[Link Video Here](#)

# DATA UNDERSTANDING

## **baris dan kolom**

Dataset memiliki 466,285 baris dan 75 kolom

## **Tipe Data**

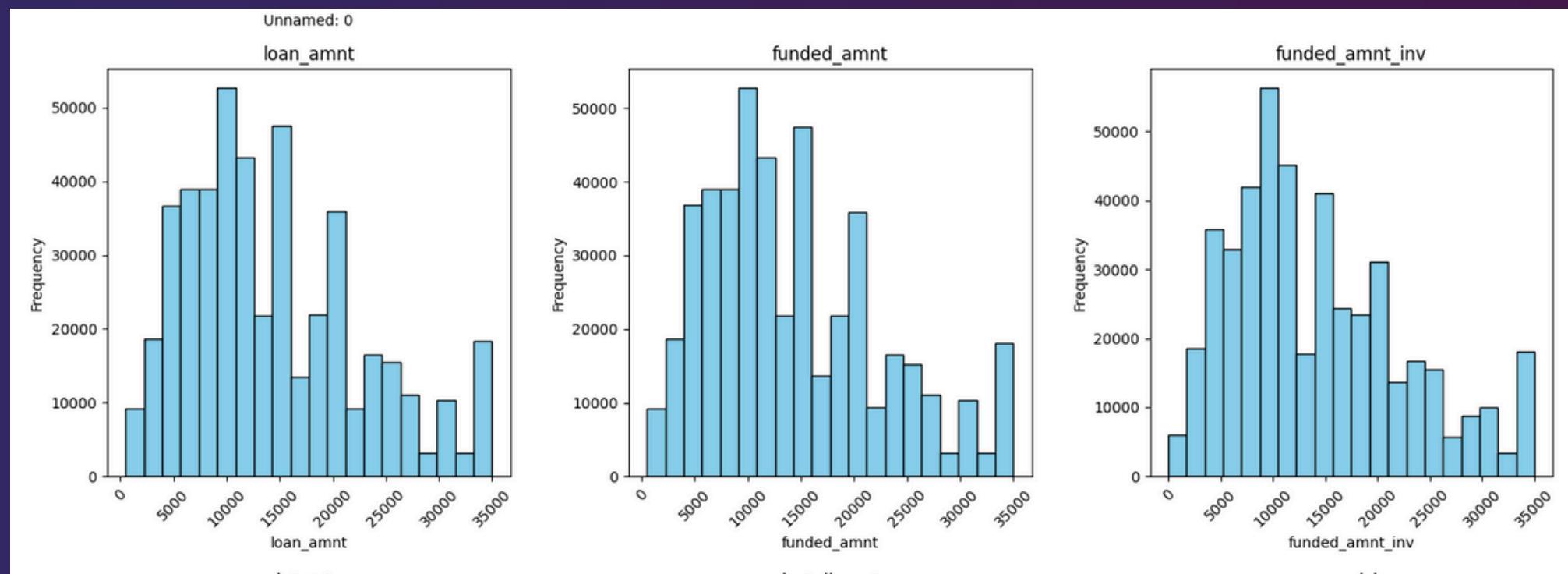
Dataset terdiri dari 46 kolom bertipe float64, 7 kolom bertipe int64, dan 22 kolom bertipe object.

## **Missing Values & Duplikasi**

Dataset tidak memiliki baris duplikat, tetapi beberapa kolom memiliki banyak nilai yang hilang, dan yang sepenuhnya kosong.

