

Laporan Analisis Data Semifinal STC Logika UI 2024

Nomor tim : **24-03-017-9**

Toby Purbojo

Joseph Hansel

1 Latar Belakang

STC Paylater merupakan suatu perusahaan BNPL (Buy Now, Pay Later) yang memperoleh pendapatan dari meminjamkan uang kepada nasabahnya. Salah satu risiko signifikan dalam proses bisnis tersebut adalah kemungkinan peminjam gagal bayar, yaitu dengan berhenti melakukan pembayaran seperti yang telah disepakati. Hal ini tentu saja akan menyebabkan kerugian finansial bagi perusahaan. Untuk mengurangi kemungkinan kerugian tersebut, sangat penting bagi STC Paylater untuk membuat keputusan mengenai siapa yang akan diberi pinjaman, suku bunga yang dikenakan dan besar pinjaman yang disetujui. Pengambilan keputusan tersebut dibantu dengan bantuan pembelajaran mesin yang mampu mempelajari pola dan karakteristik dari data yang diberikan.

Pengambilan keputusan tersebut merupakan suatu masalah kompleks yang memerlukan bantuan suatu alat yang mutakhir, salah satunya dengan bantuan pembelajaran mesin. Salah satu model pembelajaran yang digunakan adalah model *weighted k-means clustering*. Model tersebut memanfaatkan fitur-fitur nasabah sebelumnya untuk melakukan pengelompokan nasabah menjadi beberapa kelompok yang memiliki karakteristik serupa. Kelompok tersebut kemudian akan digunakan sebagai label atau variabel terikat dalam model pembelajaran mesin lainnya. Dengan adanya variabel terikat, model pembelajaran mesin dapat dilatih untuk menghasilkan model yang optimal dalam membantu proses pengambilan keputusan tersebut.

1.1 *Business Inquiries*

Dengan menggunakan model yang telah dibuat, diharapkan dapat membantu perusahaan dalam melakukan penilaian risiko seorang calon nasabah. Risiko yang dimiliki nasabah dapat digunakan dalam proses pengambilan keputusan, apakah nasabah tersebut layak diberikan pinjaman uang atau tidak.

2 Penjelasan Data

Terdapat empat buah set data yang disediakan oleh STC Paylater. Data yang disediakan tersebut meliputi *payment history*, *previous applications*, dan *loan application data (train/test)*. Data *payment history* berisi riwayat pembayaran (*installment*) nasabah untuk suatu periode tertentu. Data *previous applications* berisi fitur-fitur nasabah yang pernah melakukan peminjaman uang kepada STC Paylater. Kedua data tersebut dimodifikasi dan diolah agar dapat menghasilkan label (klaster) risiko peminjaman uang. Label tersebut akan digunakan pada data ketiga yaitu *loan application data*. Data tersebut berisi fitur-fitur calon nasabah yang ingin meminjam uang di STC Paylater. Dengan label yang telah ada, data *loan application* dapat dilakukan proses pemodelan untuk menghasilkan model optimal yang dapat memprediksi risiko dari calon-calon nasabah yang ada.

3 Eksplorasi Data, Pembersihan Data, dan Modifikasi Fitur

3.1 Eksplorasi dan Pemrosesan Data *Payment History*

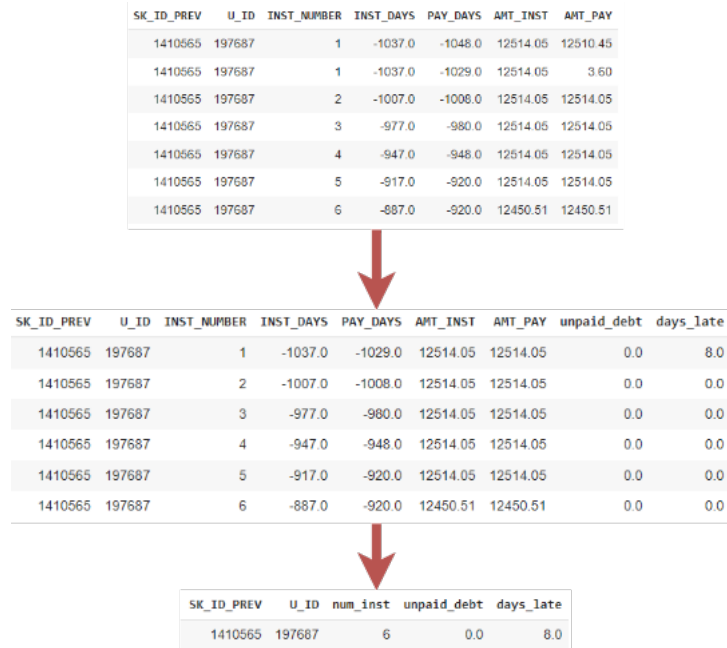
Dataset *Payment history* memiliki ukuran sebesar (2871633, 7) atau 2.871.633 observasi dan 7 variabel. Deskripsi data tersebut secara statistik diberikan pada Gambar 1. Berdasarkan gambar tersebut dan kode program, ditemukan kejanggalan pada dataset yaitu sebagai berikut.

1. AMT_INST terdapat sebanyak 64 observasi dengan nilai 0. Meskipun tagihan bernilai nol, observasi-observasi tersebut tetap dilakukan pembayaran peminjaman. 64 observasi dengan nilai nol pada tagihan akan diasumsikan lunas tagihan.
2. AMT_PAY juga terdapat nilai minimum 0. Nilai nol, Setelah ditelusuri lebih lanjut, sebuah pembayaran *installment* dapat dilakukan lebih dari satu kali.
3. Terdapat sebanyak 673 data kosong pada variabel PAY_DAYS dan 673 pada variabel AMT_PAY. Setelah ditelusuri lebih lanjut, nilai *Nan* ditemukan hanya pada pembayaran terakhir dari nasabah. Selain itu, data kosong dari variabel PAY_DAYS dan AMT_PAY ada pada observasi yang sama. Seluruh observasi dengan nilai kosong pada variabel tersebut akan dihapus karena jumlah yang tidak signifikan.

	SK_ID_PREV	U_ID	INST_NUMBER	INST_DAYS	PAY_DAYS	AMT_INST	AMT_PAY
count	2.872306e+06	2.872306e+06	2.872306e+06	2.872306e+06	2.871633e+06	2.872306e+06	2.871633e+06
mean	1.902798e+06	2.785208e+05	1.865887e+01	-1.039830e+03	-1.048684e+03	1.692881e+04	1.708792e+04
std	5.358735e+05	1.026814e+05	2.635638e+01	7.995411e+02	7.991129e+02	5.010468e+04	5.422172e+04
min	1.000020e+06	1.000090e+05	1.000000e+00	-2.922000e+03	-3.129000e+03	0.000000e+00	0.000000e+00
25%	1.435627e+06	1.893100e+05	4.000000e+00	-1.651000e+03	-1.659000e+03	4.199850e+03	3.389490e+03
50%	1.894453e+06	2.786890e+05	8.000000e+00	-8.170000e+02	-8.260000e+02	8.787330e+03	8.095050e+03
75%	2.368624e+06	3.675770e+05	1.900000e+01	-3.580000e+02	-3.670000e+02	1.661709e+04	1.597320e+04
max	2.843498e+06	4.562550e+05	2.250000e+02	-2.000000e+00	-2.000000e+00	3.371884e+06	3.371884e+06

Gambar 1: Deskripsi data *Payment History*

Berdasarkan penemuan sebelumnya, karena ditemukan bahwa sebuah *installment* dapat dibayar lebih dari sekali, maka akan dilakukan agregasi data berdasarkan pengelompokan variabel SK_ID_PREV, U_ID, dan AMT_INST. Kemudian akan dibuat dua variabel baru yaitu UNPAID_DEBT dan DAYS_LATE. Variabel UNPAID_DEBT merupakan hasil selisih antara tagihan pinjaman dengan pembayaran sebuah *installment*. Sementara, variabel DAYS_LATE adalah selisih waktu jatuh tempo *installment* dengan waktu pembayaran tagihan. Terakhir, dilakukan agregasi untuk menjumlahkan setiap *installment*. Contoh proses yang dilakukan diberikan pada Gambar 2.



