CREDIT RISK ANALYSIS

Final Task of Based Project Internship
ID/X Partner - Data Scientist

Presented by: Sabilla Farah Annisa

Sabilla Farah Annisa

A Statistics graduate with expertise in data science and analytics, skilled in MySQL, Excel, and statistical software such as R Studio, SPSS, Minitab, Gretl, and WinBugs. I am also proficient in creative tools including Photoshop, Premiere Pro, and After Effects, which I use to enhance data visualization and storytelling. Continuously eager to learn, collaborate, and create impact through data.

BUSINESS UNDERSTANDING

Company Profile

id/x partners

Sejak berdiri pada tahun 2002, id/x partners telah berkembang sebagai perusahaan konsultan yang berfokus pada penyediaan solusi data analytics dan decisioning (DAD). Layanan yang ditawarkan tidak hanya mencakup analisis data, tetapi juga terintegrasi dengan pengelolaan risiko dan strategi pemasaran, sehingga mampu mendukung klien dalam meningkatkan profitabilitas serta efisiensi bisnis.
Perusahaan ini berpegang pada nilai inti CHAMPION, yang mencerminkan orientasi pada pelanggan, integritas, kelincahan, semangat membimbing, sikap proaktif, inovasi berkelanjutan, rasa tanggung jawab, serta pengambilan keputusan berbasis data.

Introduction

Project ini bertujuan memprediksi risiko kredit dengan mengklasifikasikan peminjam ke dalam kategori Good Risk dan Bad Risk menggunakan riwayat pembayaran serta analisis machine learning.

Research Problem

Risiko kredit yang salah kelola dapat memicu kredit macet dan kerugian finansial. Penilaian yang tidak akurat bisa menyebabkan peminjam berisiko tetap diberi pinjaman atau peminjam layak justru ditolak.

Research Objectives

- Membangun model klasifikasi berbasis machine learning untuk memprediksi risiko kredit.
- Membantu lembaga keuangan mengidentifikasi peminjam berisiko gagal bayar.
- Mengurangi potensi kerugian finansial.

DATA UNDERSTANDING

Data Set

- Format : CSV
- Jenis : Berskala nominal, kategori, dan numerik.
- Isi : Dataset pinjaman kredit tahun 2007-2014 meliputi berbagai informasi terkait profil peminjam, detail pinjaman, dan status pembayaran.
- Struktur : Big Data Set (75 kolom dan 466.285 baris)

Proses Data Understanding

1. Importing Dataset & Identifikasi Struktur datanya.
2. Transformasi Data Target.
3. Membersihkan Data (Missing Values)

Tujuan Analisis

- memprediksi risiko kredit peminjam berdasarkan data historis, sehingga
- Mengidentifikasi faktor-faktor utama yang memengaruhi kelayakan kredit.
- Mendukung pengambilan keputusan pinjaman yang lebih akurat dan efisien.

Results

- Memiliki 75 kolom (data numerik dan kategorik)
- Terdapat 5 kolom yang merekam informasi berbasis tanggal.
- Terdiri dari 466.285 baris.
- Memiliki total 40 kolom missing values dan 17 kolom empty data.

	...1	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grad
1	0	1077501	1296599	5000	5000	4975.00	36 months	10.65	162.87	B	B2
2	1	1077430	1314167	2500	2500	2500.00	60 months	15.27	59.83	C	C4
3	2	1077175	1313524	2400	2400	2400.00	36 months	15.96	84.33	C	C5
4	3	1076863	1277178	10000	10000	10000.00	36 months	13.49	339.31	C	C1
5	4	1075358	1311748	3000	3000	3000.00	60 months	12.69	67.79	B	B5
6	5	1075269	1311441	5000	5000	5000.00	36 months	7.90	156.46	A	A4
7	6	1069639	1304742	7000	7000	7000.00	60 months	15.96	170.08	C	C5
8	7	1072053	1288686	3000	3000	3000.00	36 months	18.64	109.43	E	E1
9	8	1071795	1306957	5600	5600	5600.00	60 months	21.28	152.39	F	F2
10	9	1071570	1306721	5375	5375	5350.00	60 months	12.69	121.45	B	B5
11	10	1070078	1305201	6500	6500	6500.00	60 months	14.65	153.45	C	C3
12	11	1069908	1305008	12000	12000	12000.00	36 months	12.69	402.54	B	B5
13	12	1064687	1298717	9000	9000	9000.00	36 months	13.49	305.38	C	C1
14	13	1069866	1304956	3000	3000	3000.00	36 months	9.91	96.68	B	B1
15	14	1069057	1303503	10000	10000	10000.00	36 months	10.65	325.74	B	B2
16	15	1069759	1304871	1000	1000	1000.00	36 months	16.29	35.31	D	D1
17	16	1065775	1299699	10000	10000	10000.00	36 months	15.27	347.98	C	C4
18	17	1069971	1304884	3600	3600	3600.00	36 months	6.03	109.57	A	A1
19	18	1062474	1294539	6000	6000	6000.00	36 months	11.71	198.46	B	B3
20	19	1069742	1304855	9200	9200	9200.00	36 months	6.03	280.01	A	A1
21	20	1069740	1284848	20250	20250	19142.16	60 months	15.27	484.63	C	C4
22	21	1039153	1269083	21000	21000	21000.00	36 months	12.42	701.73	B	B4

Showing 1 to 23 of 466,285 entries, 75 total columns

EXPLORATORY DATA ANALYSIS

1. Data Convertion

- Melakukan data cleansing dan konversi data kategorik.
- Menghilangkan atribut yang memiliki a lot of missing value atau empty value