# EXPLORATORY DATA ANALYSIS

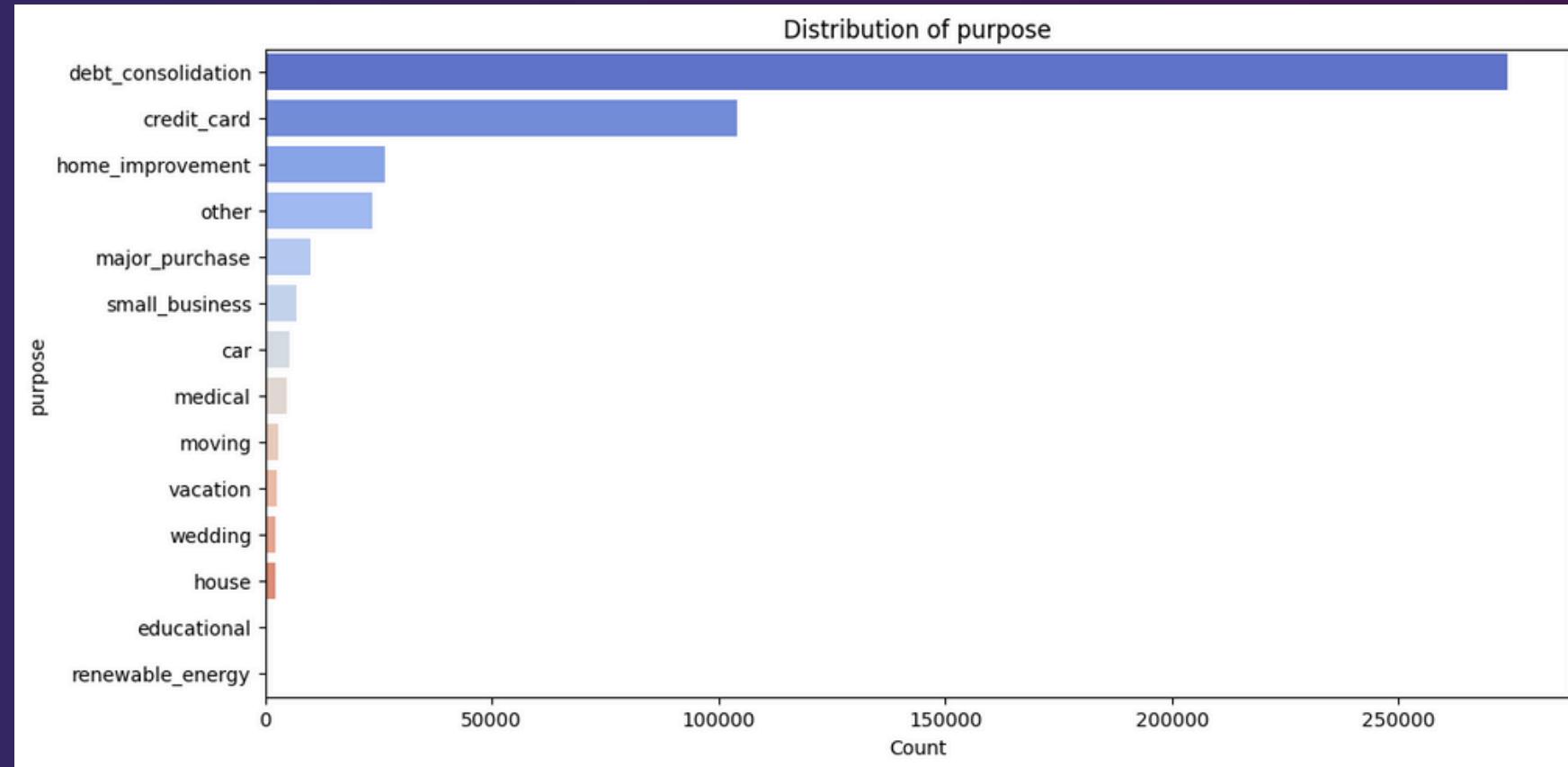
## DISTRIBUSI DATA NUMERIK



Distribusi ketiga variabel numerik—loan amount, funded amount, dan funded amount inv—menunjukkan pola yang hampir serupa, dengan mayoritas pinjaman berada di kisaran 5.000 hingga 15.000. Distribusi cenderung right-skewed, menandakan sebagian kecil pinjaman bernilai tinggi. Kemiripan bentuk distribusi menunjukkan bahwa jumlah pinjaman yang diajukan umumnya disetujui dan didanai hampir sepenuhnya oleh investor.