The diagram illustrates the aggregation process in three stages, connected by downward arrows:

- Initial Data Table:** Contains columns SK_ID_PREV, U_ID, INST_NUMBER, INST_DAYS, PAY_DAYS, AMT_INST, and AMT_PAY. It lists 6 installment records for SK_ID_PREV 1410565 and U_ID 197687.
- Intermediate Table:** Adds columns unpaid_debt and days_late. The unpaid_debt is calculated as AMT_PAY minus AMT_INST, and days_late is calculated as INST_DAYS minus PAY_DAYS.
- Aggregated Table:** Shows the final summary with columns SK_ID_PREV, U_ID, num_inst (sum of INST_NUMBER), unpaid_debt, and days_late. The values are 1410565, 197687, 6, 0.0, and 8.0 respectively.

SK_ID_PREV	U_ID	INST_NUMBER	INST_DAYS	PAY_DAYS	AMT_INST	AMT_PAY
1410565	197687	1	-1037.0	-1048.0	12514.05	12510.45
1410565	197687	1	-1037.0	-1029.0	12514.05	3.60
1410565	197687	2	-1007.0	-1006.0	12514.05	12514.05
1410565	197687	3	-977.0	-980.0	12514.05	12514.05
1410565	197687	4	-947.0	-948.0	12514.05	12514.05
1410565	197687	5	-917.0	-920.0	12514.05	12514.05
1410565	197687	6	-887.0	-920.0	12450.51	12450.51

SK_ID_PREV	U_ID	INST_NUMBER	INST_DAYS	PAY_DAYS	AMT_INST	AMT_PAY	unpaid_debt	days_late
1410565	197687	1	-1037.0	-1029.0	12514.05	12514.05	0.0	8.0
1410565	197687	2	-1007.0	-1006.0	12514.05	12514.05	0.0	0.0
1410565	197687	3	-977.0	-980.0	12514.05	12514.05	0.0	0.0
1410565	197687	4	-947.0	-948.0	12514.05	12514.05	0.0	0.0
1410565	197687	5	-917.0	-920.0	12514.05	12514.05	0.0	0.0
1410565	197687	6	-887.0	-920.0	12450.51	12450.51	0.0	0.0

SK_ID_PREV	U_ID	num_inst	unpaid_debt	days_late
1410565	197687	6	0.0	8.0

Gambar 2: Contoh pembayaran ID 1410565 sebelum dan sesudah dilakukan pemrosesan.

Hal yang perlu diingat adalah sebuah nomor unik U_ID dapat terdiri atas sejumlah nomor SK_ID_PREV unik. Hal ini penting diingat karena dataset *previous applications* tidak memiliki variabel SK_ID_PREV.

3.2 Eksplorasi dan Pemrosesan Data *Previous Applications*

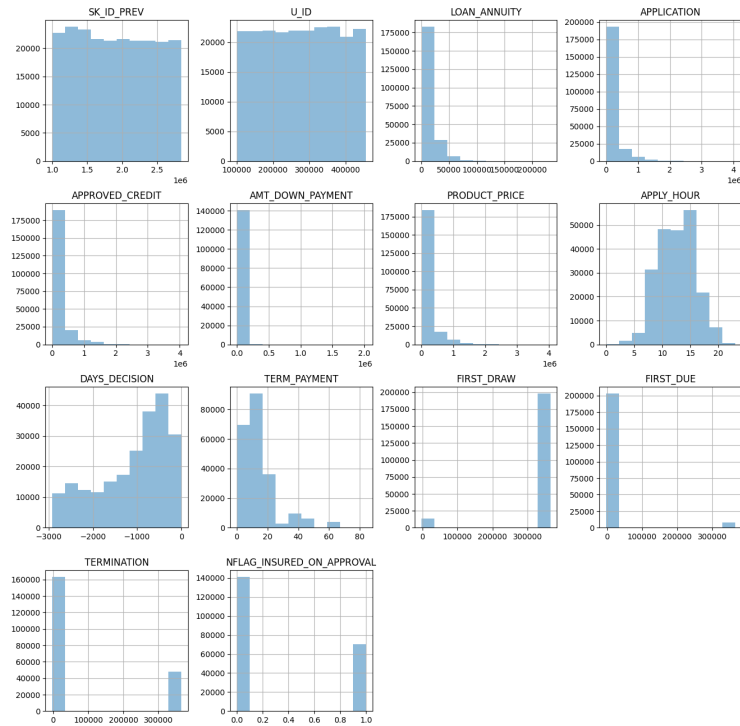
Data *Previous Applications* memiliki sebanyak 350.712 observasi dan 18 variabel bebas. Pertama, data tersebut akan disaring berdasarkan CONTRACT_STATUS berdasarkan nilai *approved*. Hal ini diketahui dari nomor unik SK_ID_PREV pada dataset *payment history*, hanya ditemukan pada data ini saat CONTRACT_STATUS bernilai *approved*. Ukuran data setelah disaring adalah 219.687 observasi dan 18 variabel.

Berdasarkan data ini, terdapat beberapa variabel yang diduga memiliki peran penting untuk pelabelan data. Berikut adalah beberapa pertimbangan variabel serta penjelasannya.

1. YIELD_GROUP merupakan variabel yang menyatakan tingkat bunga yang diberikan kepada nasabah oleh STC PayLater. Besar bunga yang diberikan dapat mencerminkan *credit score* seseorang.
2. Variabel NFLAG_INSURED_ON_APPROVAL menyatakan apakah sebuah pinjaman diasuransikan atau tidak. Sebuah pinjaman yang diasuransikan memberi jaminan kepada pihak STC PayLater, sehingga pinjaman nasabah dapat terjamin meskipun mengalami kejadian tidak terduga seperti kecelakaan.
3. Variabel PRODUCT_PRICE menyatakan harga produk yang akan dibayar menggunakan pinjaman. Variabel ini sendiri tidak memberikan informasi yang signifikan terhadap pelabelan data. Namun, *feature engineering* dengan variabel APPROVED_CREDIT dapat memberikan informasi yang penting.

Berdasarkan kode program, ditemukan sebanyak 8.280 observasi dengan nilai kosong pada variabel NFLAG_INSURED_ON_APPROVAL dan 9.222 data kosong pada variabel PRODUCT_PRICE. Selain itu, variabel APPROVED_CREDIT dan PRODUCT_PRICE terdapat nilai nol. Berdasarkan hal tersebut, tindakan yang dilakukan untuk penemuan tersebut adalah seperti berikut.