	XA	XB	XC	XD	XE	XF	XG	XH	XI	XJ	XK	XL	XM	XN	XO	XP	XQ
1	11637209	13609401	5000	5000	1	13.98	170.84	2	10	5	36000.00	0	2	1	10.90	0	
2	1615368	1887150	10400	10400	1	10.16	336.37	1	5	4	45000.00	0	11	3	29.81	0	
3	5415051	6787225	8000	8000	1	7.90	250.33	0	3	1	39000.00	0	6	3	10.55	0	
4	5784068	7215828	28000	28000	0	19.05	727.11	3	18	1	91776.00	2	6	3	13.23	2	
5	3499170	4381874	20000	20000	0	22.47	557.74	4	24	1	80000.00	2	3	3	13.44	1	
6	1142411	1383169	10000	10000	1	6.62	307.04	0	1	5	125000.00	0	2	3	3.72	0	
7	19526062	21738819	30000	30000	0	18.24	765.73	3	19	4	152000.00	2	6	3	21.77	1	
8	5761979	7193822	6000	6000	1	14.33	206.03	2	10	1	64900.00	0	6	3	19.71	0	
9	10118195	11969918	14000	14000	0	15.61	337.56	2	13	1	89999.00	2	1	4	7.00	0	
10	1584173	1853389	23000	23000	0	21.00	622.23	4	21	1	100000.00	1	10	0	11.28	1	
11	10105783	11957871	32875	32875	0	18.25	839.29	3	17	1	74000.00	2	12	1	10.04	0	
12	8620136	10391876	35000	35000	0	23.70	1000.80	5	25	1	86000.00	1	11	1	26.41	0	
13	26909658	29392683	20000	20000	0	8.39	409.28	0	4	1	225000.00	1	10	3	4.94	1	
14	14620935	16683227	10000	10000	1	15.61	349.65	2	14	5	45000.00	1	4	1	23.97	2	
15	8990817	10772310	11000	11000	1	11.99	365.31	1	7	4	42000.00	1	11	1	16.94	0	
16	2217006	2629357	6000	6000	1	12.12	199.63	1	7	1	70000.00	2	12	3	13.34	0	
17	19057077	21259757	3850	3850	1	15.61	134.62	3	15	5	52000.00	0	7	1	15.28	0	
18	4524696	5786857	3600	3600	1	15.80	126.22	2	12	5	83000.00	0	4	1	3.33	0	
19	31257622	33830880	35000	35000	0	12.99	796.18	2	10	1	115000.00	2	10	3	20.32	0	
20	13056859	15089052	4000	4000	1	13.65	136.04	2	10	5	30000.00	0	3	3	13.72	0	
21	7062349	8723960	10000	10000	0	23.10	282.48	4	23	5	58000.00	1	9	1	10.63	3	
22	14819032	16891470	8000	8000	1	12.49	267.60	1	8	5	97000.00	1	4	1	29.07	0	

Showing 1 to 23 of 10,000 entries, 25 total columns

Before

	...1	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grad
1	0	1077501	1296599	5000	5000	4975.00	36 months	10.65	162.87	B	B2
2	1	1077430	1314167	2500	2500	2500.00	60 months	15.27	59.83	C	C4
3	2	1077175	1313524	2400	2400	2400.00	36 months	15.96	84.33	C	C5
4	3	1076863	1277178	10000	10000	10000.00	36 months	13.49	339.31	C	C1
5	4	1075358	1311748	3000	3000	3000.00	60 months	12.69	67.79	B	B5
6	5	1075269	1311441	5000	5000	5000.00	36 months	7.90	156.46	A	A4
7	6	1069639	1304742	7000	7000	7000.00	60 months	15.96	170.08	C	C5
8	7	1072053	1288686	3000	3000	3000.00	36 months	18.64	109.43	E	E1
9	8	1071795	1306957	5600	5600	5600.00	60 months	21.28	152.39	F	F2
10	9	1071570	1306721	5375	5375	5350.00	60 months	12.69	121.45	B	B5
11	10	1070078	1305201	6500	6500	6500.00	60 months	14.65	153.45	C	C3
12	11	1069908	1305008	12000	12000	12000.00	36 months	12.69	402.54	B	B5
13	12	1064687	1298717	9000	9000	9000.00	36 months	13.49	305.38	C	C1
14	13	1069866	1304956	3000	3000	3000.00	36 months	9.91	96.68	B	B1
15	14	1069057	1303503	10000	10000	10000.00	36 months	10.65	325.74	B	B2
16	15	1069759	1304871	1000	1000	1000.00	36 months	16.29	35.31	D	D1
17	16	1065775	1299699	10000	10000	10000.00	36 months	15.27	347.98	C	C4
18	17	1069971	1304884	3600	3600	3600.00	36 months	6.03	109.57	A	A1
19	18	1062474	1294539	6000	6000	6000.00	36 months	11.71	198.46	B	B3
20	19	1069742	1304855	9200	9200	9200.00	36 months	6.03	280.01	A	A1
21	20	1069740	1284848	20250	20250	19142.16	60 months	15.27	484.63	C	C4
22	21	1039153	1269083	21000	21000	21000.00	36 months	12.42	701.73	B	B4

Showing 1 to 23 of 466,285 entries, 75 total columns

After

2. Descriptive Statistics

Data menunjukkan bahwa sebanyak 239.071 kreditor dijadwalkan untuk melakukan pembayaran mereka pada tanggal jatuh tempo berikutnya.