# EXPLORATORY DATA ANALYSIS

## DISTRIBUSI DATA KATEGORIK

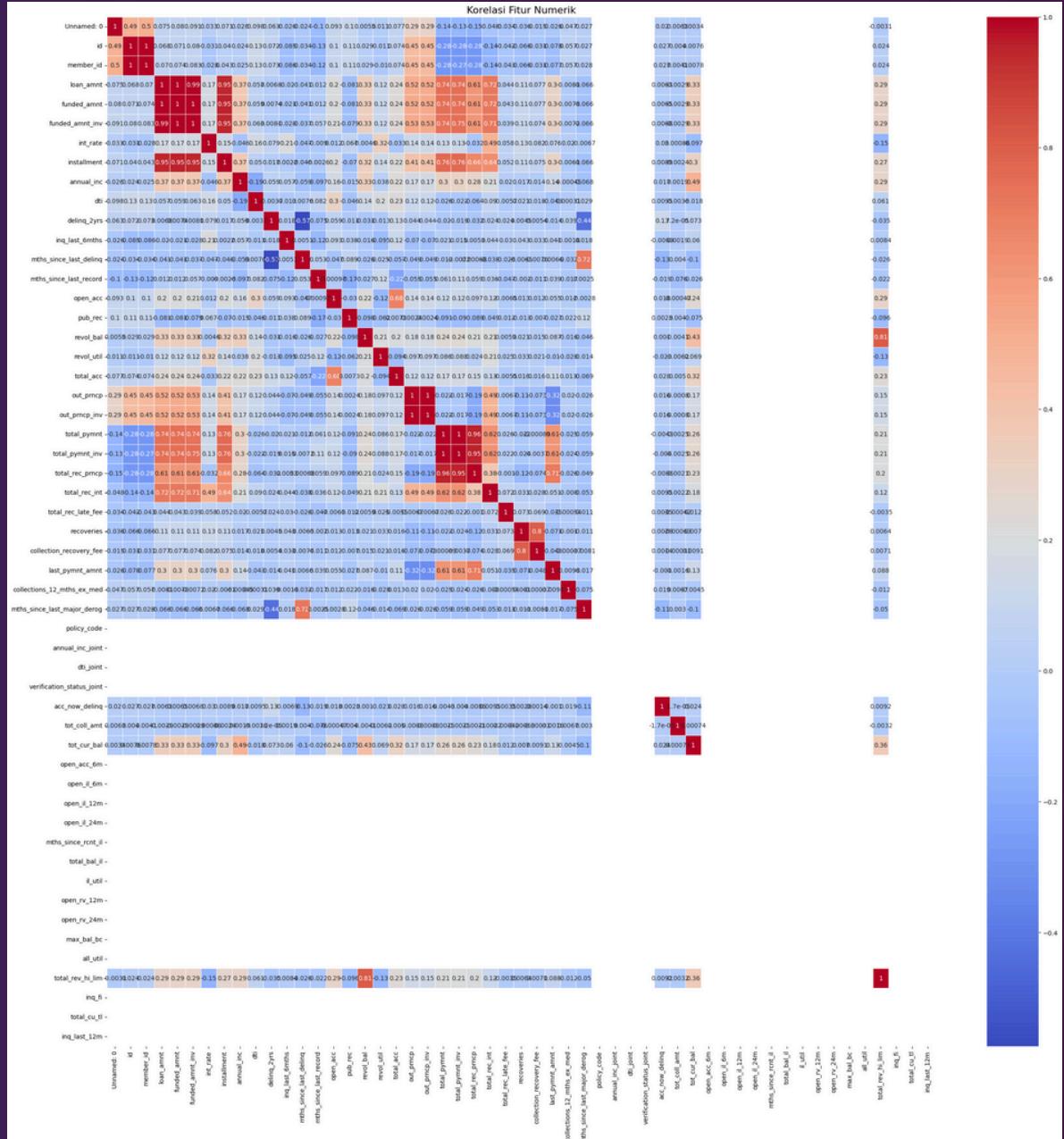


Mayoritas pinjaman digunakan untuk konsolidasi utang dan kartu kredit, sementara kategori lain seperti perbaikan rumah dan bisnis jauh lebih sedikit. Pinjaman untuk pendidikan dan energi terbarukan paling jarang diajukan.