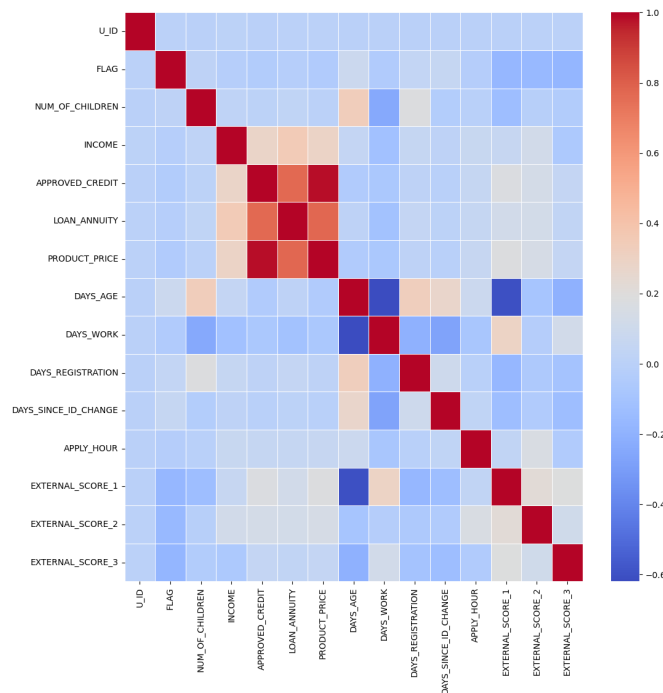
1. Nilai kosong pada variabel NFLAG_INSURED_ON_APPROVAL akan diisi dengan nilai terbanyak. Berdasarkan Gambar 3, nilai kosong akan diisi dengan 0 atau pinjaman tidak diasuransikan.
2. Dibuat variabel baru yang disebut dengan CREDIT_RATIO, variabel tersebut merupakan pembagian APPROVED_CREDIT dengan PRODUCT_PRICE. Variabel ini memberi interpretasi mengenai seberapa besar rasio pinjaman diberikan oleh bank untuk suatu harga produk yang ingin dibeli oleh nasabah. Nilai NA, infinity, dan nol pada variabel ini diisi dengan nilai 1,
3. Variabel YIELD_GROUP memiliki nilai unik *high*, *middle*, *low_normal*, *low_action*, dan NA1. Barisan dengan nilai NA1 digantikan dengan nilai terbanyak yaitu *middle*.



Gambar 3: Persebaran setiap variabel *previous applications*

3.3 Eksplorasi dan Pemrosesan Data *Loan Application Data* (Train)

Loan application data (train) memiliki 61.503 observasi data dan 23 variabel bebas. Korelasi plot dari variabel-variabel tersebut dapat dilihat pada Gambar 4.

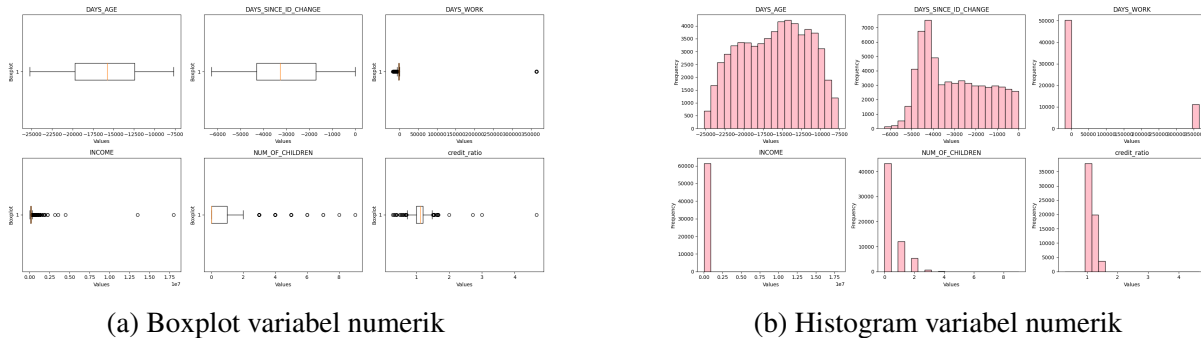


Gambar 4: Korelasi antar-variabel

Dapat dilihat bahwa variabel APPROVED_CREDIT, PRODUCT_PRICE, dan LOAN_ANNUITY memiliki korelasi yang cukup tinggi antara ketiganya. Dengan melakukan modifikasi yang serupa dengan sebelumnya, dibuat suatu variabel baru yaitu credit_score yang merupakan pembagian antara APPROVED_CREDIT dengan PRODUCT_PRICE. Selanjutnya, variabel LOAN_ANNUITY, APPLY_HOUR, DAYS_REGISTRATION, EXTERNAL_SCORE_1, EXTERNAL_SCORE_2, dan EXTERNAL_SCORE_3 dihilangkan dari data pelatihan karena dianggap tidak memiliki interpretasi yang bermanfaat dalam proses pengolahan data. Selain itu, variabel APPROVED_CREDIT dan PRODUCT_PRICE juga dihilangkan karena telah direpresentasikan dengan variabel credit_score. Dengan demikian, hanya tersisa 16 variabel dalam data pelatihan, yaitu 8 variabel numerik dan 8 variabel kategorik.

3.3.1 Variabel Numerik

Pemrosesan variabel data pelatihan dibagi menjadi dua bagian, yaitu pemrosesan variabel numerik dan pemrosesan variabel kategorik. Dapat dilihat pada Gambar 4 bahwa variabel FLAG mengalami misklasifikasi tipe data. Variabel FLAG yang menunjukkan apakah nasabah mengalami telat bayar atau tidak diklasifikasikan sebagai variabel numerik. Seharusnya, variabel FLAG merupakan variabel kategorik dengan nilai 1 yang berarti nasabah telat membayar lebih dari X hari dan 0 lainnya. Oleh sebab itu, dilakukan proses pengubahan tipe data dari *object* menjadi *number*. Boxplot dan histogram dari variabel numerik tanpa melibatkan U_ID dapat dilihat pada Gambar 5.



Gambar 5: Boxplot dan histogram variabel numerik

Selanjutnya, untuk mempermudah interpretasi, variabel yang berhubungan dengan hari yaitu DAYS_AGE, DAYS_WORK, dan DAYS_SINCE_ID_CHANGE dibagi dengan -365 agar perhitungannya dilakukan per tahun dan mengubahnya menjadi nilai positif. Variabel-variabel baru yaitu YEARS_AGE, YEARS_WORK, serta YEARS_SINCE_ID_CHANGE digunakan untuk menggantikan variabel yang lama.

3.3.2 Variabel Kategorik

Setelah selesai memproses variabel numerik, dilakukan proses modifikasi terhadap variabel kategorik. Variabel ORGANIZATION_CATEGORY yang merepresentasikan jenis organisasi

tempat nasabah bekerja dan APPLY_DAYS yang merepresentasikan hari ketika nasabah mengajukan peminjaman dihilangkan karena dianggap tidak memiliki pengaruh yang signifikan terhadap data. Banyaknya nilai unik dari variabel-variabel yang tersisa dapat dilihat pada Tabel 1.

Tabel 1: Nilai unik dari variabel kategorik

Variabel Kategorik	Nilai Unik
FLAG	2
CONTRACT_TYPE	2
GENDER	2
INCOME_CATEGORY	7
EDUCATION	5
FAMILY_STATUS	5
HOUSING_CATEGORY	6

Agar dapat diterapkan ke dalam model, variabel kategorik yang memiliki nilai unik lebih dari dua dilakukan proses *one-hot encoding*, sedangkan variabel kategorik dengan nilai unik dua dilakukan proses *label encoding*. Hasil dari *label encoding* yaitu

- FLAG (1: nasabah yang telat membayar lebih dari X hari, 0: lainnya);
- CONTRACT_TYPE (1: Revolving loans, 0: Cash loans); dan
- GENDER (1: Female, 0: Male).