Data menunjukkan bahwa mayoritas kreditor melakukan pembayaran terakhir mereka pada bulan Januari 2016. Hal ini mengindikasikan bahwa sebagian besar pinjaman memiliki aktivitas pembayaran terakhir yang terkonsentrasi di bulan yang sama.

2. Descriptive Statistics

```
> prop.table(table(sample_df$XN))

          0           1           2           3           4           5           6
0.096585759 0.479907342 0.001611441 0.395004532 0.007956491 0.002517877 0.016416558
```

- 0 = Charged Off → pinjaman gagal bayar / dihapuskan.
- 1 = Current → pinjaman masih berjalan lancar.
- 2 = Default → peminjam benar-benar gagal membayar.
- 3 = Fully Paid → pinjaman sudah lunas.
- 4 = In Grace Period → peminjam masih dalam masa tenggang pembayaran.
- 5 = Late (16-30 days) → keterlambatan antara 16-30 hari.
- 6 = Late (31-120 days) → keterlambatan antara 31-120 hari.

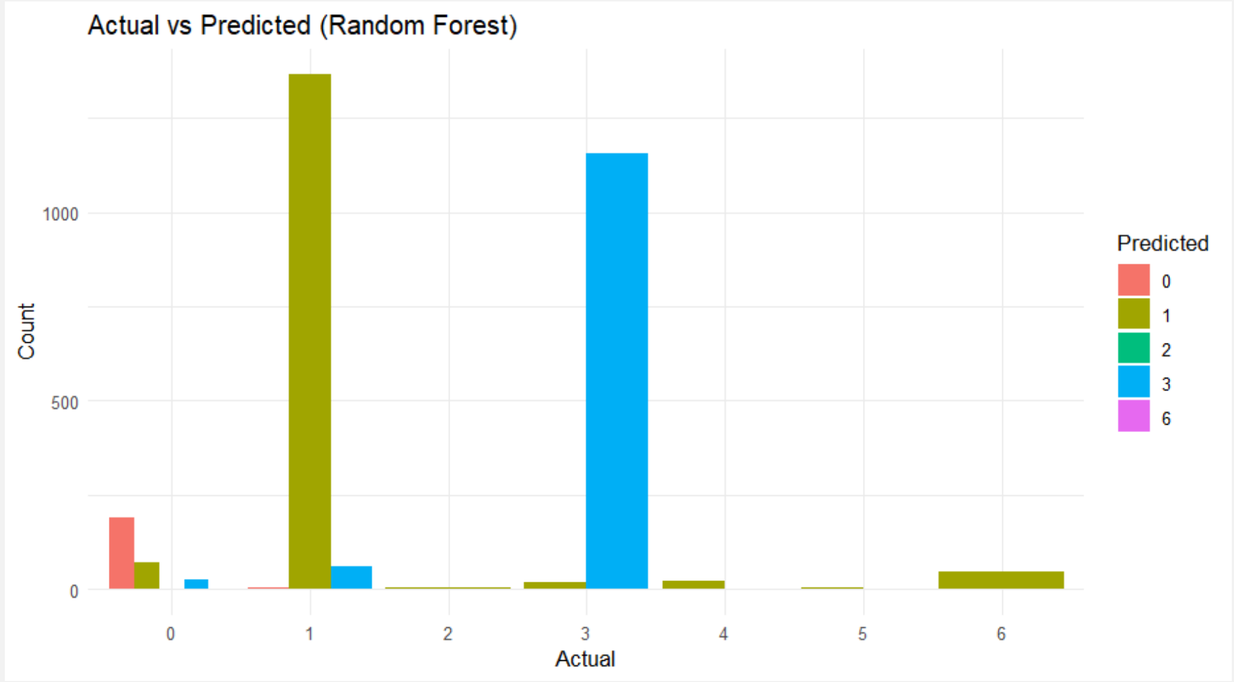
proporsi tiap kategori terhadap total data:

- Current (1) → 47.99% (paling banyak, artinya mayoritas pinjaman masih berjalan lancar).
- Fully Paid (3) → 39.50% (cukup besar, menunjukkan banyak pinjaman yang sudah lunas).
- Charged Off (0) → 9.66% (cukup signifikan, menunjukkan kredit macet yang dihapuskan).
- Late (31-120 days) (6) → 1.64% (terdapat peminjam yang menunggak lama).
- In Grace Period (4) → 0.80% (masih dalam masa tenggang).
- Late (16-30 days) (5) → 0.25% (terlambat singkat).
- Default (2) → 0.16% (sangat kecil, sudah default resmi).