# EXPLORATORY DATA ANALYSIS

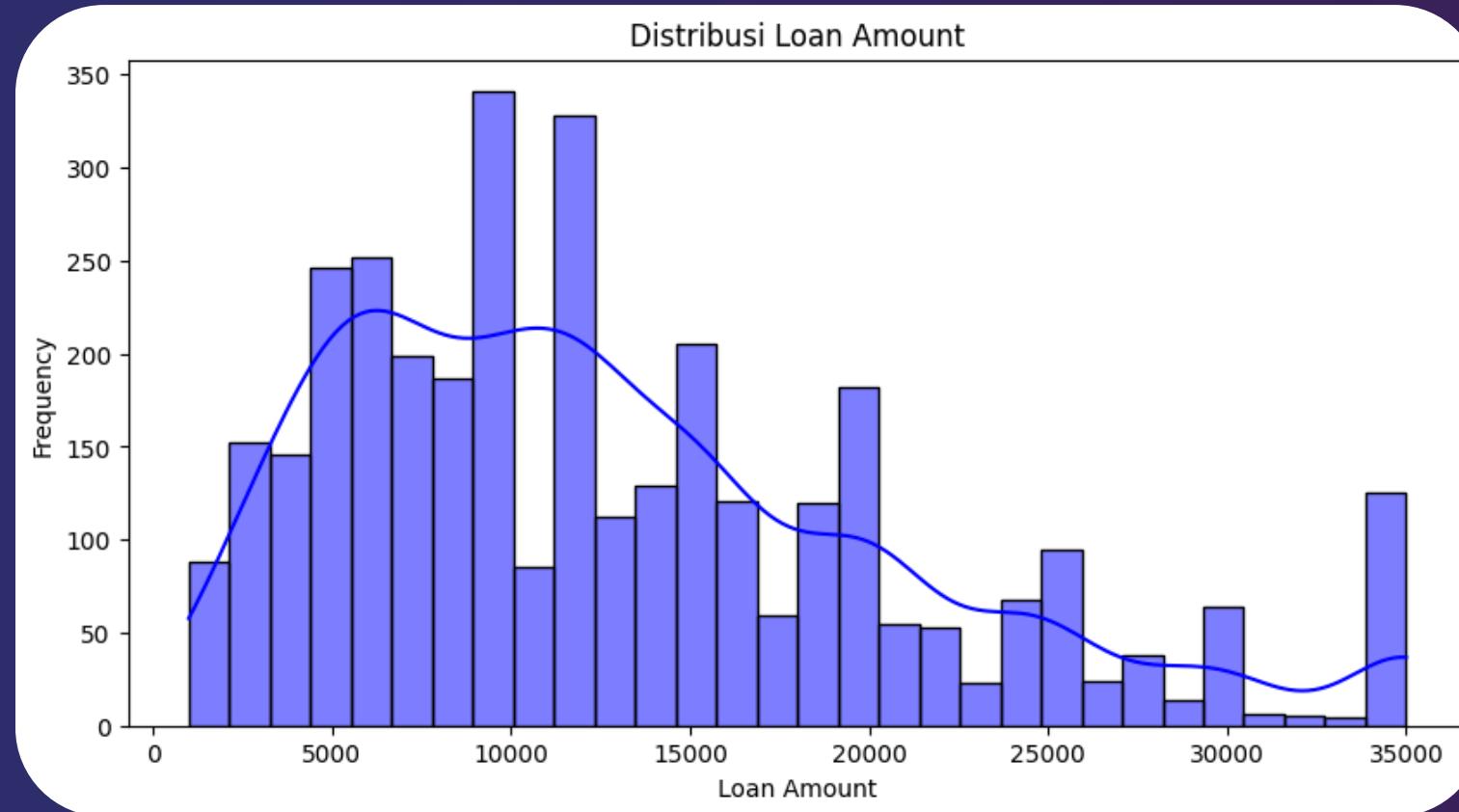
## HEATMAP KORELASI NUMERIK

Heatmap korelasi ini menunjukkan hubungan antar variabel numerik dalam dataset. Terlihat bahwa loan amount, funded amount, dan funded amount invested memiliki korelasi sangat tinggi, menandakan keselarasan antara jumlah pinjaman yang diajukan, didanai, dan diinvestasikan. Selain itu, total pembayaran dan total pinjaman juga memiliki korelasi positif yang kuat, yang logis karena semakin besar pinjaman, semakin besar total pembayaran. Di sisi lain, beberapa fitur menunjukkan korelasi rendah atau hampir nol, menandakan hubungan yang lemah atau tidak signifikan.



# EXPLORATORY DATA ANALYSIS

## UNIVARIAT



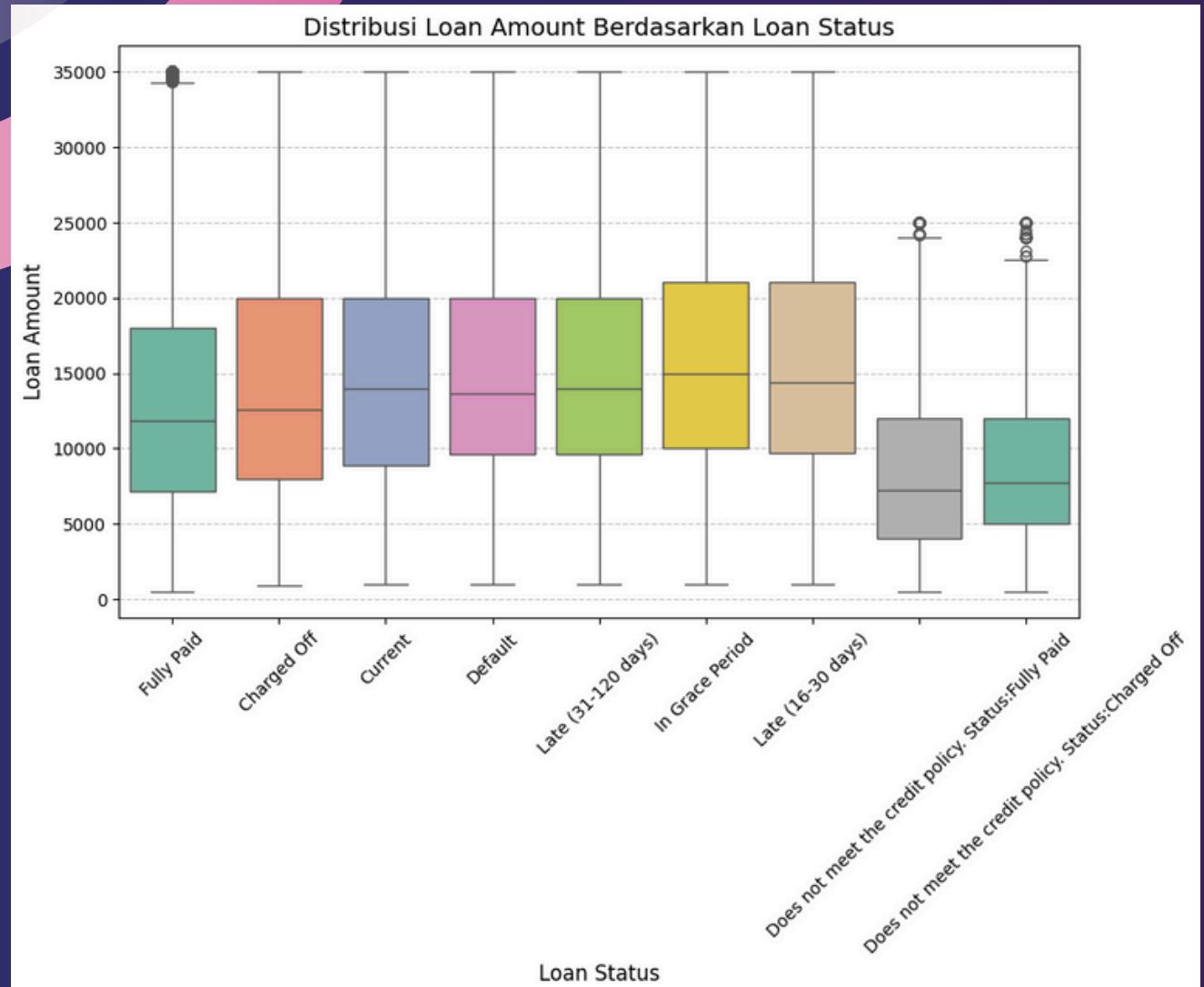
Distribusi Loan Amount ditampilkan dalam histogram dengan kernel density estimation (KDE). Grafik ini menunjukkan bahwa mayoritas pinjaman berkisar antara 5.000 hingga 15.000, dengan beberapa puncak signifikan. Distribusi cenderung right-skewed, menandakan ada sebagian kecil peminjam yang mengambil pinjaman dalam jumlah besar. Insight ini berguna dalam memahami pola pinjaman dan menentukan strategi segmentasi pelanggan.