Setelah dilakukan proses *encoding*, variabel numerik dan variabel kategorik digabungkan kembali untuk menjadi suatu data latih yang utuh.

3.3.3 Data Latih

Apabila dilakukan peninjauan lebih lanjut, dapat dilihat bahwa terdapat sejumlah observasi yang memiliki YEARS_WORK negatif. Hal tersebut tentu saja tidak masuk akal, karena YEARS_WORK merepresentasikan lamanya calon nasabah bekerja ketika mengajukan peminjaman uang. Ada sebanyak 11.253 observasi yang memiliki YEARS_WORK negatif. Untuk mengatasi hal tersebut, observasi yang memiliki YEARS_WORK negatif dihilangkan.

Berikutnya, dapat dilihat dari Gambar 5 bahwa variabel INCOME memiliki pencilan yang dapat mengganggu proses pemodelan data. Dengan demikian, observasi yang memiliki variabel INCOME lebih besar daripada 500.000 dihapus dengan harapan dapat membantu mengurangi pencilan yang dapat merusak model. Setelah seluruh variabel numerik diproses, dilakukan proses normalisasi variabel, yaitu proses mengubah nilai-nilai variabel numerik dalam suatu dataset agar memiliki skala yang seragam atau normal. Hal tersebut dilakukan dengan tujuan mengurangi sensitivitas terhadap nilai pencilan yang lain dan menghindari dominasi variabel yang memiliki nilai besar. Terakhir, dilakukan pengecekan terhadap nilai kosong dalam data latih. Karena sudah tidak ada nilai kosong, maka dapat dilanjutkan dengan proses berikutnya.

4 *Pre-processing* Data untuk Pemodelan

4.1 Penggabungan Data

Untuk memulai proses klasterisasi dilakukan, data *payment history* dan *previous applications* digabungkan berdasarkan variabel SK_ID_PREV. Penggabungan dilakukan agar mendapatkan *dataframe* baru dengan variabel-variabel yang lebih lengkap untuk mendasari proses klasterisasi. Dengan *dataframe* yg baru, dilakukan pengecekan terhadap nilai kosong. Sebanyak 8.095 observasi dibuang dari *dataframe* karena mengandung nilai kosong di dalamnya. Dengan demikian, *dataframe* yang baru terdiri dari 211.711 observasi dan 8 variabel, yaitu SK_ID_PREV, U_ID, NUM_INST, UNPAID_DEBT, DAYS_LATE, CREDIT_RATIO, NFLAG_INSURED_ON_APPROVAL, dan YIELD_GROUP.

Selanjutnya, dilakukan modifikasi terhadap beberapa variabel yang ada di dalam *dataframe*. Modifikasi pertama diterapkan terhadap variabel DAYS_LATE yang menunjukkan jumlah hari seorang nasabah telat membayar. Proses modifikasi variabel serupa dengan yang telah dilakukan sebelumnya, di mana variabel DAYS_LATE dibagi dengan 365 agar perhitungannya dilakukan per tahun. Nilai yang telah dibagi dimasukkan ke dalam kolom baru yaitu YEARS_LATE. Berikutnya, observasi dengan nilai YEARS_LATE yang lebih dari 6 tahun dihilangkan karena dianggap sebagai kasus ekstrem (pencilan) yang dapat merusak model. Modifikasi kedua diterapkan terhadap variabel UNPAID_DEBT. Sebelumnya, variabel UNPAID_DEBT merupakan variabel numerik yang merepresentasikan besar nominal seorang nasabah belum membayar hutangnya. Karena hanya sedikit sekali nasabah yang masih memiliki hutang, maka variabel UNPAID_DEBT diubah menjadi variabel kategorik yang hanya menunjukkan apakah nasabah tersebut masih memiliki hutang atau tidak. Terakhir, dilakukan proses label *encoding* terhadap variabel YIELD_GROUP yang merepresentasikan tingkat bunga yang harus ditanggung nasabah. Label *encoding* yang dilakukan yaitu 4 merepresentasikan *high*, 3 merepresentasikan *normal*, 2 merepresentasikan *low normal*, dan 1 merepresentasikan *low action*.

Setelah proses modifikasi variabel selesai, dilakukan proses agregasi terhadap seluruh observasi yang ada di dalam *dataframe*. Hal tersebut dilakukan karena satu nomor U_ID bisa memiliki lebih dari satu riwayat pembayaran. Fungsi agregasi hanya dilakukan terhadap variabel-variabel yang akan digunakan dalam proses klasterisasi, yaitu variabel UNPAID_DEBT, YEARS_LATE, CREDIT_RATIO, NFLAG_INSURED_ON_APPROVAL, dan YIELD_GROUP. Variabel UNPAID_DEBT diolah dengan fungsi agregasi maksimum, YEARS_LATE dengan fungsi agregasi rata-rata, CREDIT_RATIO dengan fungsi agregasi minimum, serta NFLAG_INSURED_ON_APPROVAL, dan YIELD_GROUP dengan fungsi agregasi modus. Contoh sebelum dan sesudah data dilakukan agregasi diberikan pada Gambar 6.

SK_ID_PREV	U_ID	num_inst	unpaid_debt	credit_ratio	NFLAG_INSURED_ON_APPROVAL	YIELD_GROUP	years_late
1000023	350748	4	0	1.0	0.0	high	0.000000
1473990	350748	7	0	1.0	0.0	middle	0.008219
1997629	350748	10	0	1.0	0.0	low_normal	0.000000

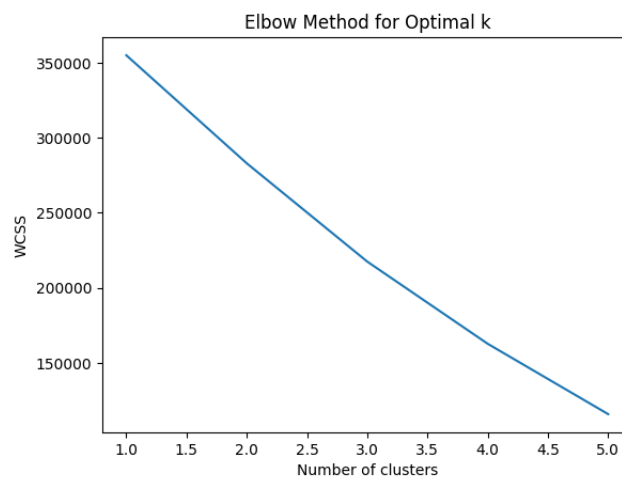
U_ID	unpaid_debt	years_late	credit_ratio	NFLAG_INSURED_ON_APPROVAL	YIELD_GROUP
350748	0	0.00274	1.0	0.0	high

Gambar 6: Sebelum dan sesudah dilakukan agregasi.