2. Descriptive Statistics

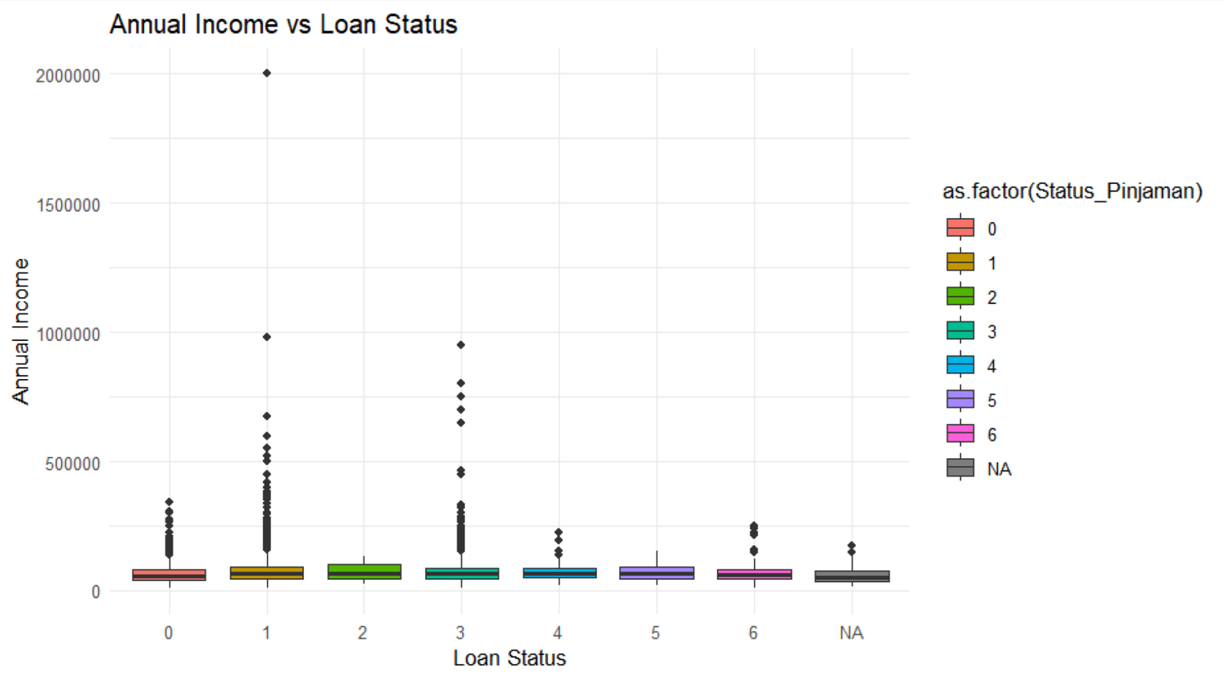
Uji Beda Rata-Rata pendapatan antar status pinjaman

Hasil uji beda rata-rata dengan ANOVA menunjukkan bahwa terdapat perbedaan yang signifikan pada rata-rata Pendapatan Tahunan antar kelompok Status Pinjaman ($F = 4.525$, $p < 0.001$). Artinya, pendapatan tahunan nasabah bervariasi secara signifikan tergantung pada status pinjamannya. Sebanyak 71 observasi dikeluarkan dari analisis karena data tidak lengkap.



2. Descriptive Statistics

BoxPlot Income dan Loan Status

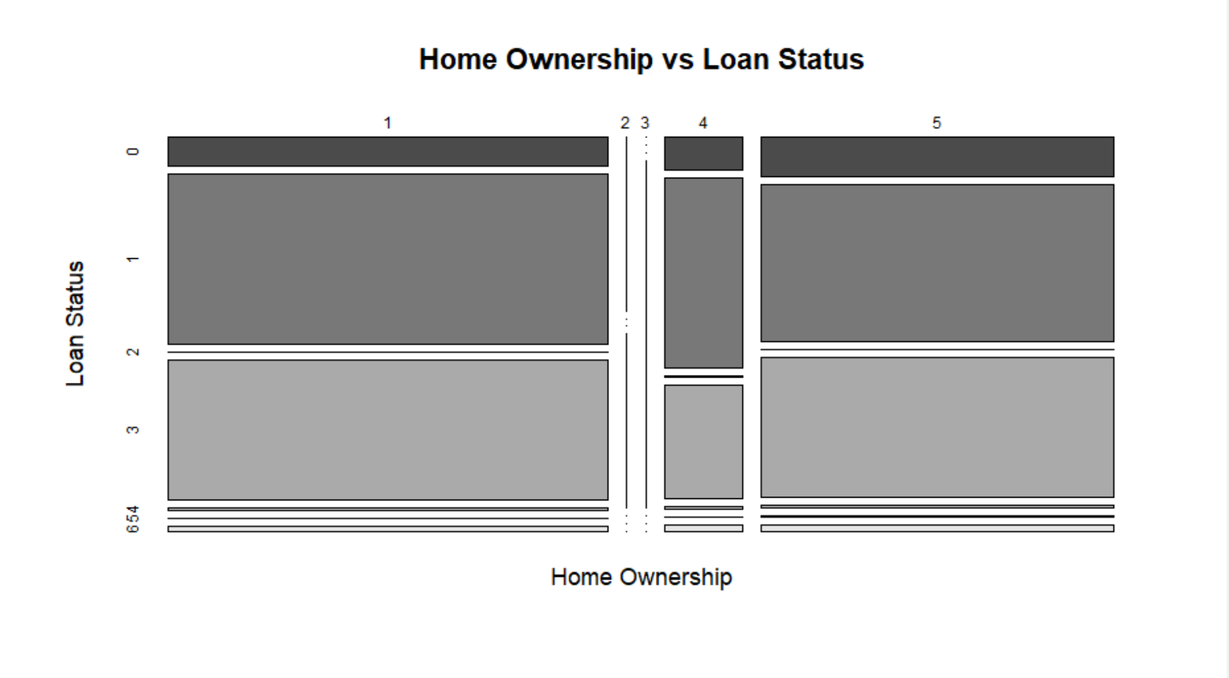


- Median pendapatan tahunan relatif mirip di hampir semua kategori status pinjaman, meskipun ada sedikit variasi.
- Terlihat banyak outlier (pendapatan sangat tinggi), khususnya pada status pinjaman 0, 1, dan 3, yang mengindikasikan sebagian kecil peminjam memiliki pendapatan jauh di atas rata-rata.
- Sebagian besar data terkonsentrasi pada pendapatan tahunan di bawah 200.000.