# EXPLORATORY DATA ANALYSIS

## BIVARIAT

### Analisis Bivariat - Loan Amount vs Loan Status

Boxplot ini menunjukkan distribusi Loan Amount berdasarkan Loan Status. Mayoritas kategori memiliki distribusi yang mirip, dengan median pinjaman berkisar antara 10.000 hingga 15.000. Namun, kategori "Does not meet the credit policy" memiliki nilai pinjaman yang lebih rendah. Outlier terlihat di hampir semua kategori, terutama pada pinjaman yang lebih besar. Analisis ini membantu memahami hubungan antara jumlah pinjaman dan status pembayaran, yang berguna dalam menilai risiko kredit.



# DATA PREPARATION

## REDUKSI DIMENSI DAN TRANSFORMASI TANGGAL

### Menghapus Fitur dengan Korelasi Tinggi

Menghapus Fitur dengan korelasi lebih dari untuk mengurangi multikolinearitas.

### Konversi Kolom Tanggal ke Selisih Bulan

Kolom tanggal dikonversi ke format datetime dengan perbaikan format.

### Menangani Kesalahan Tahun

Koreksi dilakukan untuk tahun yang salah (misalnya, tahun di atas 2025 dikurangi 100 tahun).

# DATA PREPARATION

## TRANSFORMASI KOLOM

### Mengubah 'emp\_length' ke numerik

Ekstraksi angka dari data teks dan konversi ke tipe float agar dapat digunakan dalam analisis kuantitatif.

### Mengubah 'term' ke numerik

Menghapus kata "months", lalu mengonversi ke tipe integer menggunakan pd.to\_numeric().

### Menangani nilai NaN pada 'term'

Mengisi nilai yang hilang dengan median untuk menjaga kestabilan distribusi data.

# DATA PREPARATION

## HANDLE MISSING VALUES

### Drop irrelevant column

menghapus kolom dengan missing values di atas 50% (12 kolom)

### numerik values with median

Nilai yang hilang pada kolom numerik diisi menggunakan SimpleImputer dengan strategi median.

### numerik values with mode

Nilai yang hilang pada kolom kategorikal diisi menggunakan SimpleImputer dengan strategi most\_frequent atau modus.