4.2 Klasterisasi Data untuk Pelabelan Data

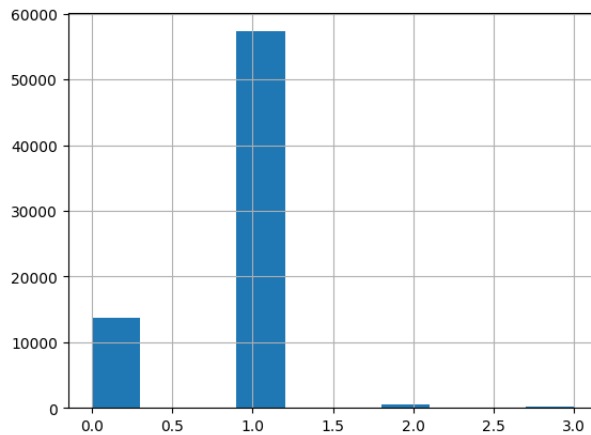
Dataframe baru yang telah selesai dimodifikasi dapat digunakan dalam proses klasterisasi. Sebelum dilakukan klasterisasi, seluruh nilai dalam *dataframe* dilakukan proses standarisasi agar memiliki rentang nilai yang sama. Proses klasterisasi dilakukan dengan menggunakan model *weighted k-means clustering*. Model *weighted k-means clustering* serupa dengan model *k-means clustering* yang biasa, tetapi perbedaannya terletak pada adanya pembobotan terhadap variabel yang diinginkan. Dalam kasus ini, variabel YEARS_LATE dikalikan dengan bobot sebesar 1,15, CREDIT_RATIO dikalikan dengan bobot sebesar 0,85, dan variabel NFLAG_INSURED_ON_APPROVAL dikalikan dengan bobot 0,95.

Untuk mencari jumlah klaster yang optimal dalam pembentukan label, digunakan *Elbow Method* yang hasilnya dapat dilihat pada Gambar 7.



Gambar 7: Elbow method untuk mencari klaster optimal

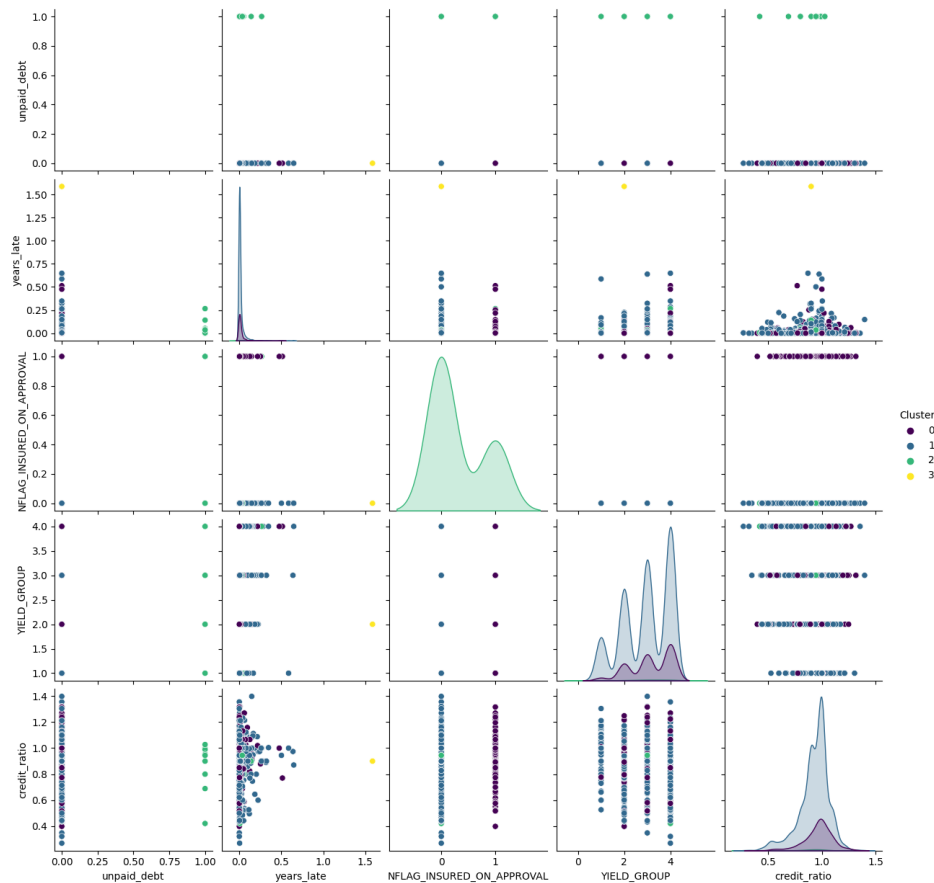
Pada nilai $k = 3$ dan $k = 4$, terlihat bahwa terdapat sedikit pembengkokan pada garis. Dengan demikian, jumlah klaster yang optimal untuk melakukan pelabelan data adalah sebanyak 3 atau 4. Walaupun demikian, penulis memutuskan untuk mengambil jumlah klaster sebanyak 4 kelompok dengan pertimbangan banyaknya variabel yang digunakan serta kemampuan interpretasi dari label yang dihasilkan. Klaster yang lebih banyak diharapkan dapat mengakomodasi variabel yang banyak serta memiliki kemampuan interpretasi yang lebih baik apabila dibandingkan dengan jumlah klaster yang lebih sedikit. Hasil persebaran label yang telah melalui



Gambar 8: Histogram pelabelan data

proses klasterisasi dapat dilihat pada Gambar 8 sementara hasil visualisasi *pairplot* diberikan pada Gambar 9. Berikut adalah interpretasi klaster berdasarkan visualisasi *pairplot*:

- Klaster 0 menggambarkan seorang nasabah yang hampir selalu membayar tagihan secara lunas dan tepat waktu. Selain itu nasabah pada klaster ini sering mengasuransikan pinjamannya. Klaster ini dapat disebut sebagai nasabah *low-risk insured*.
- Klaster 1 menggambarkan seorang nasabah yang hampir selalu membayar tagihan secara lunas dan tepat waktu. Namun, nasabah pada klaster ini jarang mengasuransikan pinjamannya. Klaster ini dapat disebut sebagai nasabah *low-risk uninsured*.
- Klaster 2 menggambarkan seorang nasabah yang sering tidak melunasi tagihan. Klaster ini dapat disebut sebagai nasabah *high-risk debt*.
- Klaster 3 menggambarkan seorang nasabah yang sering telat membayar tagihan. Klaster ini dapat disebut sebagai nasabah *high-risk debt*.



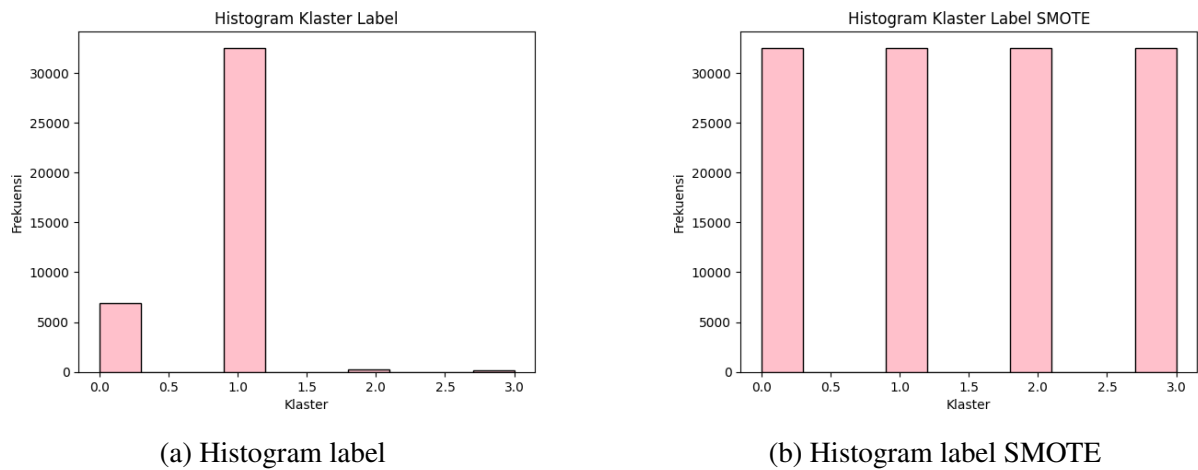
Gambar 9: Pairplot klaster data.