2. Descriptive Statistics

Asosiasi antara Status Kepemilikan Rumah dengan Status Pinjaman

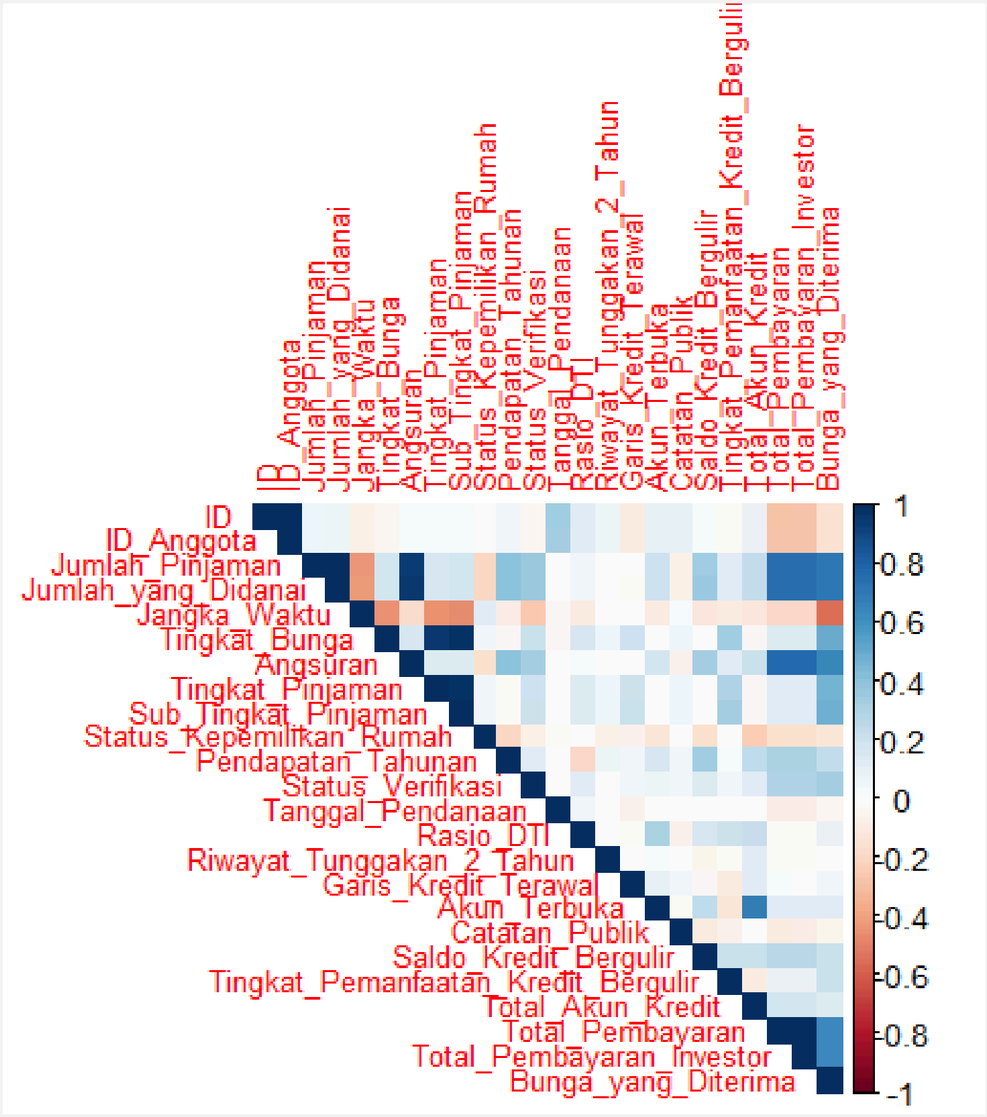
- Pemilik rumah tertentu (kategori 1 dan 5) lebih banyak terasosiasi dengan status pinjaman aktif maupun selesai (loan status menengah).
- Sementara kategori kepemilikan lain (misalnya 2, 3, dan 4) memiliki proporsi lebih kecil dalam semua status pinjaman.
- Hal ini menunjukkan bahwa kepemilikan rumah yang lebih stabil cenderung terkait dengan distribusi pinjaman yang lebih jelas dan besar, dibandingkan kategori kepemilikan rumah lain yang jumlahnya relatif sedikit.



2. Descriptive Statistics

HeatMap

- Positif Kuat: Jumlah Pinjaman, Jumlah yang Didanai, Angsuran, dan variabel total pembayaran saling berhubungan erat.
- Negatif Kuat: Tingkat Pinjaman berkorelasi negatif kuat dengan Tingkat Bunga. Artinya, grade pinjaman yang lebih baik memiliki bunga yang lebih rendah.
- Lemah: Sebagian besar variabel lain, seperti Pendapatan Tahunan dan Riwayat Tunggakan, menunjukkan korelasi yang sangat lemah atau tidak ada.