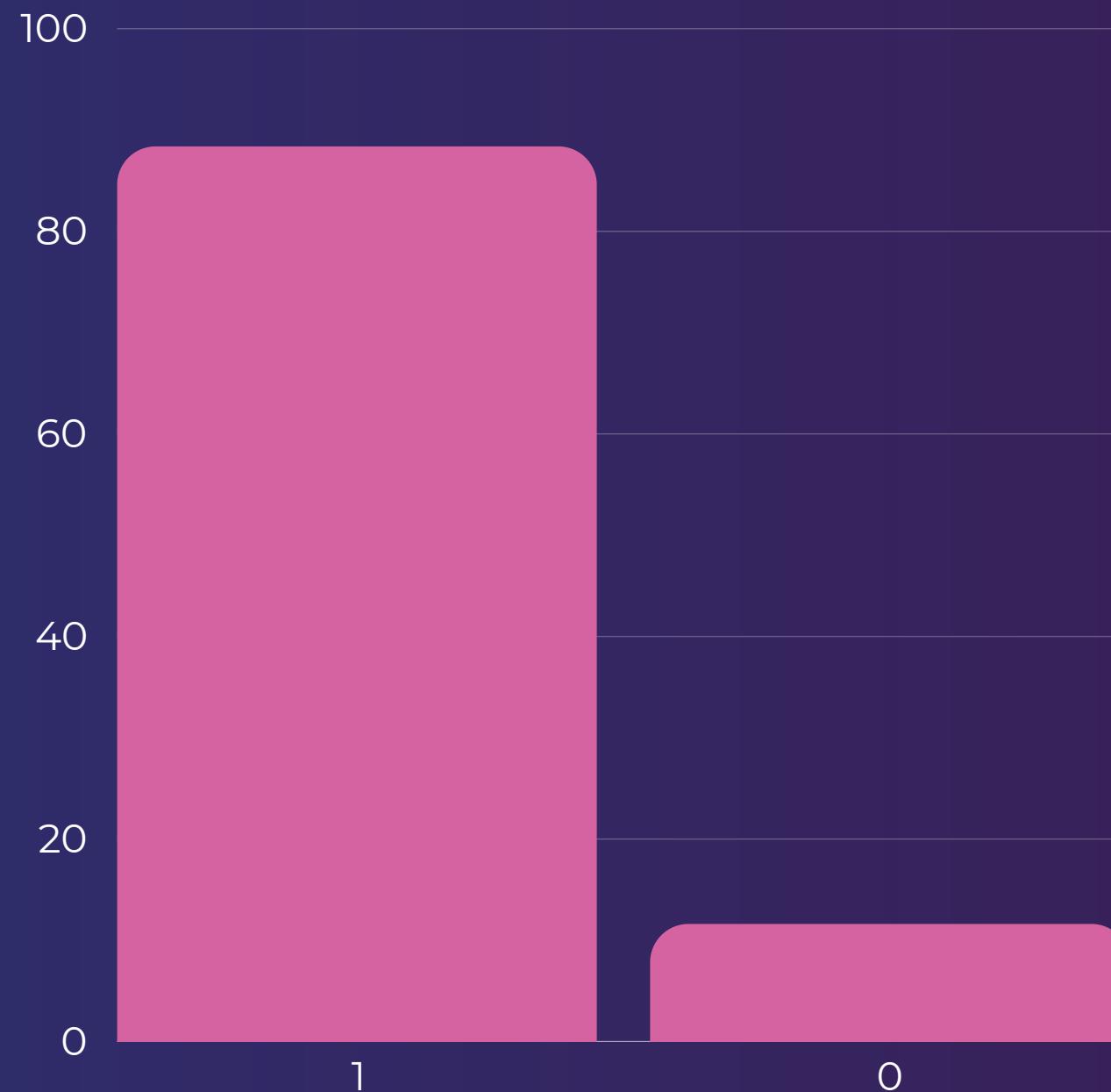
# DATA PREPARATION

## DISTRIBUSI STATUS PINJAMAN

Loan Ammount	Count
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

Dataset menunjukkan variasi status pinjaman dengan jumlah terbanyak berada pada kategori "Current" (224.226 pinjaman), yang berarti pinjaman masih berjalan dan belum jatuh tempo. Diikuti oleh "Fully Paid" (184.739 pinjaman), yang menandakan pinjaman telah lunas tanpa masalah.

Sementara itu, "Charged Off" (42.475 pinjaman) mencerminkan pinjaman yang gagal bayar dan tidak dapat ditagih kembali. Kategori lain seperti "Late (31-120 days)" (6.900 pinjaman) dan "Late (16-30 days)" (1.218 pinjaman) menunjukkan keterlambatan pembayaran dalam berbagai rentang waktu.



# DATA PREPARATION

## LABEL DATA

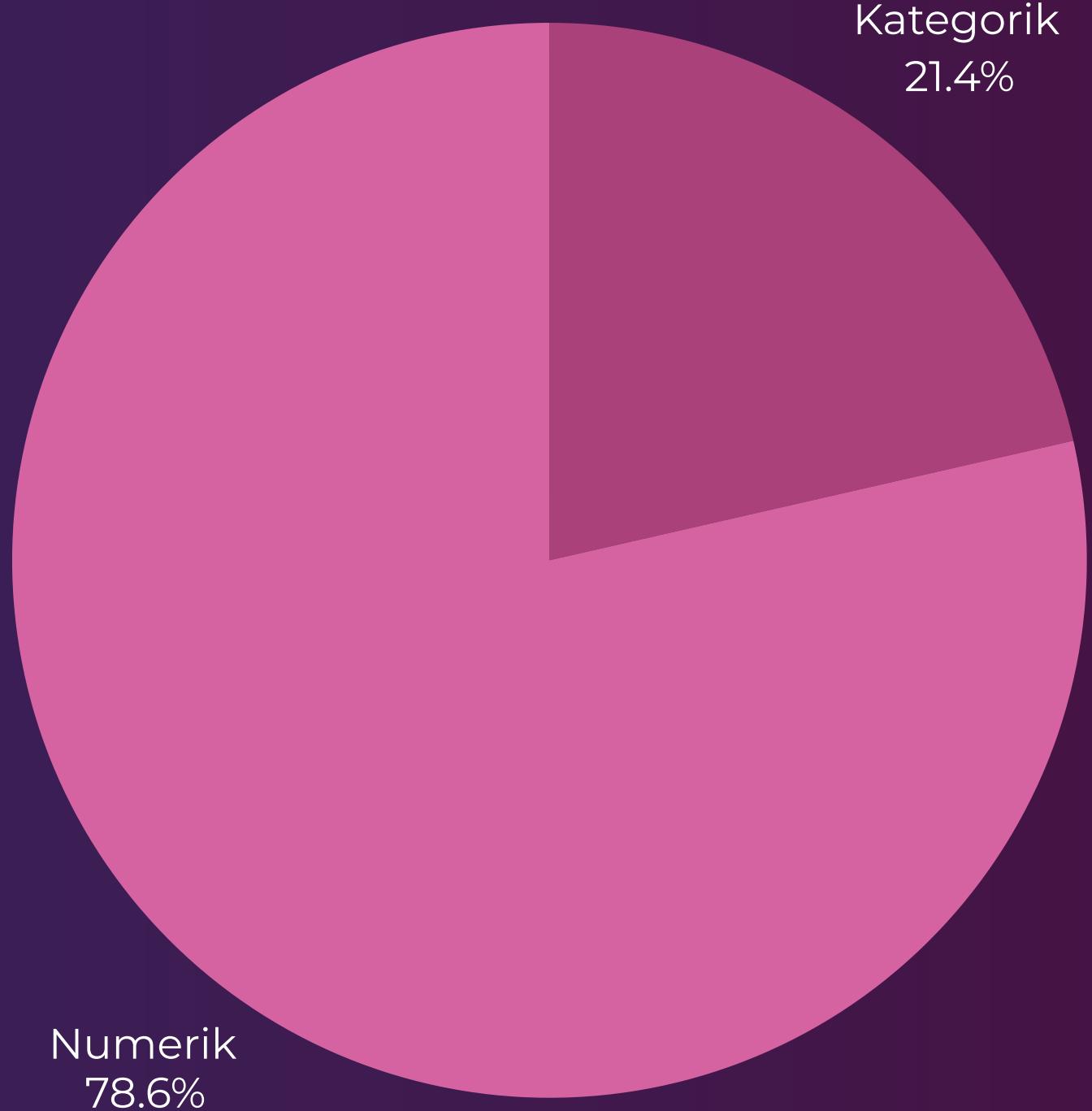
Dalam dataset, label loan\_status dikategorikan menjadi dua kelas:

- 1 (88,38%) → Menunjukkan pinjaman dalam kondisi baik, seperti Current atau Fully Paid.
- 0 (11,62%) → Menunjukkan pinjaman dengan risiko gagal bayar, seperti Charged Off atau Default.

# FEATURE SELECTION

Setelah melalui proses seleksi fitur, dataset kini terdiri dari 14 fitur, dengan 3 fitur kategorikal dan sisanya berupa fitur numerik.

- Fitur kategorikal (3) → grade, home\_ownership, initial\_list\_status
- Fitur numerik → Termasuk loan\_amnt, int\_rate, emp\_length, dti, revol\_bal, dan lainnya, yang berperan penting dalam analisis kuantitatif.



# FEATURE ENGINEERING

## One - Hot Encoding - kategorik

One-Hot Encoding mengubah fitur kategorikal menjadi variabel biner (0/1) agar dapat diproses oleh model tanpa memperkenalkan hubungan numerik yang tidak sesuai sehingga ada 15 kategorik fitur yang di encode

## Standarisasi - Numerik

Standarisasi dilakukan untuk menyelaraskan skala fitur numerik dengan mengonversinya ke distribusi dengan rata-rata nol dan standar deviasi satu. Proses ini bertujuan meningkatkan stabilitas dan kinerja model dengan menghindari dominasi fitur tertentu akibat perbedaan skala. Total fitur numerik yang distandarisasi ada 10 kolom

# DATA MODELLING

## drop loan status di feature x

```
X = df_model.drop(labels=['loan_status'],axis=1)  
y = df_model[['loan_status']]
```

## Splitting Data

```
X_train, X_test, y_train, y_test = train_test_split(X, y,  
test_size = 0.2, stratify=y, random_state = 42)
```