4.3 Penggabungan Data Input dengan Label

Setelah dilakukan pelabelan data, label tersebut digabungkan dengan data latih berdasarkan variabel U_ID yang sama. Dengan demikian, data latih memiliki satu buah variabel yang dapat digunakan sebagai variabel terikat (*output*). Data latih yang telah memiliki variabel bebas dan variabel terikat dapat digunakan dalam proses pemodelan untuk mencari model terbaik yang dapat membantu proses pengambilan keputusan peminjaman uang yang didasari oleh karakteristik nasabah.

Sebelum melakukan pemodelan data, data latih dipisah menjadi set data pelatihan dan set data validasi dengan rasio 85% dan 15% secara berurutan. Selain itu, pemisahan data dilakukan dengan menggunakan *stratify* agar proporsi kelas dalam set data pelatihan dan validasi akan tetap sama seperti proporsi kelas dalam set data aslinya.

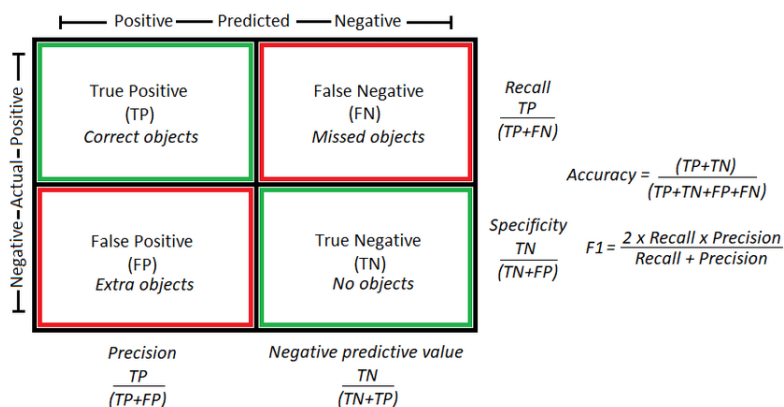
Selain itu, dilakukan metode *resampling* yaitu SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menangani ketidakseimbangan kelas. Hal ini dilakukan untuk mengurangi kebiasaan pada kelas mayoritas dan meningkatkan kemampuan model untuk mengenali pola dalam kelas minoritas. Perbedaan dari data sebelum dan setelah dilakukan *resampling* SMOTE dapat dilihat pada Gambar 10.



Gambar 10: Perbedaan data normal dengan data *resampling* SMOTE.

5 Pelatihan dan Hasil Pemodelan

Pelatihan model dilakukan menggunakan data data yang sudah dilakukan *resampling*. Selanjutnya, dilakukan pemodelan dengan sejumlah model klasifikasi yaitu *Decision Tree*, *Random Forest*, *XGBoost*, *LightGBM*, *Logistic Regression*, *Artificial Neural Network*, *Logistic Regression*, dan *AdaBoost*. Hasil prediksi dari setiap model dievaluasi menggunakan metrik *confusion matrix* seperti pada Gambar 11. Evaluasi utama yang akan digunakan pada laporan ini adalah nilai *precision*, *recall*, dan *F1-Score*. *Precision* dan *recall* adalah dua metrik evaluasi kinerja model untuk kelas tertentu, sedangkan *F1-Score* merupakan metrik gabungan antara *precision* dan *recall*.



Gambar 11: Metrik untuk klasifikasi

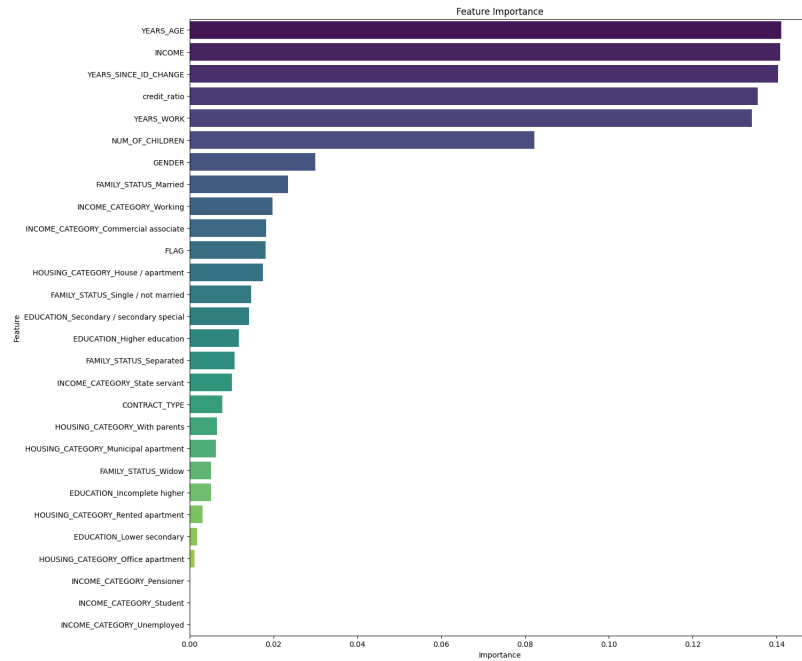
Tabel 2: Hasil prediksi dari setiap model

Model	Accuracy	F1-Score Cluster 0	F1-Score Cluster 1	F1-Score Cluster 2	F1-Score Cluster 3
Decision Tree	0.63	0.24	0.76	0.00	0.00
Random Forest	0.74	0.20	0.85	0.02	0.00
LightGBM	0.77	0.12	0.87	0.03	0.00
XGBoost	0.76	0.17	0.86	0.02	0.00
Artificial Neural Network	0.82	0.00	0.90	0.00	0.00
Logistic Regression	0.41	0.27	0.58	0.02	0.01
AdaBoost	0.52	0.24	0.68	0.01	0.01

Berdasarkan eksplorasi data sebelumnya, diketahui jika data yang digunakan tidak seimbang. Akurasi dari sebuah model tidak menggambarkan performa secara keseluruhan. Misalnya berdasarkan Tabel 2, diperoleh bahwa nilai akurasi tertinggi diperoleh dari model ANN, meskipun demikian, model ANN memprediksi seluruh data uji menjadi kluster 1. Hal ini diketahui dari nilai *F1-Score* untuk kluster 0, 2, dan 3 adalah nol. Sehingga penting metrik *F1-Score* sebagai penentu performa sebuah model.

Berdasarkan Tabel 2, model dengan akurasi tertinggi untuk memprediksi kluster 0 adalah model *Logistic Regression*, sedangkan model terbaik untuk memprediksi kluster 1 adalah *LightGBM*. Namun, tidak ada model yang dapat memprediksi kluster 2 dan 3 secara baik. Sebagai contoh, model *LightGBM* hanya memiliki nilai sebesar 0.03. Di antara semua model di atas, model *XGBoost* memiliki hasil performa paling merata.