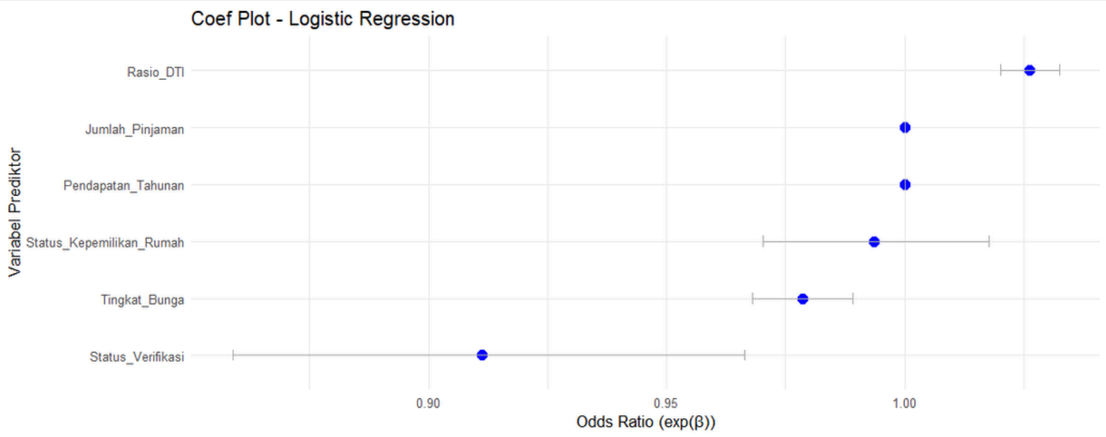
DATA MODELLING

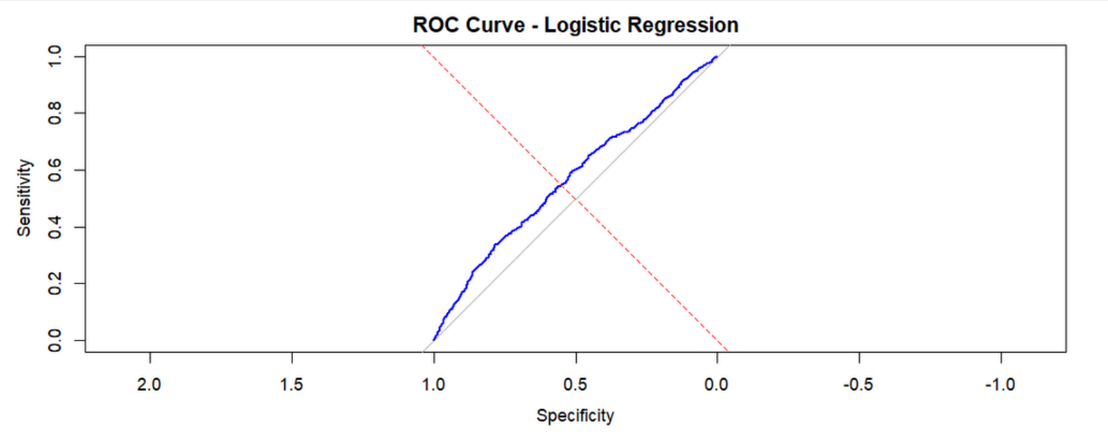
1. Train & Test Data

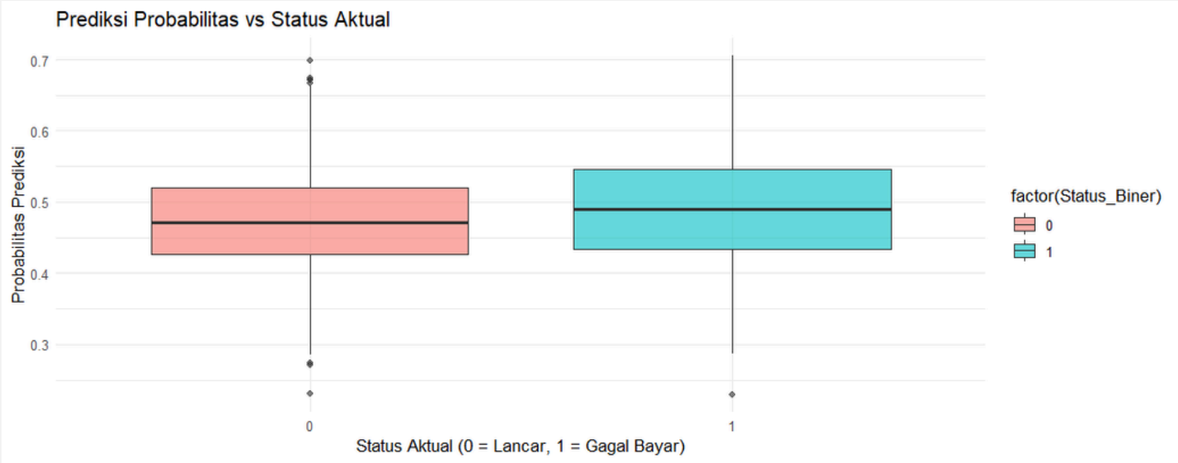
```
> # 2. Ringkasan jumlah data
> cat("Jumlah data total :", nrow(sample_dff), "\n")
Jumlah data total : 10000
> cat("Jumlah data train :", nrow(train_df), "\n")
Jumlah data train : 7005
> cat("Jumlah data test  :", nrow(test_df), "\n\n")
Jumlah data test  : 2995

> # 3. Distribusi kelas
> cat("Distribusi Status Pinjaman (Train):\n")
Distribusi Status Pinjaman (Train):
> print(prop.table(table(train_df$Status_Pinjaman)))

          0           1           2           3           4           5           6
0.096621136 0.479654925 0.001725377 0.394823868 0.008051761 0.002588066 0.016534867
> cat("\nDistribusi Status Pinjaman (Test):\n")

Distribusi Status Pinjaman (Test):
> print(prop.table(table(test_df$Status_Pinjaman)))

          0           1           2           3           4           5           6
0.096503026 0.480497646 0.001344990 0.395427034 0.007733692 0.002353732 0.016139879
```

- Kelas 0: Sekitar 9,6% dari total data.
- Kelas 1: Kelas yang paling dominan, mencakup hampir 48% dari total data.
- Kelas 2, 4, 5, 6: Masing-masing memiliki persentase yang sangat kecil, kurang dari 2%.
- Kelas 3: Sekitar 39,4% dari total data, menjadikannya kelas terbesar kedua.