# DATA MODELLING

## INISIASI MODEL

### Inisiasi Model

Inisiasi Logistic Regression, Random Forest, Gradient Boosting dan XGBClassifier

### Training Model

Training ketiga model dengan da train sebesar 80%

### Evaluasi Performa Model

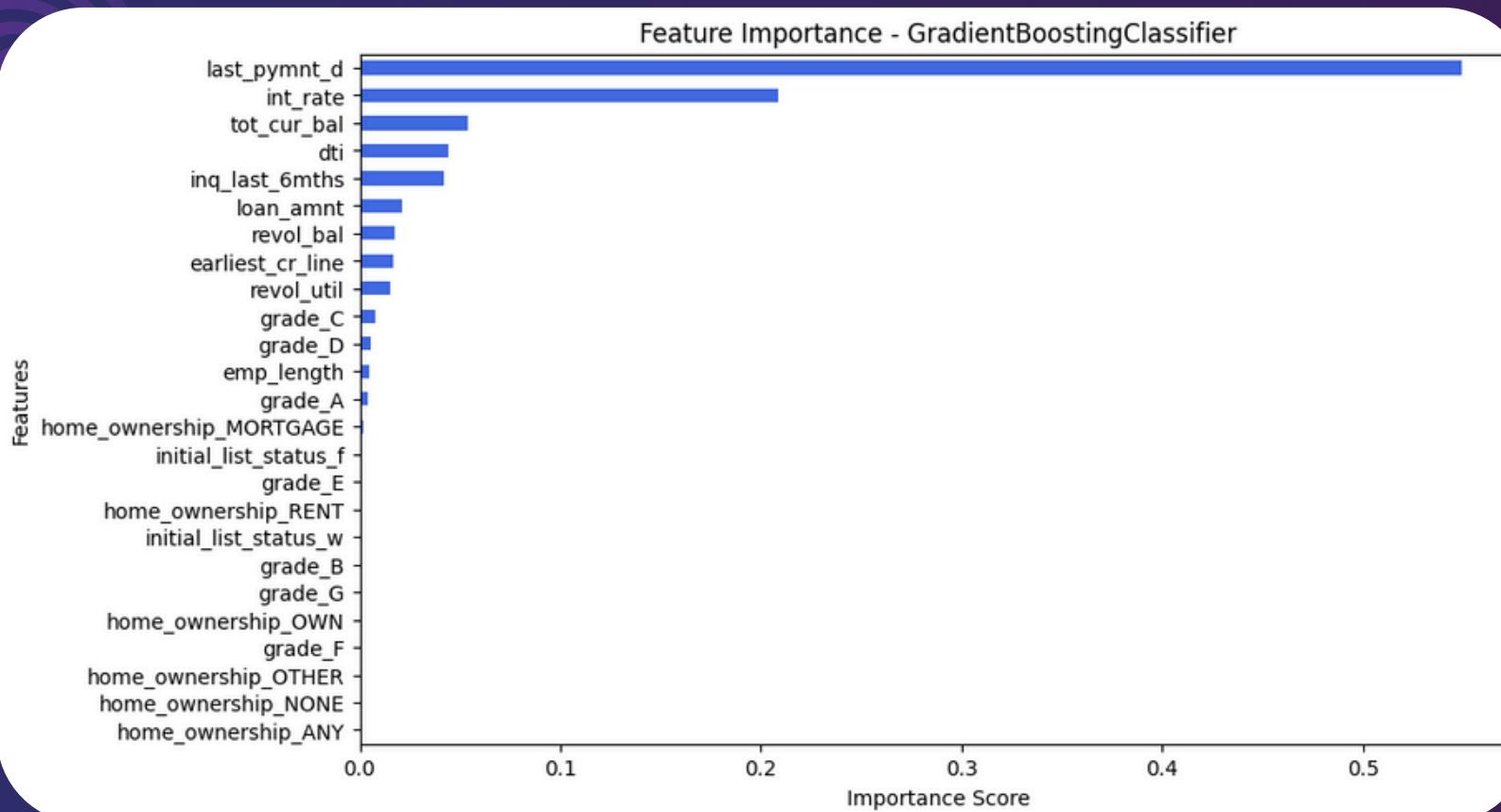
Evaluasi Performa model dengan Confussion Matrix

# MODEL EVALUATION



Berdasarkan hasil evaluasi, Gradient Boosting mencapai akurasi tertinggi (89.40%), diikuti oleh XGBoost (88.94%), Random Forest (88.83%), dan Logistic Regression (88.40%). Hal ini menunjukkan bahwa Model berbasis *boosting* (XGBoost dan Gradient Boosting) menunjukkan keunggulan dalam menangkap pola kompleks dalam data dibandingkan model linear atau berbasis pohon keputusan biasa.

Model	Akurasi
Logistic Regression	0.8840
Random Forest	0.8883
XGBoost	0.8894
Gradient Boosting	0.894



# FEATURE IMPORTANCE



Berdasarkan hasil Feature Importance dari model Gradient Boosting Classifier, fitur "last\_pymnt\_d" memiliki kontribusi terbesar dalam prediksi, menunjukkan bahwa tanggal pembayaran terakhir sangat berpengaruh terhadap keputusan model. Disusul oleh "int\_rate" yang merepresentasikan suku bunga pinjaman, serta "tot\_cur\_bal", yang mencerminkan total saldo saat ini sebagai indikator penting dalam menilai kemampuan pembayaran.

# CONCLUSION

Berdasarkan hasil Feature Importance dari model Gradient Boosting Classifier, fitur "last\_pymnt\_d" memiliki kontribusi terbesar dalam prediksi, menunjukkan bahwa tanggal pembayaran terakhir sangat berpengaruh terhadap keputusan model. Disusul oleh "int\_rate" yang merepresentasikan suku bunga pinjaman, serta "tot\_cur\_bal", yang mencerminkan total saldo saat ini sebagai indikator penting dalam menilai kemampuan pembayaran.

----- Existing Data -----		
Total Loans :	466285	100.0%
Bad Loans :	466285	100.0%
Good Loans :	0	0.0%
----- After Modeling -----		
Total Loans :	466285	100.0%
Bad Loans :	466285	100.0%
Predicted Bad Loans :	414527	88.9%
Predicted Good Loans :	51758	11.1%
Bad Loans After Prediction :	51758	11.1%
Bad Loan Growth Rate :	-88.9%	
Good Loans :	0	0.0%
Good Loans After Prediction :	414527	88.9%
Good Loan Growth Rate :	N/A%	



# THANK YOU!

DATA ANALYSIS IS KEY TO BUSINESS  
GROWTH AND SUCCESS!