Meskipun nilai *F1-Score* dari ketiga model pohon tidak berbeda jauh, model *XGBoost* memiliki nilai tengah antara model *Random Forest* dan model *LightGBM*. Model tersebut dapat diinterpretasikan memiliki akurasi yang sangat tinggi untuk memprediksi kluster 1 atau nasabah yang taat membayar tagihan dan mengasuransikan pinjamannya. Pinjaman yang diasuransikan dapat memberi jaminan kepada STC PayLater, apabila terjadi hal tidak terduga pada nasabah tersebut, STC PayLater masih akan mendapatkan bayaran tagihan. Selain itu, model dapat memprediksi kluster 0 atau nasabah yang taat membayar tagihan, tetapi jarang mengasuransikan pinjamannya. Terakhir, model hanya dapat memprediksi sekitar 2% dari nasabah yang tidak melunasi tagihan pinjaman dan sama sekali tidak dapat memprediksi kluster 3 atau nasabah yang akan membayar hutang secara telat.



Gambar 12: *Feature importance*

Gambar 13 merupakan *feature importance* yang diperoleh dari hasil pelatihan model *Random Forest*. Berdasarkan gambar tersebut, dapat dilihat bahwa beberapa variabel terpenting dalam menentukan label seorang nasabah adalah YEARS_AGE, INCOME, YEARS_SINCE_ID_CHANGE, credit_ratio, serta YEARS_WORK.

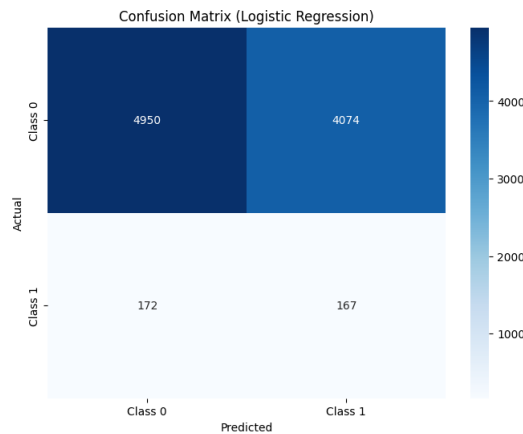
5.1 Pengembangan untuk Memodelkan Nasabah yang Tidak Taat Membayar Tagihan

Pada bagian sebelumnya, model belum dapat mendeteksi nasabah yang membayar tagihan secara lunas dan tepat waktu. Untuk mengatasi hal tersebut, akan dibuat sebuah variabel baru yaitu BAD_CUSTOMER. Seorang nasabah tergolong sebagai *bad customer* apabila jumlah hutang yang dimilikinya di atas 10 satuan uang serta memiliki rata-rata telat membayar lebih dari 0,05 tahun atau 19 hari. Variabel ini akan menggantikan label yang dibuat dengan menggunakan kluster. Dengan demikian, hasil pemodelan dapat dilihat pada Tabel 3.

Tabel 3: Hasil prediksi model

Model	Accuracy	F1-Score Good Customer	F1-Score Bad Customer
Decision Tree	0.89	0.94	0.05
Random Forest	0.94	0.97	0.04
LightGBM	0.94	0.97	0.03
XGBoost	0.89	0.97	0.03
Logistic Regression	0.55	0.70	0.07

Berdasarkan Tabel 3, dapat dipertimbangkan bahwa model terbaik untuk memprediksi seorang nasabah yang tidak taat membayar tagihan adalah *Logistic Regression*. Berdasarkan Gambar 12, model ini dapat diinterpretasikan jika 49% dari seluruh nasabah yang tidak taat membayar tagihan berhasil diidentifikasi. Akan tetapi, sebanyak 45% dari nasabah yang taat membayar tagihan diidentifikasi sebagai nasabah yang tidak taat.



Gambar 13: Hasil prediksi model *Logistic Regression*

6 Kesimpulan dan Saran

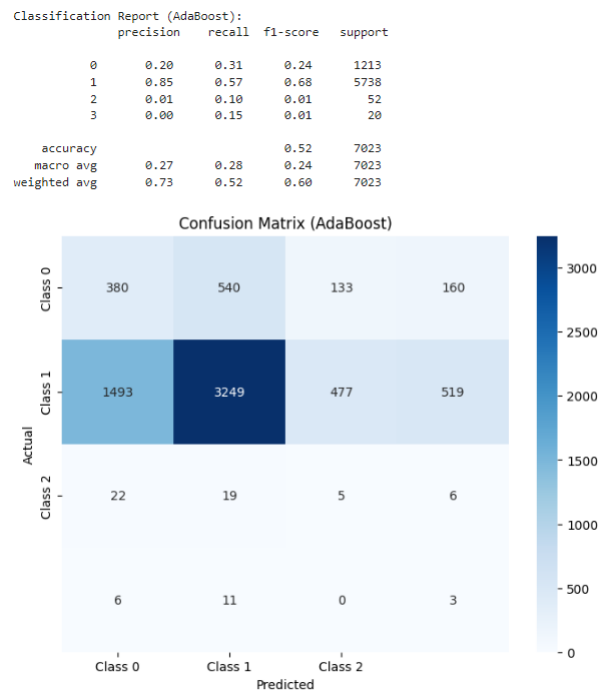
Berdasarkan laporan ini, dapat disimpulkan beberapa hal berikut:

1. Dengan memanfaatkan dataset *payment history* dan *previous application*, dapat dilakukan klusterisasi yang mencerminkan nasabah dengan kriteria tertentu. Hasil klusterisasi ini kemudian dapat berfungsi sebagai label dari data training. Data tersebut dapat diklasifikasi menjadi *low-risk insured*, *low-risk uninsured*, *high-risk debt*, dan *high-risk late payment*.
2. Hasil pemodelan data training dengan label klaster terbaik diperoleh oleh model *XGBoost*. Model tersebut dapat mengidentifikasi nasabah yang membayar tagihan secara tepat secara baik, tetapi memiliki kekurangan memprediksi nasabah yang tidak taat membayar tagihan.
3. Hasil pemodelan data training dengan label *BAD_CUSTOMER* terbaik diperoleh oleh model *Logistic Regression*. Model tersebut dapat mengidentifikasi sebanyak 49% nasabah yang tidak taat membayar tagihan. Namun, sebanyak 45% nasabah yang taat membayar tagihan salah teridentifikasi.

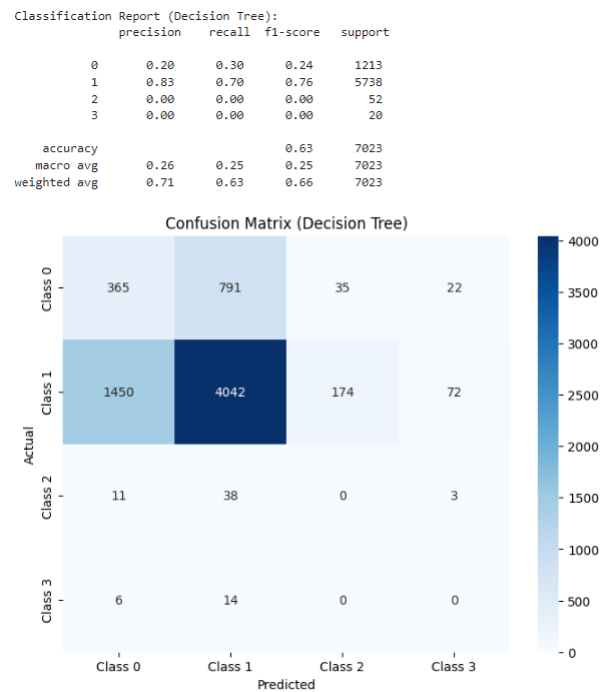
Selain itu, adapun saran yang perlu dilakukan untuk memperoleh hasil yang lebih baik:

1. Melakukan analisis training data lebih lanjut untuk mengetahui variabel yang memiliki pengaruh lebih besar sebagai penentu nasabah yang *low risk* atau *high risk*.