2. REGRESION


Coef Plot - Logistic Regression
Rasio_DTI
Jumlah_Pinjaman
Pendapatan_Tahunan
Status_Kepemilikan_Rumah
Tingkat_Bunga
Status_Verifikasi
Variabel Prediktor
0.90
0.95
1.00
Odds Ratio (exp(β))


ROC Curve - Logistic Regression
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
Specificity


Prediksi Probabilitas vs Status Aktual
Probabilitas Prediksi
0.3
0.4
0.5
0.6
0.7
0
1
Status Aktual (0 = Lancar, 1 = Gagal Bayar)
factor(Status_Biner)
0
1

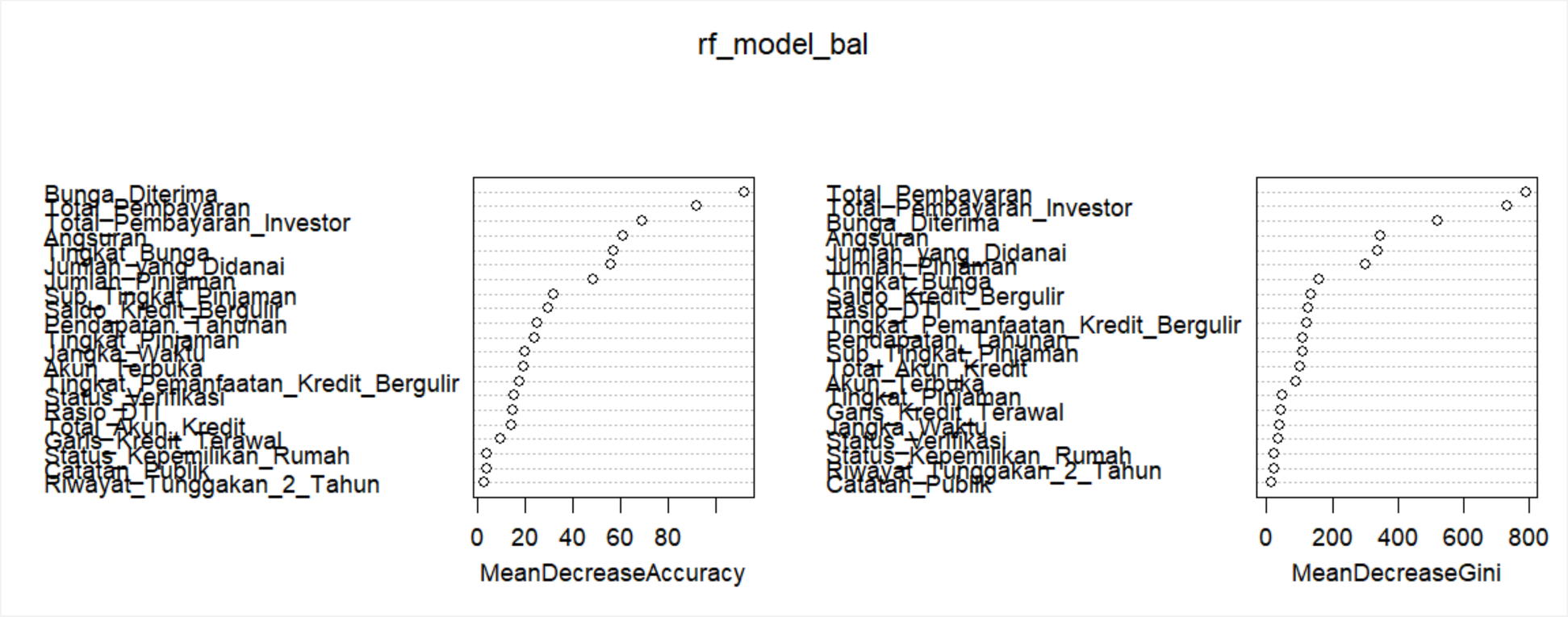
2. REGRESION

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 650 534
         1 366 435

               Accuracy : 0.5466
                 95% CI : (0.5244, 0.5687)
    No Information Rate : 0.5118
    P-Value [Acc > NIR] : 0.00104

                  Kappa : 0.089

 Mcnemar's Test P-Value : 2.597e-08

            Sensitivity : 0.4489
            Specificity : 0.6398
         Pos Pred Value : 0.5431
         Neg Pred Value : 0.5490
             Prevalence : 0.4882
         Detection Rate : 0.2191
   Detection Prevalence : 0.4035
      Balanced Accuracy : 0.5443

       'Positive' Class : 1
```

```
=== Confusion Matrix ===
          Reference
Prediction   0   1
         0 650 534
         1 366 435
```

- Accuracy: 54.7% (hanya sedikit di atas baseline 51.1%).
- Kappa: 0.089 → kesesuaian prediksi sangat lemah.
- Sensitivity (Recall): 44.9% → banyak gagal bayar tidak terdeteksi.
- Specificity: 63.9% → lebih baik mendeteksi nasabah lancar.
- Precision Positif: 54.3% → prediksi gagal bayar masih sering salah.
- Balanced Accuracy: 54.4% → hampir setara dengan tebakan acak.
- McNemar's Test signifikan → ada bias prediksi ke salah satu kelas.

3. RANDOM FOREST



3. RANDOM FOREST

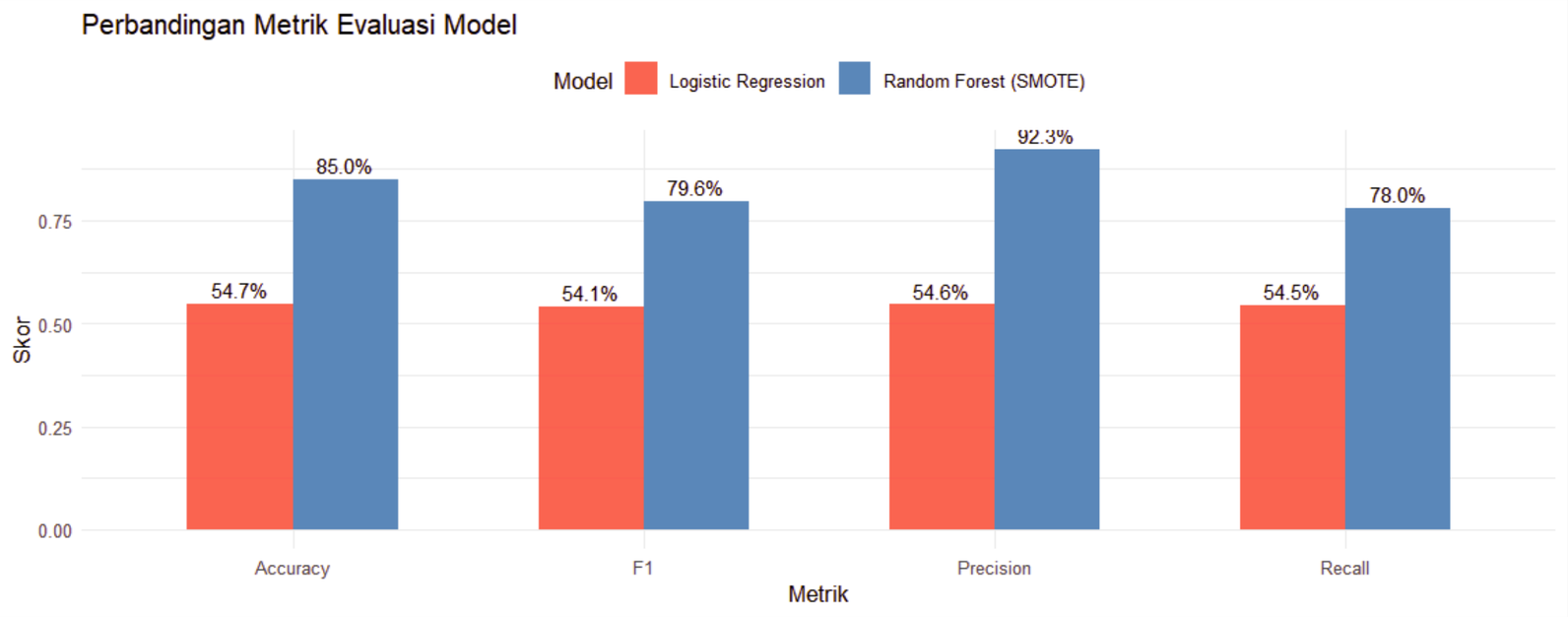
```
Precision : 0.9232
Recall    : 0.7802
F2-Score  : 0.796
> accuracy
[1] 0.8503778
```

- Model Random Forest dengan SMOTE menghasilkan performa sangat baik untuk klasifikasi status pinjaman.
- Dengan Precision tinggi (92%), model aman dari false positive, dan Recall 78% cukup baik untuk menangkap risiko.
- Secara praktis, lender bisa menggunakan model ini untuk mengurangi kerugian akibat pinjaman gagal bayar, meskipun masih ada sekitar 22% risiko gagal bayar yang tidak terdeteksi.

EVALUATION

MODEL EVALUATION

AUC = 0.5683999



- Accuracy: Logistic (54.7%) jauh di bawah Random Forest (85%).
- F1-Score: Logistic (54.1%) vs Random Forest (79.6%) → RF lebih seimbang.
-

- Accuracy: Logistic (54.7%) jauh di bawah Precision: Logistic (54.6%) vs Random Forest (92.3%) → RF lebih tepat dalam memprediksi gagal bayar.
- Recall: Logistic (54.5%) vs Random Forest (78%) → RF lebih banyak mendeteksi kasus gagal bayar.

CONCLUSION

Hasil

- Berdasarkan EDA, terlihat adanya peningkatan jumlah pinjaman bermasalah pada tahun 2014.
- Sebagian besar pinjaman yang macet mengalami keterlambatan dalam rentang 10–30 bulan.
- Hasil uji ANOVA menunjukkan bahwa usia kredit memiliki pengaruh signifikan terhadap status pinjaman buruk (berpotensi macet).

Model Terbaik

- Berdasarkan evaluasi, Random Forest dengan SMOTE memberikan hasil terbaik dengan akurasi 85%, precision 92%, recall 78%, dan F1 79%.
- Dibandingkan Logistic Regression, Random Forest jauh lebih baik dalam mendeteksi pinjaman berisiko tinggi.
- Jika tujuan utama adalah meminimalkan risiko salah klasifikasi (False Negatives), Random Forest lebih disarankan.
- Model ini efektif untuk mendukung keputusan kredit agar lebih tepat dan mengurangi potensi gagal bayar.

THANK YOU

I appreciate the opportunity to present this project. I am confident that my skills in Data Analysis, and I am available to discuss this further.

Git Hub : https://github.com/sabillafarah/Loan-Classification.git

File Project :
https://drive.google.com/drive/folders/1kxmF8qVh26merO0S-uBcrjGhhztqZRIn?usp=drive_link

081252501012

sabillafarannisa@gmai.com

@Sabilla Farah Annisa