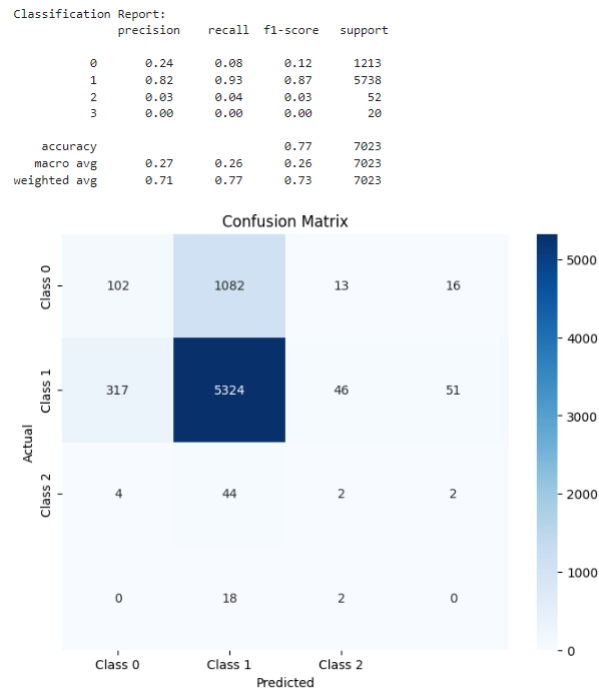
7 Lampiran Gambar



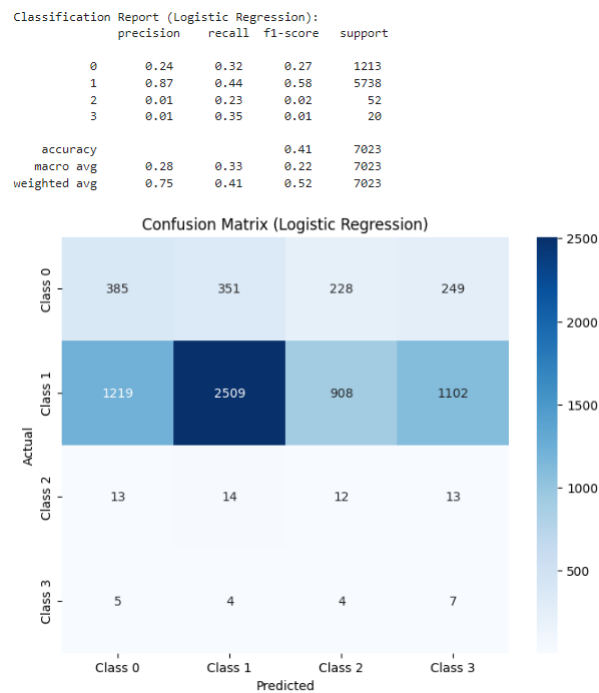
Gambar 14: Hasil prediksi model *AdaBoost*



Gambar 15: Hasil prediksi model *Decision Tree*

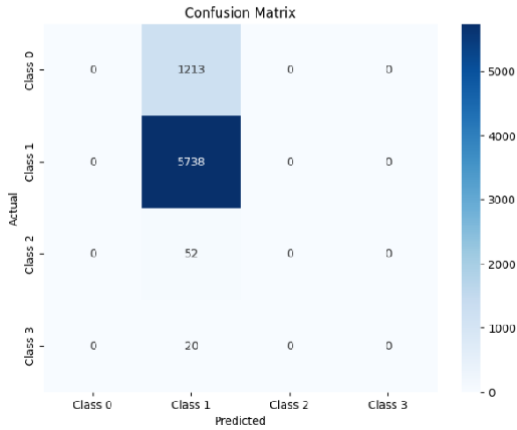


Gambar 16: Hasil prediksi model *LightGBM*



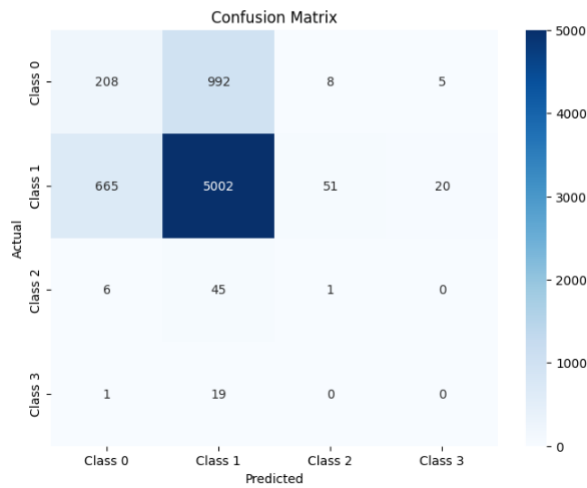
Gambar 17: Hasil prediksi model *logistic regression*

Classification Report:				
	precision	recall	f1-score	support
0	0.00	0.00	0.00	1213
1	0.82	1.00	0.90	5738
2	0.00	0.00	0.00	52
3	0.00	0.00	0.00	20
accuracy			0.82	7023
macro avg	0.20	0.25	0.22	7023
weighted avg	0.67	0.82	0.73	7023



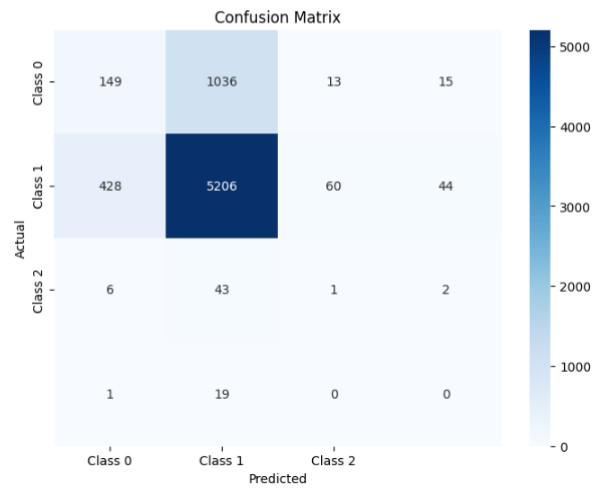
Gambar 18: Hasil prediksi model *Neural Network*

Classification Report:				
	precision	recall	f1-score	support
0	0.24	0.17	0.20	1213
1	0.83	0.87	0.85	5738
2	0.02	0.02	0.02	52
3	0.00	0.00	0.00	20
accuracy			0.74	7023
macro avg	0.27	0.27	0.27	7023
weighted avg	0.72	0.74	0.73	7023



Gambar 19: Hasil prediksi model *Random Forest*

Classification Report:				
	precision	recall	f1-score	support
0	0.26	0.12	0.17	1213
1	0.83	0.91	0.86	5738
2	0.01	0.02	0.02	52
3	0.00	0.00	0.00	20
accuracy			0.76	7023
macro avg	0.27	0.26	0.26	7023
weighted avg	0.72	0.76	0.74	7023



Gambar 20: Hasil prediksi model *XGBoost*